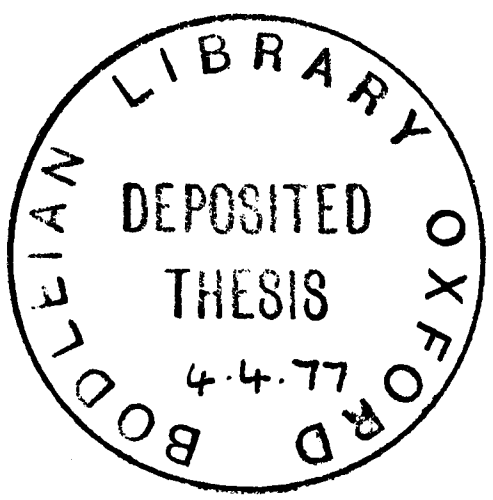
THE STRUCTURE OF SIMULTANEOUS EQUATION ESTIMATORS

by

A. Hughes Hallett

Submitted for the degree of D.Phil. at the University of Oxford,

October 1976.



for

Claudia

ABSTRACT

The thesis is concerned with developing a coherent theory of estimation suitable for the standard simultaneous equation econometric model. Particular attention is paid to the case of large models and many of the estimators considered may be used when the sample size is inadequate, i.e. there are more predetermined variables than observations in the model. However, the analysis is not restricted to this case. A great many new estimators have been proposed in recent years but little is known of their properties or performance. I consider iterative estimators (under a classical model specification) from the repeated least squares and instrumental variable groups, as well as conventional and maximum likelihood estimators. Included at the same time are four separate methods of obtaining maximum likelihood estimates. The questions to be answered for each estimator are:

(a) Have they the properties of consistency and efficiency? Are their statistical properties affected in any way by the numerical algorithm used to obtain them?

(b) Under what parameters of the model and data does each estimator exist?

(c) What is their suitability or ease of use from a computational view point?

(d) Which estimator entails the smallest numerical burden?

(e) What are the relations between them and to conventional maximum likelihood estimators? Is a family structure of estimators available?

Finally a comparison is made to provide, in given circumstances, a preference ordering over the available estimators. This is done from the answers to the above questions, after a discussion using known results to show the impossibility of making any general comparisons on small sample properties.

I can view each estimator as an iterative process with a pre-specified or infinite number of iterates and a particular starting point. I therefore begin by reviewing the consistency efficiency, optimality, availability and starts of the estimators. They are then classified according to iterative algorithms; and a new algorithm is suggested which is computationally superior.

There is a substantial analysis of the iterative solution of equation systems. New convergence theorms, theorems on optimal acceleration parameters, orderings, etc. and comparison theorems are developed here. Both linear and nonlinear iterations are considered, but for large systems linear iterations appear adequate and are generally superior computationally. Thus the analysis of the thesis extends to nonlinear estimation techniques and iterative estimators involving nonlinear iteration. Next relations with maximum likelihood estimators are developed and a complete classification and ranking by asymptotic efficiency is provided. The iterative schemes involved in the estimators may then be analysed for convergence conditions and optimal convergence conditions for their existence and availability are established and it is possible to make recommendations on the best way of using any iterative estimator, and, with some numerical analysis, to provide a preference ordering among them. Maximum likelihood estimators are also included in the computational analysis. The general conclusion is that iterative estimators, appropriately used, are always competitive with conventional methods and except for very small models are superior from the computational point of view. A number of numerical examples are presented illustrating the effectiveness of these results. A sample of nine well known models of various sizes has been used for this purpose.

The main contributions are intended to be:

(1) the classification and statistical properties of the estimators considered;

(2) several new theorems on the convergence, comparison and optimal use of various iterative solutions to systems of equations;

(3) the first computational analysis of alternative estimators;

(4) a coherent estimation theory for the simultaneous equation model;

(5) a preference ordering over the estimators.

The theory is important in itself and because most of the large, well known models have so far been estimated inconsistently and inefficiently for lack of such a theory. They remain however the typical models in econometrics. Finally the whole structure and theory of the estimators considered is summarised in an estimator generating function.

ACKNOWLEDGMENTS

I must acknowledge encouragement from many quarters. Nuffield College, Oxford, provided the very stimulating atmosphere in which I first began to study the problems treated here; and in particular John Fleming "caused" me to address myself to them. Chris Gilbert and other colleagues at Bristol have been most helpful in bringing the ideas to fruition.

The major vote of thanks goes to Professor J.A.C. Brown who, as supervisor, maintained a voice of calm throughout the clamour of ambitious mathematics. Mark Salmon was a major influence. He effectively launched my analysis by convincing me (in many conversations held over a long period of unlikely but convivial circumstances) of the practical importance of the problem. Although I retain complete responsibility for the results presented here, two people are responsible for the fact that any were produced at all. Mike Lowcock supervised the computer in calculating what was intended, rather than let anarchy reign. Pat Shaw typed the manuscript - part of it several times - so satisfactorily that I could accept no other.

Bristol,

August 1976.

CONTENTS

PART I

		Page No.
CHAPTER 1:	INTRODUCING THE ESTIMATION PROBLEM	
1.1	Introduction	1
1.2	Estimation problems: maintained hypotheses	4
1.3	Equation solution algorithms	7
1.4	Signposting	8
CHAPTER 2:	CONVENTIONAL ESTIMATION - THE BACKGROUND KNOWLEDGE	
2.1	Introduction	10
2.2	Model and notation	10
2.3	Conventional estimators	19
2.4	Information classes	39
2.5	Computation	44
CHAPTER 3:	THE OBJECTIVES AND PROCEDURE	
3.1	Introduction	53
3.2	Selection criteria	54
3.3	The empirical models	58
3.4	Empirical differences due to estimation technique	59
3.5	Conclusions	64

PART II

		Page No.
CHAPTER 4:	ALTERNATIVE ESTIMATORS	
4.1	Introduction - recursive systems	66
4.2	Conditions for consistency	66
4.3	Generalised inverses	72
4.4	Fisher's SOIV method	73
4.5	Principle components	75
4.6	Prior knowledge methods	80
4.7	Iterative estimation	84
4.8	Summary	92
CHAPTER 5:	A MAXIMUM LIKELIHOOD INTERPRETATION OF ITERATIVE ESTIMATION	
5.1	LIML as an instrumental variable estimator	94
5.2	The asymptotic equivalence of iterative and maximum likelihood estimators	97
5.3	An identity between iterative and maximum likelihood estimators	98
5.4	Two illustrations	101
5.5	A summary of results	101
5.6	Conclusions	105
CHAPTER 6:	GENERAL INTERDEPENDENT SYSTEMS - WHAT IS THE TRUE FAITH?	
6.1	Introduction	107
6.2	Definitions	107
6.3	Inconsistency of FP estimates	110
6.4	An extension of theorem 1	114
6.5	Comments	118
6.6	Comparison with IIV estimates	119
6.7	GEID conditions for other estimators	122
6.8	Other GEID conditions	123
6.9	Autoregressive models and serial correlation	124

Page No.

CHAPTER 7: EXTENSIONS OF ITERATIVE ESTIMATION THEORY

7.1 Introduction 127

7.2 Some results for (2.41), (5.10) and (4.27) 128

7.3 A k-class FIML estimator 130

7.4 Discussion 136

7.5 Asymptotic inefficiency of limited information estimators when full information conditions apply 139

7.6 Autoregressive iterative estimators 142

PART III

CHAPTER 8: AN ANALYSIS OF THE PROBLEM OF COMPLETE MODEL SOLUTION I

8.1 The use of algorithms 149

8.2 The current knowledge 150

8.3 Stability conditions for polynomials 153

8.4 The Jacobi and JOR methods 158

8.5 Gauss-Seidel and SOR methods 165

8.6 Alternative methods for calculating α^* 174

8.7 Ordering and normalising problems 176

8.8 Conclusions 184

CHAPTER 9: EXTENSIONS OF SOR THEORY

9.1 Introduction 185

9.2 Comparison theorems 186

9.3 Tests for the superiority of SOR 191

9.4 Bounds for testing 192

9.5 Extensions of SOR 198

9.6 Approximate α^* strategies 206

9.7 Permuted matrices 214

Page No.

CHAPTER 10: COMPLETE MODEL SOLUTION II: NONSTATIONARY METHODS

10.1 Introduction 216

10.2 Richardson's stationary method 217

10.3 Aitken's double Seidel process 227

10.4 Alternating direction implicit methods 230

10.5 Nonstationary methods: first order 233

10.6 Semi-iterative methods 241

PART IV

CHAPTER 11: A SURVEY OF ITERATIVE ESTIMATORS AND THEIR STRUCTURE

11.1 Introduction 256

11.2 The family structure of an iterative estimator 257

11.3 The independent decisions 260

11.4 A new iterative scheme: a digression 263

11.5 The independent decisions (cont.) 268

11.6 Relaxing assumptions 276

11.7 A convention on naming estimators 276

CHAPTER 12: CONVERGENCE ANALYSIS IN ITERATIVE ESTIMATION

12.1 Introduction 278

12.2 A framework for general convergence analysis 279

12.3 Convergence conditions - FP 291

12.4 Convergence conditions - IIV 302

12.5 The QRF instrument case 308

12.6 Global convergence and nonlinearity 313

12.7 Sphericality transforms 316

12.8 Existence theorems 321

Page No.

CHAPTER 13: THE COMPUTATION OF ITERATIVE ESTIMATORS

13.1 Introduction 324

13.2 Identities in the specification 325

13.3 The computational burden of the general iterative estimator 338

13.4 Comparisons in the computation of iterative estimator 348

13.5 Comparing (27) and (29) type estimators 358

13.6 Comparison of iterative estimators with LIML and FIML 367

13.7 OLS computation and QR transformation 374

Appendix 1: Numerical tables 376

Appendix 2: Decision rules including initial calculations 391

PART V

CHAPTER 14: ON THE USE OF ITERATIVE ESTIMATORS

14.1 Introduction 394

14.2 Guaranteed convergence 394

14.3 Plural convergence points 405

14.4 Apparent convergence problems 409

14.5 Optimal equation orderings 410

CHAPTER 15: NONLINEAR MODELS

15.1 Introduction 412

15.2 The background analysis 414

15.3 Nonlinearity in the parameters 417

15.4 Nonlinearity in the variables 419

15.5 General nonlinearities 423

15.6 Statistical properties 427

15.7 Conclusions 428

		Page No.
CHAPTER 16:	THE USE OF FINITE SAMPLE PROPERTIES	
16.1	Introduction	429
16.2	Mean square error estimation	430
16.3	The existence of moments	433
16.4	Exact bias and variances	434
16.5	Approximate finite sampling theory	439
16.6	Monte Carlo evidence	443
16.7	Conclusions	444
CHAPTER 17:	THE ESTIMATOR GENERATOR FUNCTION AND CONCLUSIONS	
17.1	Introduction	446
17.2	The estimator generator function	446
17.3	Algorithm structure specification	451
17.4	Preference ordering among estimators	453
APPENDIX:	SOME MATRIX THEORY AND STATISTICAL REFERENCES	
A.1	Matrix factors which preserve stability	454
A.2	Cramer's generalised theorem and an extension	460
A.3	Nilpotent matrices	467
BIBLIOGRAPHY		

PREFACE

My interest in the subject of this study started when I was reviewing the problems of estimating large econometric models with a restricted availability of data. This is still the typical situation in econometric model building. My interest was sharpened by the uncertainty over the properties of the new estimators and because with so many estimators to choose from, a preference ordering of some kind was now essential. Although "Iterative estimation theory" might be a suitable subtitle, the theory would not be complete without setting it within the context of conventional estimation theory. The important results concern the relations between estimators, their statistical properties, and their computational and numerical analysis. I have therefore attempted something more general than is usual.

CHAPTER 1: INTRODUCING THE ESTIMATION PROBLEM

1.1 Introduction

Undoubtedly the major contribution of econometrics has been to provide the means to identify and measure the important relationships between economic variables. In this study we are concerned with any process that can be described by a set of stochastic simultaneous equations so that the results have wider application than economics itself; although modelling of economic processes is the purpose envisaged here.

Interrelationships of economic variables have been displayed in a wide variety of econometric models, measuring the size and significance of the dependence of variables on one another from a set of historical data. Many national econometric models have been constructed. The earliest were those by Tinbergen for the Netherlands and U.S.A. The success of these models together with the considerable advances in macroeconomic theory following the "Keynesian Revolution" has inspired the construction of a whole range of national econometric models. In the United States, very simple models first estimated by Klein led to the medium size models of Klein-Goldberger and Michigan University, both of which have been updated and expanded in several versions. Recently there has been a strong tendency for the size and detail of models to grow and, with the acceptance that national econometric models are here to stay for planning purposes, we have entered the era of large models. In this class are the Brookings model with several hundred equations and exogenous variables requiring the estimation of several thousand coefficients; the Wharton model in its various versions between 70 and 200 equations; the Data Resources Inc. model, the OBE model and the FRB-MIT-Penn models all with between 100 and 500 equations.

In Britain, at present, the range is more restricted. An early small model was constructed by Klein and others at Oxford; a model is maintained by

the Treasury; a model of about 100 equations, used for forecasting, has been constructed by the London Business School and a medium size model is under construction at the University of Southampton. Models of this kind are in use in most other countries of the world, notably the Netherlands, Canada, Sweden, Germany, France, Japan, Israel, Taiwan, Greece, etc. The final step of linking various national models to provide an explanation of world trade flows has been undertaken in Project LINK.

Models have also been constructed for developing economies such as India or Columbia requiring a rather different specification. Models of dependent or regional economies have also been attempted, including models of Puerto Rico and Michigan State.

The models cited above are not the only examples of simultaneous equation models in popular use. Commodity models, often also quite large and detailed in their explanation of different components of the supply and demand of the given commodity, and perhaps its immediate competitive and complementary goods, are widely used to explain movements in the given market. Such diverse commodities as tea, coffee, cocoa, jute, water melons, meat, sugar and onions have been treated. Metals such as copper, tin, zinc, tungsten, etc. have also been treated. The markets for manufactured goods have been modelled, e.g. cars, sweets, and specific industries such as rail transport have also been considered. Again models have been constructed, with considerable detail, of particular sectors of an economy. Perhaps the most prominent are those of the financial sector (such as the St. Louis model) or the well known complete demand systems models, or those of the housing market, or even the American textile industry. And there are many more examples which might be cited.

The potential applications for econometric estimation techniques are therefore enormous. The most straightforward use of a model is simply to display an explanation of the behaviour of an economic variable in terms of its relationship with other variables. It is frequently remarked that it is

relatively easy to understand a single relationship, but the possible effects of several simultaneous relationships acting in concert on a variable, perhaps with feedback influences, is next to impossible to grasp. However the explanations are presented - whether in terms of estimated coefficients, by calculating and tracing of effects of a shock to the system, or by simulation of various test conditions - they will certainly require each coefficient to be estimated as accurately as possible since these are the points of interest. We naturally aim for unbiased and efficient estimators in the most general circumstances of data and model specifications; or failing that, those that are asymptotically so in terms of the density functions of the estimated coefficients. In other words, the statistical maintained hypothesis contained in the estimator used is itself of no interest. The focus of attention is quite obviously on the economic model and what it can tell us, and we want the estimator which least involves its own maintained hypothesis in the maintained hypothesis of the economic specification.

Similarly if the use of the model is to test hypotheses of economic behaviour, we equally require the separation of hypotheses. Hypothesis testing is most simply carried out directly on the coefficients or indirectly in relations between coefficients, on reduced or final form coefficients. Evidently these tests should ideally be unbiased and efficient conditionally on the economic maintained hypothesis, uncluttered by the statistical one. Forecasting likewise requires to be conditioned only on the economic maintained hypothesis and exogenous information. This proposition obviously extends to the use, frequently made by governments, planning agencies and businesses, of simulating likely outcomes of various policy packages or changes in economic conditions. Multiplier analysis for policy effectiveness and any attempts to stabilise or control the economy similarly point to the same requirement.

It is therefore particularly important that we should be sure of distinguishing between economic and statistical maintained hypotheses. If possible

they should not be confused at all in building an econometric model; and at least the statistical maintained hypothesis should be separated out and relaxed as far as possible. This is the central thesis in the following work and requires estimators to be examined for their properties, and classified by them, together with the necessary statistical assumptions.

1.2 Estimation problems: maintained hypotheses

Single equation models may be estimated by ordinary least squares (OLS). Simultaneous equation models are more suitably estimated by two stage least squares (2SLS), three stage least squares (3SLS), limited information maximum likelihood (LIML) or full information maximum likelihood (FIML). Any model may be estimated by instrumental variable techniques (IV) suitably used (Aitken's generalised least squares, GLS, or its IV version - GIV - may be used where required in place of OLS or IV). These are the traditional econometric estimators. Individually they are well understood. Their statistical properties are well known, as are the assumptions those properties require.

Recently there has been an alarming increase in the number of simultaneous equation estimators proposed in econometric literature, and most are in use. In this new group are estimators which involve generalised inverses, structurally ordered instrumental variables (SOIV), principal components, or prior information. Also particularly important are the iterative estimators of the fixed point and iterated instrumental variable families (broadly defined to include those iterated to convergence and those with a prespecified number of iterations such as LIVE and FIVE in Brundy and Jorgenson (1971)). In each case there are many different computational algorithms which produce an estimator with the same basic structure (see Chapter 4). Largely unknown are the statistical properties of these estimators relative to those conventional ones, at least in terms of consistency and, particularly, asymptotic efficiency. It is also of interest to see what assumptions are necessary to produce those

properties in each estimator, i.e. if the statistical maintained hypothesis is more stringent, or hopefully, more relaxed, than for conventional estimators. Further it is important to know what influence, if any, the different algorithms used in a given estimator have on either point.

It would be beneficial to establish the relationship that any of these new estimators has to any of the conventional estimators; as well as what relationships these new estimators have one to another. In fact there are few comparison theorems for conventional estimators in the literature; and none for the new group of estimators, nor for the new group compared to the conventional ones. It is, for example, well known that LIML and 2SLS or FIML and 3SLS are asymptotically equivalent and that FIML and 3SLS are at least as efficient as LIML and 2SLS respectively. It is also well known that FIML $\equiv$ LIML $\equiv$ 3SLS $\equiv$ 2SLS for just identified equations; and 3SLS $\equiv$ 2SLS in exactly identified equations or those with structural errors are uncorrelated with any other. Beyond this nothing is established.

A search for comparison, or equivalence, theorems naturally introduces the possibility of selecting a preferred estimator in given circumstances. A preference ordering may be based on statistical properties as is implied above and in the thesis of the separation of maintained hypotheses. However, other factors such as computational efficiency, numerical stability and accuracy, relaxation of necessary assumptions, robustness, or ease of programming may well enter into the calculation of the preference ordering.

Another natural by-product of comparison theorems would be the establishment of a structure embracing all the estimators. This has been attempted (Chow (1964)) for conventional estimators. However that particular scheme depends on minimising Euclidean norms of various matrices relative to various others and it is not easy to see how the conventional estimators fit in, and it is quite impossible to fit any unconventional estimators in to it. Another scheme is that of Hendry (1976) which treats estimators as numerical approximations to FIML. It is suggested here that this scheme is more complicated

than necessary and does not fully show the relations between the estimators or the degree of the approximations involved. Instead an estimator generating function is proposed with the property that by specifying various values of the parameters any estimator, considered here, is produced. Having then divided the estimators among asymptotically equivalent groups, computational, convergence and "availability" analysis, together with a summary of the necessary statistical maintained hypothesis, is used to derive a selection scheme for the preferred estimator.

A particular problem in estimation has been introduced by the specification of large models. It is obvious, on consideration of the estimating equations, that conventional estimators break down whenever the sample size is less than the number of predetermined variables in the model. It will be clear that conventional estimators as traditionally applied are unavailable.

The formal definition of this problem of inadequate samples is as given above (negative degrees of freedom) - but a less formal definition would include a very low number of degrees of freedom since it is then likely that the actual statistical properties of various estimators are far from the asymptotic ones claimed for them.

In view of the prevalence of large models (the Brookings, Wharton, Michigan and other large American models mentioned all suffer from this problem) it seems sensible to analyse particularly those estimators which are available in inadequate samples. This analysis is important because large models are likely to be more frequently specified in the future in order to provide a more detailed description of the components of an economy. Explanations of the movements of highly disaggregated variables, of particular markets or industries, are now within reach and are considered desirable in themselves and to give detailed forecasts. Aggregate forecasts contain too little useful information being virtually in the class of naive models, and the models generating them suffer from well known aggregation biases. Aggregate models

cannot adapt to reflect structural changes, and in particular the incidence of government policy instruments such as tax rate movements. It is sometimes noted that a very detailed model is required to portray the effects of instruments, which may be very different for different instruments - say, contrasting the effects of changing surtax rates and family allowance rates.

1.3 Equation solution algorithms

A general proposition of literary criticism is that a piece of writing should argue for both a major and more obvious theme, as well as a secondary more submerged theme. This study attempts to honour this basic principle. The major theme is that discussed in the separation of hypotheses and the structure and analysis of the larger set of estimators. The minor theme is an analysis of iterative methods of simultaneous equation solution - linear or nonlinear. Some of the methods are well known to econometricians but many will be quite new. However, a theoretical examination of even the well known methods has never been made available to econometricians. A knowledge of the convergence properties and how to use the acceleration parameters optimally, plus comparison theorems for the various methods are required. The knowledge available elsewhere is due to research for applications in the natural sciences, where as a result of carefully designed controlled experiments it is possible to deal only with equation systems having a positive definite and symmetric iteration matrix, or one that is block cyclic.

Here convergence theorems, comparison theorems, results on the use of acceleration parameters and so on, are reworked for the case of an arbitrary real iteration matrix, as occurs in econometric systems. The results obtained are used at many points in the analysis of iterative estimators. This secondary theme is therefore offered as no less a contribution than the estimation theory itself.

1.4 Signposting

The argument and analysis presented here falls into five distinct phases. Part I is introductory and contains three chapters of which this is the first. The second contains a specification of the model and its assumptions, how far they may be relaxed, and a summary of the conventional estimators; their statistical and computational properties and comparison theorems. This chapter contains no new results but is set up to provide a benchmark for subsequent theory in a form for easy reference. The third chapter contains the selection criteria for preferred estimators, a justification for using asymptotic statistical properties and some empirical justification for searching for a desired estimator, based on nine well known econometric models.

Part II introduces the unconventional estimators, and variants of those estimators, and establishes their statistical properties. Chapter 4 introduces each basic estimator and indicates any difficulties with it in terms of the selection criteria. Several estimators are filtered out at this stage leaving only the FP and IIV families for further analysis. Chapter 5 establishes a maximum likelihood interpretation of IIV estimators, asymptotical equivalence with FP and treats their statistical properties and sampling variances. Chapter 6 investigates the possibility of relaxing the statistical maintained hypothesis in the form of the General Interdependent (GEID) system - and reaches negative conclusions. Chapter 7 extends the analysis so far, to provide an estimator of each type for every information class, including autoregressive estimators in the presence of lagged dependent variables. Various formulations are discussed. Further results include a linear FIML estimator and simplified proofs of some comparison theorems in maximum likelihood estimators.

Part III deals with solution algorithms for systems of simultaneous equations. Chapter 8 derives convergence conditions and optimal acceleration parameters and equation orderings for the well known Jacobi and Gauss-Seidel

methods. Chapter 9 attempts to simplify the acceleration parameter analysis for Gauss-Seidel methods by a variety of approximations and also provides some approximate tests of its superiority over Jacobi in a general case. Chapter 10 deals with nonstationary methods and concludes the semi-iterative accelerations superimposed on any earlier method guarantees convergence at reasonable speeds.

Part IV returns to iterative estimation, combining the results of Parts II and III. First Chapter 11 surveys the iterative estimator algorithms, and starts; and sets up a family structure for FP and IIV estimators. A new algorithm is suggested. Chapter 12 then obtains convergence criteria for each family member for both families - aiming for both local and global criteria. Chapter 13 investigates the computational burden of each estimator of the iterative families, and having selected the most economical (they were all equal on the other selection criteria) compares it with conventional estimators. The final part treats some remaining loose ends in the theory. Chapter 14 contains some practical results on how to use iterative estimators so that they converge and treats problems of plural convergence points, apparent convergence and equation ordering. Chapter 15 generalises all the estimators for nonlinear model specifications and indicates the preferred estimator. Small sample statistical properties are reviewed in Chapter 16 to see if any improvements to estimator selection can generally be made over the scheme so far - with negative conclusions. Finally Chapter 17 contains the final estimator classification by means of the estimator generating function from which they can all be obtained in asymptotically equivalent groups; and summarises the preference ordering on estimators.

CHAPTER 2: CONVENTIONAL ESTIMATION - THE BACKGROUND KNOWLEDGE

2.1 Introduction

The purpose of this chapter is to set the scene for the subsequent analysis. The precise form of the model, which we wish to estimate, is specified together with the set of assumptions which apply to it. Then the conventional estimators, which, to date, have been almost exclusively used for such models, are defined and their statistical properties summarised. Starting points are provided for investigation of the other estimators that have been proposed. Additional motivation is provided by the unavailability of estimators in inadequate samples, and by suggestions that an estimation strategy is necessary, since we find in Chapter 3 the same model estimated with different estimators, can have markedly different properties affecting economic policy analysis, control, growth patterns and stability.

A standard notation is established in Section 2 for the conventional simultaneous equation econometric model. The notation will be used throughout, although it is amplified at a few points. The conventional estimators are treated in Section 3. All the information here is standard econometric theory. Finally, we review the information content of each estimator in Section 4 and make some remarks on the computational problems in Section 5.

2.2.1 The model and notation

The first problem in econometric work is the specification of a functional relationship between a particular variable y_i whose behaviour we wish to explain and a set of explanatory variables X_i such that the observation y_{ti}, at time t, is generated by

$$y_{ti} = f_i(X_{ti}, u_{ti}) \quad t = 1, \ldots T \tag{1}$$

Generalising, we have, overall, a set of M variables Y_t, whose behaviour is to be explained, and a set of K independent variables Z_t, where Y_t and Z_t are an M-row vector and a K-row vector respectively. It is implied that we have M simultaneous relationships, in the manner of (1), with $Y_t \cap Y_{ti} \neq \psi$ but $y_{ti} \notin X_{ti}$ all $i = 1 \ldots M$[1] and $Z_t \cap X_{ti} \neq \psi$ some $i = 1 \ldots M$, summarised as

$$Y_t = F(X_t, U_t) \quad t = 1 \ldots T \tag{2}$$

where ψ denotes the empty set. $F(\cdot,\cdot)$ is any M x (K + M) matrix valued functional independent of t, possessing continuous nonvanishing derivatives of at least first order, with rows $f_i(\cdot,\cdot)$. The set of dependent variables Y_t has i^{th} element y_{ti} whereas the set of predetermined variables has i^{th} element z_{ti}. u_{ti} is the i^{th} element, corresponding to the i^{th} equation, of the M-row vector of unobserved stochastic disturbances representing the effect on Y_t, of influences otherwise unexplained in (2). We partition $X_{ti} = [Y_{ti} : Z_{ti}]$, and correspondingly $X_t = [Y_t : Z_t]$. Let Y_{ti} and Z_{ti} be r_i and s_i row vectors respectively.

The relationships (2) may be theoretical or hypothesised characterisations of the working of some aspects of the economy to be modelled. Thus they will be investigated with a sample of data, size T, from the T-dimensional Euclidean space E_T - the observation space on which (2) is defined. Since Y_t are completely determined within (2) by Z_t, U_t, we may define them as endogenous variables whose distribution is completely specified by given values Z_t and the distribution of U_t. It follows the variables Z_t may be of two kinds - lagged endogenous and exogenous variables. The latter are determined completely outside the system and may or may not be stochastic;

---

1. The assumption is therefore are no relationships of the form $y_i = f_i(Z_{ti}, U_{ti})$ which is justified in the discussion of recursivity.

we assume for such variables $E(Z_t' U_s) = 0$ all $t,s = 1 \ldots T$. The former, by our definition of endogenity, preserve the concept of contemporaneous independence. We assert plim $T^{-1}(Z_t' U_s) = 0$ $t,s = 1 \ldots T$ because $E(Z_{t-j}' U_t) = 0$ all $j = 1 \ldots t-1$; $t = 1 \ldots T$ (Mann and Wald 1943). Therefore the general property

$$\text{plim } T^{-1} (Z_t' U_s) = 0 \quad t,s = 1 \ldots T \text{ for all } Z_t \tag{3}$$

is used throughout without further distinguishing between the two sorts of predetermined variable. Finally, (2) contains an arbitrary normalisation in each equation. It is argued that some normalisation will naturally emerge in the modelling of each economic process; that is in the selection of each y_{ti}. In any case a normalisation is always possible, if arbitrary. We continue with it given and arbitrary[1] but make some references to altering it.

These definitions of variables agree with Tinbergen's (1952) usage for the theory of economic policy. Targets and irrelevant variables are endogenous, whereas instruments and data variables are predetermined. Equations may be classified in the same manner. Behavioural (and technical) relations between stochastic are as described by (2). Identities and definitional are relations in which $u_{ti} \equiv 0$, $t = 1 \ldots T$ and $f_i(\cdot,\cdot)$ are known precisely a priori. Let there be G behavioural relations whose parameters and functional form $f_i(\cdot,\cdot)$ are to be estimated. Until further notice (Chapter 13 when identities are handled) we assume G = M.

For most of the analysis we specialise specification (2) to a relation linear in the parameters with additive disturbances;

---

1. and could therefore be changed, but not without consequences for the distribution and properties of U_t. The reader is referred to Mosbaek & Wold (1970) for symmetric and asymmetric model specifications. The general specification would be $F(X_t,U_t) = 0$.

$$Y_t\beta + Z_t\Gamma = X_tD = U_t \quad t = 1 \ldots T \tag{4}$$

or summarised for the whole sample

$$Y\beta + Z\Gamma = XD = U \tag{5}$$

This assumption will apply, for simplicity, except where it is relaxed (in Chapter 15) to the general form (2) where problems of nonlinearity are the focus of attention. (5) does not imply a linear specification for it allows nonlinearity[1] in those elements of Z_t which are functions of original variables. Here Y is a T x M matrix of observations on endogenous variables, with rows Y_t; Z is a T x K matrix of observations on predetermined variables, with rows Z_t; $X = [Y \vdots Z]$; U is a T x M matrix of stochastic disturbances with rows U_t. β and Γ are M x M and K x M matrices of unknown parameters to be estimated and $D' = [\beta' \vdots \Gamma']$. However some parameters are restricted to be zero by the a priori knowledge used in specifying the model. A second notation is convenient. From (5) we obtain the i^{th} equation as

$$y_i = Y_i\beta_i + Z_i\gamma_i + u_i = X_i\delta_i + u_i \quad i = 1 \ldots M \tag{6}$$

where y_i is a T-column vector and Y_i, Z_i, X_i are T x r_i, T x s_i and T x (r_i+s_i) matrices. Noting $\delta_i' = (\beta_i' \vdots \gamma_i')$ is the $r_i + s_i$ row vector of coefficients, a priori nonzero, to be estimated in the specification of the i^{th} equation we have

---

1. The problem of nonlinear endogenous variables is treated in Chapter 15.

$$\left.\begin{array}{l} D_{ji} = -\beta_{ji} \quad \text{if } y_j \in X_i \\ \qquad \text{and } j \neq i \\ \quad = 1 \qquad \text{if } i = j \\ \\ \text{and } D_{k+M,i} = -\gamma_{ki} \text{ if } z_k \in X_i \\ \\ \text{but } D_{j,i} \text{ and } D_{k+M,i} = 0 \quad \text{otherwise} \end{array}\right\} \; i,j = 1 \ldots M; \quad k = 1 \ldots K \qquad (7)$$

We assume every equation of (2) and (5) to be identified; that is both rank and order conditions are satisfied. From the latter, therefore, $q_i \geq 0 \quad i = 1 \ldots M$ where $K - s_i = r_i - 1 + q_i$.

Either specification (2) or (5) is called the structural form. If (5) holds a restricted reduced form

$$Y = Y^{*} + V \qquad (8)$$

exists where $Y^{*} = Z\Pi$, and $\Pi = -\Gamma\beta^{-1}$ with $V = U\beta^{-1}$, provided β is nonsingular. We distinguish the unrestricted reduced form obtained from the projection of Y on the Z space, i.e. $Y^{+} = Z\Pi^{+} = Z(Z'Z)^{-1}Z'Y$ where

$$Y = Y^{+} + V^{+} \qquad (9)$$

2.2.2 The assumptions summarised:

In order to be able to study the properties of an estimator applied to (5), we must make certain assumptions about that system. The following assumptions are made, being those almost always used in econometric texts (e.g. Dhrymes (1970)). Several amendments to this list are possible for many of the estimators considered and these possibilities are indicated.

Let (8), at time t, be rewritten with an obvious partitioning of Z_t

$$Y_t = \sum_{j=1}^{\ell} Y_{t-j}\,\Pi_j + Z_t^*\,\Pi_{\ell+1} + V_t \qquad (10)$$

$$\text{given } (\Pi_j)_{ik} = \Pi_{ik} \text{ if } y_{t-j,i} = z_{tk} \quad \left.\begin{array}{l} i = 1 \ldots M \\ k = 1 \ldots K \\ j = 1 \ldots \ell; \end{array}\right.$$

$$0 \text{ otherwise}$$

becoming

$$Y_{(\ell)t} = Y_{(\ell)t-1}\,\Pi_1^* + Z_t^*\,\Pi_2^* + (V_t\ 0 \ldots 0) \qquad (11)$$

if $Y'_{(\ell)t} = \begin{pmatrix} Y_t' \\ \vdots \\ Y'_{t-\ell+1} \end{pmatrix}$; $\Pi_1^* = \begin{pmatrix} \Pi_1 I_M & & & \\ \vdots & \ddots & & 0 \\ \vdots & & \ddots & \\ \Pi_\ell & 0 & & I_M \end{pmatrix}$; $\Pi_2^* = (\Pi_{\ell+1}\ 0 \ldots 0)$;

where ℓ is the maximal lag in lagged endogenous variables and Z_t^* contains all K_1 current and lagged exogenous components from Z_t. We assume <u>at most</u>:

(a) $E(U) = 0$

(b) $E(u_i u_i') = \sigma_i^2 I \quad i = 1 \ldots M$

(c) $E(U'U) = \Sigma$, a symmetric, finite, positive definite matrix.

(d) $|x_{ti}| < \chi^*$ some finite bounding constant, $i = 1 \ldots M + K$; $t = 1 \ldots T$

(e) $E(z_{ti} U_s) = 0$ all $t,s = 1 \ldots T$; $\quad i = 1 \ldots K_1$ and

$E(y_{t-j,i} U_t) = 0$ all $j = 1 \ldots t - 1$; $\quad t = 1 \ldots T$; $\quad i = 1 \ldots M$.

(f) Stability: $\rho(\Pi_1^*) < 1$; $\rho(\cdot)$ denotes the spectral radius of that matrix.

(g) These are no linear dependencies in Z or X_i any $i = 1 \ldots M$.

(h) In (5) β is nonsingular; i.e. no equation is redundant, or any that are have been discarded.

(i) $q_i \geq 0$ $i = 1 \ldots M$ and assumptions on specification at (2).

(j) The structural errors are mutually identically independently distributed $U \sim N(0, \Sigma \otimes I)$.

2.2.3 Relaxation of Assumptions:

However often we can assume less than (a) - (j). Assumptions (a), (c), (d), (f) and (i) are never altered. Problems of serially correlated errors are treated in many places, and an autoregressive version of each estimator analysed is provided. In those cases assumption (b) is never adhered to, but $E(u_i u_i') = \Sigma_i$ $i = 1 \ldots M$ is taken; Σ_i symmetric positive definite. Then any results given for any nonautoregressive estimator are valid only within that implied information class (see Section 4 below) - analogous to properties quoted in any text book for limited information estimators within their information class when full information conditions could apply. Since no other exposition follows this idea of properties within an information class for the autoregressive case, results here are, instead, worded as if (b) held; and then reworked for the autoregressive estimator as if it did not, first assuming in these cases no lagged dependent variables in Z_t. This tactic has the distinct advantage that we do not attempt to handle the simultaneous presence of lagged dependent variables and serially correlated errors within the general model. The general analysis then allows $y_{t-j,i} \in Z_t$.

Assumption (e) is replaced by the weaker plim $T^{-1}(Z'U) = 0$ throughout, as mentioned above. In Chapter 6 (GEID systems) it is further weakened to either plim $T^{-1}(Y^*_i{}' u_i) =$ plim $T^{-1}(Z_i' u_i) = 0$ $i = 1 \ldots M$, or the same with v_i replacing u_i.

Assumption (g) may, for all iterative estimators, be replaced by no linear dependencies in Z_i $i = 1 \ldots M$. It is retained as it is for any estimator which makes use of Π^+; it is strengthened to the requirement of no linear dependences in X when the existence of a likelihood function is postulated.

Assumption (h) is not applicable for the form (2). Neither the equivalent requirement that (F'(Y*)), where F'(·) denotes the matrix of first order partial derivatives (here evaluated at Y*), nor β be non-singular is actually necessary for many iterative estimators. However the condition of no redundant equations is such an obvious one for a sensibly specified model, so it is retained.

Assumption (j) is only used when a likelihood function is postulated and therefore only in the derivation of maximum likelihood estimators. For no other estimator is it required. It is required for estimators shown to be equivalent (asymptotically or not) to maximum likelihood ones in so far as the equivalence is only valid if the maximum likelihood estimator may be derived from the likelihood function, but not in the sense of an approximation to it in the absence of the likelihood function. Non-maximum likelihood estimators are in no way affected by the removal of assumption (j), and maximum likelihood estimators could be obtained if other distributions were assumed; but in either case these equivalences would be invalid.

It is clear that the distribution in assumption (j) is not required to derive the asymptotic distributions of conventional test statistics for the estimated model. This follows from Theorems A2.5, A2.6 and A2.7 whose regularity conditions are satisfied by the model and assumptions here.

2.2.4 Comments

(i) Assumptions (a) - (j) form two groups. One group ((a), (b), (c), (e) and (j)) provide the statistical properties of an estimator. The other group ensures the existence of the estimators considered. It is necessary to bear in mind in what follows the variety of relaxations that are appropriate for each estimator. In particular we make no assumption on the sample size T apart from it being greater than the minimum required for the existence of the estimator. For the model (6) i = 1 .... M, OLS requires $\max_i r(X_i) = \max_i (r_i + s_i) < T$

whereas it will turn out that for 2SLS we need $r(Z) = K < T$; for 3SLS $\max (r(Z) = K,M) < T$; for LIML and FIML the existence of a likelihood function needs $K + M < T$, but for computations $K < T$ or $K,M < T$; for IIV and FP family members $\max_i (r_i + s_i) < T$ or $\max_i (r_i, s_i) < T$ according to the algorithm. Therefore the degrees of freedom used by each estimator are stated and no assumption is made on T. Those observations are very similar to the possible relaxations in assumption (g). The two sets of requirements should in fact be treated together.

(ii) Assumption (f) (stationarity of Y_t) is a subset of assumption (d), but there appears to be confusion as to the purpose of it. The proof that the conditions in assumption (e) imply plim $T^{-1}(Z'U) = 0$ is well known (see Theorem 1, p.547 Malinvaud (1970)) and is usually achieved in conjunction with stationarity. However nonstationarity does not imply inconsistency.[1] Rewriting (11) we have

$$Y_{(\ell)t} = \sum_{k=0}^{\infty} Z^*_{t-k}\, \Pi_2^*(\Pi_1^*)^k + \sum_{k=0}^{\infty} V_{t-k}(\Pi_1^*)^k \qquad (12)$$

which, given the shape of Π_2^*, implies, if $\rho(\Pi_1^*) > 1$ and $|z_{tj}| < \chi^*$ $j = 1 \ldots K$, $|y_{ti}| \to \infty$ as $T,k \to \infty$, $i = 1 \ldots M$. The consequence is that our estimators cease to be defined. OLS on (6) implies

$$\hat{\delta}_i = (X_i'X_i)^{-1} X_i'y_i \quad i = 1 \ldots M \qquad (13)$$

so that in this case the inverse moment matrix cannot exist if $T \to \infty$, because $Y_t \cap X_{ti} \neq \psi$. It will be shown that all the other estimators mentioned in this study replace the moment matrix by either

---

1. See Sargan (1958) for proof in the case of a nonsimultaneous equation model. Naturally if we can solve the problem of consistent estimation of the coefficients β_i $i = 1 \ldots M$, then his result will carry over for (5).

$X_i^{*\prime}X_i^*$; $X_i^{*\prime}X_i$; $\hat{X}^{*\prime}(\Sigma \bar{(x)} I)^{-1}\hat{X}^*$; or $\hat{X}^{*\prime}(\Sigma \bar{(x)} I)^{-1}\hat{X}$, where $X^* = [Y^* : Z]$ and $\hat{X} = \text{diag}(X_i)$. Because $Y^*_{(\ell)t} = \sum_{k=0}^{\infty} Z_{t-k}\Pi_2^*(\Pi_1^*)^k$, then these estimators are also undefined. If however assumption (d) and hence by implication assumption (f), holds then $\left|\frac{\Sigma_t x_{ti} x_{tj}}{T}\right| \leq \chi^{*2}$ (or the equivalent for the other estimators) and the existence of the estimators is assured as $T \to \infty$, and we may then talk of asymptotic properties.[1]

2.3 The conventional estimators:

The conventional estimators for (5) are defined as those traditionally used; that is OLS, 2SLS, 3SLS, LIML or FIML. The list is far from covering all methods proposed but others, including variants of those mentioned, will be treated as unconventional estimators and most analysed here. The distinction is that the statistical and computational properties of the conventional estimators, as well as some (not all) of the relationships between them, are well known. However those for the unconventional are not well known; nor is their relationship to the conventional estimators. We summarise those properties of conventional estimators.

We need a theorem due to Wold (1963). Let the sample moments be

$m_k = \sum_{t=1}^{T} x_{tk}$, $m_{kj} = \sum_{t=1}^{T} x_{tk}x_{tj}$ and the corresponding theoretical ones be

$\mu_k = E(x_k)$, $\mu_{kj} = E(x_k'x_j)$.

Theorem 1:

OLS regression on the i^{th} equation of (4), by minimising $||u_i||_e^2$ where $||\cdot||_e$ denotes the Euclidean norm, yields consistent estimates of β_i and γ_i iff

---

1. if $\rho(\Pi_1^*) > 1$ Anderson (1959) has shown that the moments of the distribution of the estimators do not exist, although their t statistics are distributed as usual.

(a) $y_i = X_i\delta_i + u_i$ is a predictor specification; $E(y_i|X_i) = X_i\delta_i$ $i = 1....M.$

(b) Plim T^{-1} $m_k = \mu_k$ and plim T^{-1} $m_{kj} = \mu_{kj}$, $k,j = i,1 .... r_i+s_i, i=1....M$

where x_k, $x_j \varepsilon [X_i \vdots y_i]$ (ergodicity).

(c) Det $\{\mu_{kj}\} \neq 0$, $k,j = 1 .... r_i + s_i$.

Proof:

Wold (1963). #

Comment:

Wold's proof shows the predictor specification to be a necessary condition whereas ergodicity and (c) are sufficient.

2.3.1 Non Simultaneous Estimation:

The OLS estimator for (5) is defined by (13). Since $Y \cap X_i \neq \psi$ the predictor specification mentioned fails, $E(y_i|X_i) \neq X_i\delta_i$ $i = 1 .... M.$ The violation of a necessary condition implies inconsistent estimates. However as Fisher (1965) points out if β is lower triangular, Σ is a diagonal matrix and $E(u_iu_i') = \sigma_i^2 I$ $i = 1 .... M$, or if β is lower triangular, Π_j and $E(UU_\theta')$ strictly[1] lower triangular where $U_\theta = \{U_{t-\theta}\}$ all $\theta \geq 0$ and $j = 1 .... \ell$, then OLS will be consistent. However we have made no such assumptions. Alternatively if direct ($\beta_{ij} \neq 0$ some $j \neq i$) and indirect ($\sigma_{ij} \neq 0$ some $i \neq j$) simultaneity are weak, we could appeal to Wold's proximity theorem (Fisher 1965) to claim "near" consistency. A third case is noted in Chapter 6.8 below which depends on a precise a priori knowledge of the parameters of the system. This knowledge is not available a priori so we are not justified in assuming it. Estimates are therefore inconsistent.

---

1. i.e. $(\Pi_j)_{kq} = 0$ if $q \geq k$ as opposed to $q > k$.

As a digression, suppose $\beta_i = 0 \quad i = 1 \ldots M$. Then

Theorem 2: (Gauss-Markov)

OLS estimates are (i) unbiased, (ii) consistent, (iii) minimum variance among linear unbiased estimators (efficient) if Σ diagonal. Then $V(\hat{\delta}_i) = \sigma_i^2 (X_i'X_i)^{-1}$ if X_i is fixed in repeated samples.

Proof:

Dhrymes (1970) Theorem 4.1. #

Theorem 3:

If $u_i \sim N(0, \sigma_i^2 I)$ then OLS and maximum likelihood estimates coincide.

Proof:

Dhrymes (1970) Theorem 4.2. #

Theorem 4:

If $E(u_i u_i') = \Sigma_i$ the efficient estimator is its autoregressive version (GLS)

$$\hat{\delta}_i = (X_i'\Sigma_i^{-1} X_i)^{-1} X_i' \Sigma_i^{-1} y_i \quad i = 1 \ldots M \tag{14}$$

with $V(\hat{\delta}_i) = (X_i'\Sigma_i^{-1} X_i)^{-1}$ if X_i are fixed.

Proof:

Dhrymes (1970) Theorem 4.3. #

Theorem 5:

If $E(U'U) = \Sigma$ the efficient estimator is its full information version

$$\hat{\delta} = (\hat{X}' \phi^{-1} \hat{X})^{-1} \hat{X}' \phi^{-1} y \tag{15}$$

where $\delta = \text{vec}\ (\delta_i)$; $\hat{X} = \text{diag}\ (X_i)$; $y = \text{vec}\ (y_i)$ and $\phi = \Sigma \otimes I$; $\text{var}\ (\hat{\delta}) = (\hat{X}'\ \phi^{-1}\ \hat{X})^{-1}$ if $\hat{X}$ is fixed (more usually known as seemingly unrelated regressions - Zellner (1962)).

Proof:

Dhrymes (1970) Theorem 4.4. #

In fact efficiency in (14) and (15) follows directly by Lemma A2.3 since Dhrymes (1970) Theorem 4.2 shows in their respective circumstances they are maximum likelihood estimators. Consequently the variance formulae show the respective Cramer-Rao lower bounds.

We retain the new notation of Theorem 5 together with (4) and (6).

Remarks:

Identities do not enter the computations. In view of the moment matrices in (13), the degrees of freedom used are $T - (r_i + s_i)$ $i = 1 \ldots M$ (Dhrymes (1970) Theorem 4.2) and the estimator exists if $T > \max_i (r_i + s_i)$.

Optimality:

The use of OLS may be justified in terms of its optimality; that is if $\hat{y}_t = g(z_i)$ for some estimation procedure $g(\cdot)$ then OLS minimises $\sum_{t=1}^{T} (y_t - \hat{y}_t)^2$. More generally for a quadratic loss function

$$L = (\hat{y} - y)'\ W(\hat{y} - y) \tag{16}$$

and $W = H'H$ is symmetric and positive definite we have the risk

$$R = \sum_{t=1}^{T} E[h_t'(\hat{y} - E(\hat{y}))]^2 + \sum_{t=1}^{T} [h_t'(E(\hat{y}) - y)]^2 \tag{17}$$

h_t is t^{th} row of H.

Then (see Malinvaud (1970) Chapter 5.9) the GLS estimator is minimax and admissable in the class of linear estimators[1]; if further Z_t is normally distributed it is admissable in the class of regular unbiased estimators. Clearly it is the minimum distance estimator for distance function (16). All the estimators considered here have optimality in some sense, and when nonlinear systems (2) are considered we continue with minimum distance estimators, choosing a suitable W, to maintain such optimality.

We now revert to the original model specification (16) and (5).

An alternative to (13) is instrumental variable estimation. In instruments N_i $i = 1 \ldots M$ are used we have

$$\hat{\delta}_i = (N_i' X_i)^{-1} N_i' y_i \quad i = 1 \ldots M \tag{18}$$

Provided

$$\left.\begin{array}{ll} \text{(a)} & 0 < \text{plim } T^{-1}(N_i'X_i) < \infty \\ \text{(b)} & \text{plim } T^{-1}(N_i'u_i) = 0 \end{array}\right\} \quad i = 1 \ldots M \tag{19}$$

dictates the choice of N_i then,

Theorem 6:

The estimates (18) are consistent and imply, asymptotically, $\sqrt{T}(\hat{\delta}_i - \delta_i)$

$$\sim N(0, \sigma_i^2 \text{ plim } T^{-1} (N_i' X_i)^{-1} \; N_i' N_i \; (X_i' N_i)^{-1}) \tag{20}$$

Proof:

Dhrymes (1970) Chapter 6.4 and Theorem A2.3. #

Again we create autoregressive and full information analogues of (18) corresponding to (14) and (15) by using instruments $\Sigma_i^{-1} N_i$ and $\phi^{-1}\hat{N}$, where $\hat{N} = \text{diag } (N_i)$, respectively (amending therefore the variance in (20) at the same time).

---

1. Minimax minimises the maximum risk; admissable implies it is unique in doing so.

Comment:

The identities do not enter the calculations; the existence of (18) and degrees of freedom used are identical to the OLS case.

Remark:

We might plan to overcome the problem of inconsistent OLS estimates by a judicious choice of N_i $i = 1 \ldots M$ such that (19) holds. The problem so far is a lack of optimality - we have no criterion to select the optimal instruments from the many that satisfy (19). In view of (20) any such set N_i implies consistency, so we take that which minimises the asymptotic variance. An alternative plan would minimise the small sample bias in (18) which comes from $N_i \neq X_i$. However the exact sampling distribution of $(\hat{\delta}_i - \delta)$ is not known - and its approximation (Sargan and Mikail (1971)) is sufficiently complicated to render this plan infeasible. Indeed for many cases of interest the first moment of this distribution will not even exist (see Chapter 16). Dhrymes shows that the asymptotic genralised variance in (20) may be written

$$|\text{cov}\ (\hat{\delta}_i)| = \frac{\sigma_i^{2s_i}}{|X_i' X_i|} \cdot \frac{1}{r_{(X_i,N_i)}} \qquad (21)$$

where $r_{(X_i,N_i)}$ is the sample coefficient of vector correlation between X_i and N_i. Therefore since Y_i not Z_i caused the failure of the predictor specification, Z_i should be contained in N_i and only Y_i replaced.

Remark:

Not only does OLS fail to have any desirable asymptotic statistical properties in the estimation of (5), but also, since X_i then has a distribution, all the variance formulae in Theorems 2, 4, 5 are invalid. Although an expression for the exact variance of the estimate is available for some special cases of (5), it cannot be used in any general estimation theory (e.g. Sawa 1972). Any such possible extensions to the estimation theory derived from asymptotic properties are considered in Chapter 16, together with the possibilities of comparing estimators on their small sample properties.

2.3.2 Simultaneous Estimation:

The 2SLS estimator of (5) is defined by

$$\hat{\delta}_i = [X_i'Z(Z'Z)^{-1}Z'X_i]^{-1}X_i'Z(Z'Z)^{-1}Z'y_i \quad i = 1 \ldots M \tag{22}$$

Then

Theorem 7:

Estimates (22) are (i) consistent (ii) asymptotically efficient,[1]

$$\text{(iii)} \quad \sqrt{T}(\hat{\delta}_i - \delta) \overset{a}{\sim} N(0, \sigma_{ii} \text{ plim } T^{-1}[X_i'Z(Z'Z)^{-1}Z'X_i]^{-1}) \tag{23}$$

if Σ is diagonal with i^{th} element σ_{ii}. Furthermore if $U_i \sim N(0, \sigma_{ii}I)$ then (22) and the LIML estimator ((32) below) are asymptotically equivalent.

Proof:

Dhrymes (1970) Theorem 4.11, proposition 7.3 with Lemma A2.3. #

Theorem 8:

Estimator (22) may be rewritten as

$$\hat{\delta}_i = (N_i'X_i)^{-1}N_i'y_i \quad i = 1 \ldots M \tag{24}$$

where $N_i = X_i^+ = [Y_i^+ : Z_i]$.

Proof:

Expand the matrix and vector in (23) with the partitioning in X_i. If we replace

$$(Z'Z)^{-1}Z'Z_i = \begin{pmatrix} I \\ \cdots \\ 0 \end{pmatrix} \tag{25}$$

---

1. This implies Var $(\hat{\delta}_i) \to$ Cramer-Rao minimum variance bound, as $T \to \infty$, all i. See Lemma A2.2.

wherever it occurs, we have (24) expanded with the partitioning in both N_i and X_i, using the notation of (9). #

Corollary:

(23) may also be rewirtten as

$$\hat{\delta}_i = (X_i^{+\prime} X_i^{+})^{-1} X_i^{+\prime} y_i \quad i = 1 \ldots M \tag{26}$$

since $X_i^{+\prime} X_i = X_i^{+\prime} X_i^{+} + X_i^{+\prime} (V_i^{+} : 0)$ and from (9) $V_i^{+} = [I - Z(Z'Z)^{-1}Z']V_i$; hence $X_i^{+\prime} V_i^{+} = 0$ in view of $Y_i = Z\Pi_i$ and (25). #

Remark:

It is well known that 2SLS may be treated equally as an OLS estimator (26) or an IV estimator (24). We proceed with the form (26) although either could be used. From the corollary, (20) reduces to

$$\sqrt{T}(\hat{\delta}_i - \delta_i) \overset{a}{\sim} N(0, \sigma_{ii} \text{ plim } T^{-1}(X_i^{+\prime} X_i^{+})^{-1}) \quad \# \tag{26a}$$

Theorem 9:

If $E(u_i u_i') = \Sigma_i$, the efficient and consistent estimator is its autoregressive form

$$\hat{\delta}_i = (X_i^{+\prime} \Sigma_i^{-1} X_i^{+})^{-1} X_i^{+\prime} \Sigma_i^{-1} y_i \quad i = 1 \ldots M \tag{27}$$

Proof:

Dhrymes (1970) Theorems 4.3 and 4.7 and (26), #

Theorem 10:

If $E(U'U) = \Sigma$ the efficient and consistent estimator is its full information form

$$\hat{\delta} = (\hat{X}^{+\prime}\ \phi^{-1}\ \hat{X}^{+})^{-1}\ \hat{X}^{+\prime}\ \phi^{-1}\ y \qquad (28)$$

where $\hat{X}^{+} = \text{diag}\ (X_i^{+})$.

Proof:

Dhrymes (1970) Theorems 4.4 and 4.7 and (26). #

Theorem 11:

Asymptotic variances of $\hat{\delta}_i$ and $\hat{\delta}$ from (27) and (28) are given by plim $T^{-1}\ X_i^{+\prime}\ \Sigma_i^{-1}\ X_i$ and plim $T^{-1}\ \hat{X}^{+\prime}\ \phi^{-1}\ \hat{X}$ respectively and moreover both may be treated as IV estimators, asymptotically normally distributed.

Proof:

For (27) $(X_i^{+\prime}\ \Sigma_i^{-1}\ X_i^{+})^{-1} = X_i^{+\prime}\ \Sigma_i^{-1}\ X_i$ in the manner of the corollary to Theorem 8, considering all i = 1 .... M by theorem 4.3 Dhrymes (1970). For (28) $\hat{X}^{+\prime}\ \phi^{-1}\hat{X}^{+} = \hat{X}^{+\prime}\ \phi^{-1}\hat{X}$ since $\hat{X} = (I \otimes X)S$ where $S = \text{diag}\ (S_i)$ and $X_i = XS_i$. Then

$$\hat{X}^{+\prime}\ \phi^{-1}\hat{X}^{+} = S'(\Sigma^{-1} \otimes X^{+\prime}X^{+})S = S'(\Sigma^{-1} \otimes X^{+\prime}X)S \qquad (28a)$$

by corollary to Theorem 8 and the result follows by (20), (26a) (28a). #

Corollary

(28), which is 3SLS, may be treated either as a GLS or (Generalised) IV estimator. It is also asymptotically equivalent to FIML if $U \sim N(0,\phi)$ (Dhrymes (1970) remark (i) p.371). (26) and (28) are consistent, asymptotically normal and asymptotically efficient with follows from Theorems 11, 12 and 6 with (19) (or Theorems 4, 5, A2.5 and Lemma A2.3). #

Comments:

(a) Consistent and asymptotic efficient estimates result from 2SLS or 3SLS applied to individual subsystems, of a block recursive model, in isolation. A block recursive model is defined as (5) where β is lower block diagonal and either Σ block diagonal, or Π_j and $E(UU'_\theta)$ strictly lower block diagonal for

$\theta \gtrless 0$, $j = 1 \ldots \ell$ (Fisher 1965). No such assumptions apply, however, to (5).

(b) 3SLS dominates 2SLS in the sense of minimum asymptotic variance of estimates except where (i) the i^{th} equation is just identified, (ii) $\sigma_{ij} = 0$ $i \neq j = 1 \ldots M$, for which equations 2SLS $\equiv$ 3SLS (Dhrymes (1970) Theorem 4.13).

(c) As a consequence of (b) for 3SLS, just identified equations or those with $\sigma_{ij} = 0$ $i \neq j = 1 \ldots M$ should be estimated as a separate subsystem, by 2SLS. Identities do not enter the estimation process of either 2SLS or 3SLS.

(d) The unused degrees of freedom for 2SLS or 3SLS are of no interest for hypothesis testing in the absence of small sample distributions. However, from (8) $\hat{u}_i = V^+\hat{\beta}_{(i)}$ where $\hat{\beta}_{(i)}$ is the i^{th} column from $\hat{\beta}$. Therefore $\hat{u}_i' \hat{u}_i = \theta_i' \begin{pmatrix} V'QV & 0 \\ 0 & 0 \end{pmatrix} \theta_i$ where θ_i is a vector with non-zero elements $\delta_i + (X_i^{+\prime} X_i^+)^{-1} X_i^{+\prime} u_i$ but zeros and unity to match $\hat{\beta}_{(i)}$ and $\hat{\Gamma}_{(i)}$; where $Q = I - Z(Z'Z)^{-1}Z'$ an idempotent matrix.[1] Given an orthogonal matrix P

$$\hat{u}_i' \hat{u}_i = (\theta_{1i} VP)' \begin{pmatrix} I_{T-K} & 0 \\ 0 & 0 \end{pmatrix} (PV\theta_{1i}) = \sum_{j=1}^{T-K} \xi_j^2, \text{ for } \theta_i' = (\theta_{1i}' : \theta_{2i}').$$

Therefore 2SLS and 3SLS have T - K degrees of freedom. Clearly the existence of (22) requires T > K whereas (28) requires T > max (K,M).

Optimality:

There are several possible interpretations of 2SLS (and 3SLS). By (9) $Y_i = Y_i^+ + V_i^+$, the simplest explanation is that Y_i is purged of its stochastic part which causes the failure of the predictor specification. The deterministic part used in its place in OLS on

$$y_i = X_i^+ \delta_i + u_i + V_i^+ \beta_i \quad \text{so} \quad E(y_i \mid X_i^+) = X_i^+ \delta_i \qquad (29)$$

---

1. Q is identical for 3SLS since (15) $\equiv$ (13) when every equation has the same explanitory variable set which occurs in (9) (Zellner (1962)), but q_i must be amended to reflect (28).

Unless Σ is diagonal any u_i $i = 1 \ldots M$ must be purged from Y_i^+. Therefore 2SLS has the following optimality properties as a least squares estimator:

(a) it minimises the sum of squares $\sum_t W_{ti} = \sum_t (u_{ti} + V_{ti}^+ \beta_i)^2$ of the transformed system (29); it therefore minimises the residual variance of (6) as a whole, where some is contained within Y_i. W is that matrix in (16).

(b) For the transformed system (29) it is a minimax estimator admissable in the class of linear estimators. Further since expectation is a linear operator it is also therefore minimax and admissable in the class of linear estimators for (6).

(c) It minimises the differences of the residual variances; for (4) less that of y_i regressed on Y_i,Z (Johnston (1972)).

Clearly Y^+ is an estimate of Y^*. It is tempting to suggest that 2SLS is a minimum distance or maximum likelihood estimator.[1] However what we need is either form of estimator for either (6), or

$$y_i = Y_i^* \beta_i + Z\gamma_i + (u_i + V_i\beta_i) \quad i = 1 \ldots M \tag{30}$$

where all the stochastic disturbances are collected together. If Y* can be found, we may obtain either by OLS on (30). This is the starting point of our investigation of iterative estimators. (30) is Wold's reformulated interdependent (REID) system. REID is motivated by three other ideas. If we wish to retain the predictor specification so that OLS can be consistently used, and (5) is the correct specification, then Y_i must be replaced by a linear combination of Z. In view of (8) the minimum change is to transform the system to (30). Then OLS on (30) gives BLU estimates of the parameters of (2.5) since theorem 1 holds by assumption (e), and

$$E(y_i \mid X_i^*) = X_i^* \delta_i = E(y_i \mid Z) \quad i = 1 \ldots M \tag{31}$$

---

1. It is trivially both for system (29)

is the predictor specification. Secondly, then the firstterm of (17) vanishes, which reduces the risk function provided h_t is not increased too much. Thirdly it is the minimum distance estimator for (16) where the true W with $W_{ij} = (u_i + V_i\beta_i)'(u_j + V_j\beta_j)$, being unknown is approximated by continual reevaluations simultaneously with parameter estimation. A single estimate as in 2SLS is not used.

2SLS and 3SLS as IV estimators have all of the above optimality properties. However they have their own optimality property from (21). The first condition of (19) implies that N_i must be functions of variables in the system, if the system is correctly specified to include all the expalantory variables. The second condition implies Y_i are barred as an argument for N_i.

Theorem 12:

Among the class of all linear combinations of Z, any $N_i = [ZP_i \vdots Z_i]$, where plim $P_i = \Pi_i$, or $\phi^{-1}\hat{N}$ for the full information analogue, $\hat{N} = \text{diag } N_i$, produces consistent and asymptotically efficient estimates by (18).

Proof:

Because (19) holds, then remarks following (21) and Dhrymes (1970) Theorem[1] 6.2. Efficiency follows (19), (20) and Lemma A2.3. #

Naturally therefore 2SLS and 3SLS are optimal since their instruments satisfy (19) and Theorem 12; but there are many other equally good instrument sets, including[2] $N_i = X_i^*$. This is another starting point for our investigation of iterative estimators.

If there are more potential instruments than variables to be replaced (as here $K > r_i$ all i) then it is optimal in the sense of asymptotic efficiency to choose those linear combinations of all possible instruments. These linear combinations are given by Theorem 12 (see Malinvaud Chapter 20.5 and Sargan

---

1. Actually adapted to note (21) is valid only asymptotically so we have to take the probability limit of his result - something he fails to do.
2. Which would overcome the dangers of collinearity and degrees of freedom loss in Π^+.

(1958)). No further improvements are possible by the inclusion of more instruments.

It should be noted that Theorem 12 is actually a statement of two major theorems of Brundy and Jorgenson (1971) and is offered as an alternative proof of them.

2.3.3 Maximum likelihood methods:

For (5) they are relatively less developed. The LIML estimator is given as the eigenvector (normalised to have $\hat{\beta}_{ii} = -1$) associated with the minimal eigenvalue λ in the eigenproblem

$$|\lambda W_{ii} - W_{ii}^{*}| = 0 \qquad (32)$$

Here $W_{ii} = Y_i'[I - Z(Z'Z)^{-1}Z']Y_i$, and

$$W_{ii}^{*} = Y_i'[I - Z_i(Z_i'Z_i)^{-1}Z_i']Y_i \qquad (33)$$

Then $$\hat{\gamma}_i = -(Z_i'Z_i)^{-1}Z_i'(y - Y_i\hat{\beta}_i) \qquad (34)$$

Theorem 13:

LIML estimates are consistent and asymptotically efficient when ϕ is diagonal.

Proof:

Theorem A2.7 (Also Theorem 7). #

Their asymptotic equivalance to 2SLS was mentioned and no other relationships appear to be known at this stage. Identities do not enter the computations. The degrees of freedom unused appear to be T - K; in any case the estimator exists only if T > K by (33). Asymptotic variance formulae for estimates may

be presented in several forms but those given for 2SLS will suffice. An auto-regressive version does not appear to have been suggested, although the full information version FIML is well known.

Under assumption (j) model (5) has the associated log likelihood function,[1]

$$L(\beta,\Gamma,\Sigma) = C + {}^{T}/_{2} \log \det \Sigma^{-1} + T \log |\det \beta| - {}^{T}/_{2} \operatorname{tr}[T^{-1}\Sigma^{-1}D'X'XD] \qquad (35)$$

From (35) under the restriction Σ is diagonal, and with a suitable transformation, we obtain the single equation LIML estimator (Dhrymes (1970), Theorem 7.2). For FIML we assume no restrictions of Σ and only that (5) is identified and proceed directly. Conventionally (35) is concentrated with respect to Σ yielding,

$$\hat{\Sigma} = D'X'XD \qquad (36)$$

which implies we chose $\hat{\beta},\hat{\Gamma}$ to maximise

$$L_C(\beta,\Gamma) = C^* + {}^{T}/_{2} \log \det (D'X'XD) + T \log |\det \beta| \quad ; \qquad (37)$$

that is solve the necessary conditions

$$\frac{\partial L_C}{\partial \beta} = \frac{\partial L_C}{\partial \Gamma} = 0 \qquad \text{for } \beta,\Gamma \qquad (38)$$

Evidently the unknown β,Γ enter (37) in a highly nonlinear fashion which is still present in (38). No explicit expressions for $\hat{\beta},\hat{\Gamma}$ are therefore given. We can, however, specify numerical algorithms either to maximise (37) or solve (38).

---

1. $C = TG \log \sqrt{2\pi}$ a constant irrelevant to the maximisation following. The form here is a little unconventional but will facilitate subsequent analysis.

(36) is obtained from

$$\left.\begin{aligned} \frac{\partial L}{\partial \Sigma^{-1}} &= T\Sigma - D'X'XD = 0 \\ \text{which together with} \qquad & \\ \frac{\partial L}{\partial D} &= [T(\beta')^{-1} - Y' \vdots -Z'] XD\Sigma^{-1} = 0 \end{aligned}\right\} \qquad (39a\text{-}b)$$

form the necessary conditions to maximise (35). Eliminating T, (39b) implies the necessary condition

$$\begin{bmatrix} -Z' \\ (\beta')^{-1}\Gamma' \vdots Z' \end{bmatrix} XD\,\Sigma^{-1} = 0 \qquad (40)$$

which (Hausman 1974,1975) may be written as the IV estimator

$$\hat{\delta} = (N'\hat{X})^{-1}N'y, \text{ where } N = (\hat{\Sigma} \otimes I)^{-1}\hat{X}^*, \text{ and } \hat{X}^* = \text{diag } X_i^*, \qquad (41)$$

if the normalisation of (5) is imposed and the ordering in (40) is rearranged since we need only retain these elements comprising the a priori unknown (nonzero) coefficients in β,Γ.

An iterative algorithm which solves (41) for β,Γ, which enter non-linearly, is rather obvious. (41) also allows several useful characteristics of FIML, which are not possible to derive directly from (38), to be demonstrated in a straightforward manner. This unconventional approach to FIML is therefore appropriate to our analysis of iterative estimators - Chapter 5. Meanwhile we will treat (38) and its properties as the conventional FIML.

Theorem 14:

The solutions $\hat{\beta},\hat{\Gamma}$ to (38) or (41) are (i) consistent, (ii) asymptotically equivalent to the 3SLS estimator (28), (iii) asymptotically efficient; (iv) $\sqrt{T}\ (\hat{\delta} - \delta) \overset{a}{\sim} N(0, \sigma^2(N'\hat{X})^{-1})$.

Proof:

(41) and (38) are yield equivalent solutions.

(a) Using (41): (i) conditions (19) hold for instrument set N. (ii) given (i) plim $T^{-1}N'\hat{X}$ = plim $\hat{X}^{+'}\ \phi^{-1}\hat{X}$ = plim $T^{-1}\hat{X}^{*'}(\Sigma \otimes I)^{-1}\hat{X}$ (noting (41) and (28a)) is immediate; as it is if $\hat{\phi}^{-1}$ replaces ϕ^{-1} in the practical estimating version of (28); where $\hat{\phi}$ is a consistent estimate of ϕ. (iii) Theorems 12 or 10. (iv) (20) with (28a).

(b) From (38) (i) Theorem 3.7 Dhrymes (1970) (ii) - (iv) Dhrymes remark (i) p.371 and Theorem 11. #

Comments:

FIML is better understood than LIML. For (38);

(a) FIML estimates may be obtained with or without the elimination of identities from the model (Malinvaud (1970) Chapter 19.6) - see also Chapter 14 below.

(b) FIML estimates may be obtained from independent estimation of independent subsystems of equations if they exist, i.e. if the model is block recursive. (Klein (1969) indicates it, Chow (1974) proves it rigorously including single equation maximum likelihood estimators).

(c) Dhrymes (1973) shows for a just identified equation FIML ≡ 3SLS ≡ ILS. Since in this case 3SLS ≡ 2SLS, no gain in asymptotic efficiency materialises over LIML. It is not clear if the case of $\sigma_{ij} = 0$, $i \neq j = 1 \ldots G$ the same is true (as with 3SLS and 2SLS); nor is it clear if such equations could be estimated in a separate block, whereas, naturally, just identified ones should be (but see Chpaters 5, 14).

Remarks:

It is not possible to assess the degrees of freedom used as this depends on the algorithm used to solve (38) or (41). Either method requires Σ non-singular which implies in both cases[1]

$$T > K + G \tag{42}$$

is required for the likelihood function (35) to exist and have a well defined maximum. (42) applies to both FIML and LIML; although we may actually use algorithms which work when $T \lesssim K + G$.[2]

It will be understood in what follows that any estimator $\hat{\delta}_T$, obtained with some T, is equivalent to the maximum likelihood ones $\hat{\delta}$ if $\hat{\delta}_T \equiv \hat{\delta}$ all $T = K + G + 1, \ldots. \infty$, but $\hat{\delta}_T$ uses identical estimating equations when $T \lesssim K + G$. No further distinction will be made between actual and equivalent maximum likelihood estimators and we concentrate on conditions for the existence of an algorithm.

Finally:

Theorem 15:

The maximum likelihood estimate $\hat{D}$, with log likelihood function L(D), is the unique, consistent and sufficient estimate, if $L(\hat{D}) \gtrsim L(D)$ (i.e. the global maximum). Its variance is consistently estimated by $-\left[T^{-1}\frac{\partial^2 L}{\partial D \partial D'}\right]^{-1}_{D=\hat{D}}$.

Proof:

Theorems 3.7 and 3.8 Dhrymes (1970). #

---

1. if M > G we can always eliminate the identities first.

2. One might argue that LIML estimates under the restriction Σ = diag $(\delta_i' X_i' X_i \delta_i)$, which is always nonsingular if $\delta_i' X_i' X_i \delta_i \neq 0$ so (42) does not apply. This philosophy places the restriction within the maintained hypothesis, since to test the diagonality without (42) implies no general alternative hypothesis exists. The restriction is not regarded as such here and would be at variance with the asymptotically equivalent 2SLS maintained hypothesis. Diagonality must remain a testable hypothesis to leave the option open of increasing efficiency if it does not hold.

Optimality:

The optimality of maximum likelihood estimators is obvious. LIML may also be thought of as the least variance ratio estimator, i.e.

$$\text{minimises } \lambda_i = \frac{\beta_i' W_{ii}^* \beta_i}{\beta_i' W_{ii} \beta_i} \quad i = 1 \ldots M \tag{43}$$

From (8), (26) and the corollary to Theorem 8,

$$\lambda_i = \frac{\hat{w}_i' \hat{w}_i}{\hat{v}_i' \hat{v}_i} \tag{44}$$

where $\hat{w}_i = y_i - Z_i \hat{p}_i$ a partial reduced form.

From (37) we have for FIML

$$L_C = C^* - \tfrac{T}{2} \log \det [\beta'^{-1} D' X' X D \beta^{-1}] \tag{45}$$

so, from (8), (45) is maximised when the determinant of the generalised variance

$$\beta'^{-1} D' X' X D \beta^{-1} = V'V \tag{46}$$

is minimised.

2.3.4 Systems of estimators:

It is well known that OLS, 2SLS and LIML can be summarised in the k-class system

$$X_i'(I - kQ)X_i \delta_{i(k)} = X_i'(I - kQ)y_i \quad i = 1 \ldots M \tag{47}$$

OLS, 2SLS and LIML correspond to k = 0, 1, $\hat{\lambda}_i$ respectively: $\hat{\lambda}_i$ the minimal root of (32) and $Q = I - Z(Z'Z)^{-1}Z'$. Moreover $\underset{T\to\infty}{\text{plim}}\ \hat{\lambda}_i = 1$, $\hat{\lambda}_i \geq 1$ and $T \log \hat{\lambda}_i \overset{a.}{\sim} \chi^2_{K-r_i}$ (proof: Dhrymes (1970) propositions 7.2 and 7.4).
A full information analogue exists

$$\hat{X}'(\hat{\Sigma}^{-1} \otimes (I - kQ))\hat{X}\ \delta_{(k)} = \hat{X}'(\hat{\Sigma}^{-1} \otimes (I - kQ))y \qquad (48)$$

(Savin 1973). It is obvious k = 0, 1 gives GLS, 3SLS respectively; and if Σ is diagonal we can obtain (47). We have in general

Theorem 16:

$\underset{T\to\infty}{\text{plim}}\ (k - 1) = 0$ and $\underset{T\to\infty}{\text{plim}}\ \sqrt{T}\,(k - 1) = 0$ are sufficient for $\delta_{i(k)}$ and $\delta_{(k)}$ to be consistent and asymptotically efficient.

Proof:

Savin (1973). #

2.3.5

In (14), (15), (27) and (28) Σ_i and Σ are assumed known. If they are not,

Theorem 17:

(a) $\hat{\Sigma}_i$, $\hat{\Sigma}$ consistent estimates of Σ_i, Σ do not alter the properties of (14), (15), (27) and (28).

(b) $\hat{\Sigma}_i$,Σ may be obtained in a 2-step procedure with $\hat{u}_i = y_i - X\hat{\delta}_i$ and $\hat{\Sigma}_i = \hat{u}_i\hat{u}_i'$ or $\hat{U} = Y - \hat{X}\hat{D}$ and $\hat{\Sigma} = \hat{U}'\hat{U}$, i.e. (13) as a preliminary for (14), (15) or (26) for (27), (28) provided $y_{t-j,i} \notin X_i$, $j > 0$ in (14), (27) and either $E(u_{ti}\, u_{sj}) = 0$ all $t \neq s$ $i,j = 1 \ldots M$ or $Y_{t-j} \notin Z$, $j > 0$ for (15), (28).

Proof:

Dhrymes (1970) Theorem 4.7. #

Theorem 18:

(a) If the 2-step procedure in Theorem 17(b) for (14), (15), under the conditions there, is iterated to convergence we get identically the maximum likelihood estimator.

(b) It will converge.

Proof:

(a) From (41) specialised to $\beta_i = 0$ all $i = 1 \ldots M$.

(b) Sargan (1964) if they result from a likelihood function, Telser (1964) if not. #

Similar, not identical, results are available for (27) and (28) involving iterative estimation (since $Y^* \neq Y^+$) in Chapter 7. It is, however, obvious that 2SLS and 3SLS may be treated as quasi-maximum likelihood estimates of (35) where prior estimates Y^+ and $\Sigma^+ = U^{+'} U^+$ are used for Y and Σ.

A number of suggestions have been made if the conditions of Theorem 17(b) fail. In the first case for the simultaneous model[1], the minimum distance estimator, (17), with

$$y_{ti} = y_{t-1,i}\rho_i + (X_{ti} - X_{t-1,i}\,\rho_i)\delta_i + e_i \qquad (49)$$

when $u_{ti} = u_{t-1,i}\,\rho_i + e_{ti}$ where $E(e_i e_i') = \sigma_{e_i}^2 I$, could be used. A numerical algorithm could be specified to search the parameter space for estimates satisfying the implicit restrictions. Simpler would be Durbin's (1960) method in which the restrictions are ignored. System (49) implies a reduced form

$$Y = Y_{-1}\beta R\beta^{-1} - Z\Gamma\beta^{-1} + Z_{-1}\Gamma R\beta^{-1} + e\beta^{-1} \qquad (50)$$

---

1. if $\beta_i = 0$ direct maximum likelihood would be possible.

which may be consistently estimated by OLS[1] where $U = U_{-1}R + e$; $Y^{+} = Y_{-1}\Pi_1^{+} + Z\Pi_2^{+} + Z_{-1}\Pi_3^{+}$ being the result. Replacing X_i by X_i^{+} from (50) in (49) we may proceed by OLS again (a 2SLS variant). If

$$E(e'e) \tag{51}$$

is not diagonal we may apply GLS to (49) with X_i^{+} from (50) to give a 3SLS variant. This method would be particularly recommended for good results by the study of Grilliches and Rao (1969). Clearly similarly transformed X_i^{*} could be used in IV form (41) since plim $T^{-1}X^{*\prime}e = 0$ where $X^{*} = (Y^{*} \vdots Z)$ and Y^{*} is the systematic part of (50). This may be shown to give a FIML variant, upon specification of the likelihood function of this case. Dhrymes et al. (1974) have suggested some other IV estimators which must be used iteratively. Therefore this aspect will be taken up again in iterative estimation (Chapter 7) where also LIML estimates for this case are available. One of these two methods, or direct likelihood maximisation itself must be employed since (5), (6), or Cochrane-Orcutt transformations of them, yield inconsistent and inefficient estimates.

2.4 Information Classes

Definition:

An information class is the set of all possible pairs $\{F(X,U;D);\ p(U;\theta)\}$ for which a particular given statistic $h = h(X,\theta)$ is jointly complete and sufficient.

Here X,D,U are as before and $F(\cdot,\cdot,\cdot)$ denotes a model whose parameters we wish to estimate. $p(\cdot,\cdot)$ denotes a joint probability density function, with parameter set θ, characterising U. The definition of information class

---

1. There could be s shortage of degrees of freedom here, making the approach infeasible. We can circumvent this - see Chapter 7.

is more restrictive than it need be in other contexts, but the pairs mentioned cover the whole model space of (2) and h is the only kind of statistic we will analyse. In fact if h is jointly sufficient then so is any 1-1 transform of h (Theorem VII.3 Mood, Graybill and Boes (1974)). Here we apply the definition only for a minimally jointly sufficient h. Completeness ensures that every function of h is an uniformly unbiased and minimum variance estimate of its expected value, and is therefore unique for that parametric function (Rao (1973) Chapter 5a.2). Therefore,

Theorem 19:

h, in an information class, is an estimator which is unique, uniform, unbiased and attains the minimum variance bound.[1]

Proof:

Theorems VII.10 and VII.17; Mood, Graybill and Boes (1974). #

Theorem 20:

h is an estimate of the parameters a model (2) or (6) with log likelihood function $L(\cdot,\cdot)$ then cov(h) has a lower bound given by the Cramer-Rao minimum variance bound - $E\left(\frac{\partial^2 L}{\partial D\, \partial D'}\right)^{-1}$ which is independent of the method of estimation. Those estimators which attain the bound are functions of the sufficient statistic.

Proof:

Rao (1973) Chapter 5a.3. #

Directly from Theorems 19 and 20 we can identify information classes by differing minimum variance bounds.[2] Secondly sufficiency implies the optimality

---

1. Minimum variance implies both $Var(h_i) \leqslant var(h_i^*)$, all $h_i^* \neq h_i$, for each element of h and the concentration ellipsoid of h is contained within that of h* - or $|cov(h)| \leqslant |cov(h^*)|$ being generalised variances.

2. Estimators which do not achieve a minimum variance bound are relatively inefficient.

of minimum variance within an information class, as well as in exhausting all the information in a given sample. Further note,

Lemma 1:

If in Theorem 20 a minimum variance bound estimator exists, then $L(\cdot,\cdot)$ admits of a sufficient statistic, and then h is a function of it.

Proof:

Dhrymes (1970) Lemma 3.23 and corollary 3.4. #

Considering only asymptotic properties, we know form the results quoted in the previous section for any regular $p(\cdot,\cdot)$, as $T \to \infty$ 2SLS, 3SLS, LIML, FIML estimates and their autoregressive variants are unbiased, equivalent functions of the sufficient statistic and therefore satisfy the conditions on h; and hence Lemma 1 and Theorems 19,20. Thus their properties are quoted for the information class implied by their asymptotic variance bound. (13) and (18) however do not satisfy the conditions placed on h for models (2) or (6).

If two statistics h and h* exist satisfying Theorems 19 and 20 where Var(h*) - var(h) is positive semi-definite then h is said to be associated with a higher information level and h to dominate h*. Otherwise they are of the same information class. An information class does not imply any estimator necessarily achieves the minimum variance bound. (Definition A1). We can establish a hierarchy of information classes. Of the estimators considered so far, as $T \to \infty$ we identify four classes

(1) limited information (LI),

(2) autoregressive (AR),

(3) full information (FI),

(4) full information autoregressive (FI-AR).

This follows from (28), (41) dominating (26) and (32); (27) dominating (26)

and (32); and the estimator implied by (51) dominating (28), (41).[1]

We cannot however say anything about information classes or levels for finite T. Moreover comparing (26) and (28) we have asymptotic variances of estimators

$$(\hat{X}^{+\prime}\Sigma^{*-1}\hat{X}^{+})^{-1}\hat{X}^{+\prime}\Sigma^{*-1}\phi^{*}\Sigma^{*-1}\hat{X}^{+}(\hat{X}^{+\prime}\Sigma^{*-1}\hat{X}^{+})^{-1} - (\hat{X}^{+\prime}\phi^{-1}\hat{X}^{+})^{-1}\hat{X}^{+\prime}\phi^{-1}\phi^{*}\phi^{-1}.\ \hat{X}^{+}(\hat{X}^{+\prime}\phi^{-1}\hat{X}^{+})^{-1} \tag{52}$$

where Σ^{*} = diag (Σ_i) and ϕ^{*} is the true error covariance matrix including the serial correlation elements. (52) may only be positive or negative definite iff

$$|\phi^{-1}\hat{X}^{+}\{\hat{X}^{+\prime}\phi^{-1}\phi^{*}\phi^{-1}\hat{X}^{+}\}^{-1}\hat{X}^{+\prime}\phi^{-1}\Sigma^{*}\{\hat{X}^{+}[\hat{X}^{+\prime}\Sigma^{*-1}\phi^{*}\Sigma^{*-1}\hat{X}^{+}]^{-1}\hat{X}^{+}\}^{-1}\Sigma^{*}-\lambda_i I| = 0 \tag{53}$$

implies either $\lambda_i < 1$ or $\lambda_i > 1$ all i = 1 .... M + K.[2] Clearly in the absence of further restrictions on ϕ, Σ^{*}, we cannot say if this is true and thus cannot rank information level (2) relative to (3) above. Otherwise information levels may be ranked as increasing in the order given.

We characterise each information level by implied minimum variance bounds, which coincide if $\phi^{*} = I$. Alternatively each could also be characterised by the <u>sphericality transformation</u>[3] matrix L where $\phi^{*} = LL'$, i.e. $E\{(\text{vec } LU)(\text{vec } LU)'\} = I_{MT}$. This is contrasted with the <u>orthogonality transformations</u> giving the projection of Y onto Z space (or any estimate of it, e.g. Y^{+}).

---

1. Proofs in order; Theorem 4.13 Dhrymes (1970) with Theorem 4.12 ; Theorem 4.3 Dhrymes (1970) with Theorem 4.4; both Theorems applied sequentially.
2. By Dhrymes (1970) proposition A.18 and Corollary A.6.
3. Since in any circumstances ϕ^{*} is positive semidefinite L exists, triangular and nonsingular.

We have the position summarised to date:

Information level: ↓ Estimation strategy and transformations: ⟶

	LS non-orthogonal : orthogonal		IV non-orthogonal : orthogonal		ML non-orthogonal : orthogonal	
(1)	OLS	2SLS	IV	Lyttkens (1971) Brundy & Jorgenson (1971)	OLS	LIML
(2)	GLS	2SLS(AR) (GIVE)	GIV	GIVE	-	-
(3)	Zellner (1962) seemingly unrelated	3SLS	-	GIVE Brundy & Jorgenson (1971)	GLS	FIML
(4)	Parks (1967)	(GIVE) 3SLS (AR)	-	GIVE	-	-

GIVE refers to the package, Hendry (1972), of IV estimators.

In every case full iteration on covariance matrices Σ^*, ϕ or ϕ^* is possible if the conditions of Theorem 17b apply. (c.f. Theorem 18). If lagged dependent variables are included then a suitable starting regression e.g. Durbins method, enables estimation to be completed as implied here, with the same asymptotic properties. Therefore we know the results of iterating on covariance matrices, but not from iterating on the systematic part of Y. The latter proposal, as an investigation of iterative estimators, will be our first objective.

There is obviously a fifth information class, namely the inclusion of a priori information. However, this case is not amenable to the previous definitions since different information "levels" result for a given model according to the degree of precision of that prior knowledge. In these information classes all the information exists in the data, but we make a decision whether to use it. In contrast, a priori information may or may not be to hand with useful precision (c.f. Meyer & Kuh (1957)) which precedes the decision as to whether we want to use it. Therefore this "information class" is of rather a different nature.

2.5 Computation

In the literature to date, one finds virtually no interest in the computational aspects of econometric estimation. However the subject should be of some interest for its own sake. Also if several estimators are available with identical asymptotic properties, but unknown properties in finite samples, relative computational burdens are grounds for discrimination which should enter into the selection of an estimator. Naturally when we consider iterative estimation, then computational considerations become vital.

For LIML we can only employ (31). Let $Z'Z = RR'$ where R is nonsingular and triangular (by assumption Z is of full rank).

Let $Q_i = R^{-1}Z'X_i$; $p_i = R^{-1}Z'y_i$ and $r_i = R^{-1}Z'u_i$ then

Theorem 21:

The computationally efficient 2SLS estimator is $\hat{\delta}_i = (Q_i'Q_i)^{-1}Q_i'p_i$.

Proof:

Calculating R^{-1} involves fewer operations than $(Z'Z)^{-1}$. If R is calculated by the method of Nagar (1968) it involves fewer operations than the multiplications $Z\hat{\Pi}_i$ in (26), $r_i \geq 1$, $i = 1 \ldots M$. In (22), calculating R involves fewer operations than the multiplication of $X_i'Z$ by $(Z'Z)^{-1}Z'X_i$. The remaining operations are in common in the two comparisons respectively. #

Corollary 1:

(28) is most efficiently obtained as $\hat{\delta} = (\hat{Q}^{*\prime}\hat{Q}^{*})^{-1}\hat{Q}^{*\prime}p^{*}$ where $\hat{Q}^{*} = \text{diag } R^{*-1}Q_i$ and $P^{*} = R^{*-1}P$, $P = (P_1 \ldots P_M)$, $p^{*} = \text{vec } (P^{*})$. $\Sigma = \hat{r}'\hat{r}$, $\hat{r} = (\hat{r}_1 \ldots \hat{r}_M)$, and $\hat{\Sigma} = R^{*\prime}R^{*}$; R^{*} is nonsingular, triangular and obtained by Nagar's method. #

Throughout we assume the sequence of calculations is performed to minimise the total required to obtain an estimator.

2.5.1 FIML Computation:

There are several ways of obtaining FIML estimates from (37) or (38). All are also valid for the minimum distance estimation of (17) and thereafter. We may solve equations (38)[1] iteratively by one of many well known methods (e.g. Jacobi, Gauss-Seidel, SOR). These methods are analysed in detail in part III. The resulting FIML estimates are analysed for computational properties in part IV. This method is the main interest in subsequent analysis.

Traditionally direct maximisation of (36) has been favoured by econometricians. There are three approaches - direct search methods, iterative methods involving the evaluation of derivatives and those that do not. A selection of methods of each kind is briefly reviewed in Walsh (1975). Search methods always

---

1. In the case of model (5) these equations are (41). For model (2) we require the analytic derivative of the likelihood function - see Chapter 15.

produce a solution but also require an extremely large number of function evaluations in $n = \sum_{i=1}^{M}(r_i+s_i)+\frac{1}{2}M(M+1)$ dimensions when a grid with a sufficiently fine mesh of evaluation points is used. It is generally recognised for the general problem (37), gradient methods will be superior. The best known gradient methods are Newton-Raphson, Davidon-Fletcher-Powell, and Powell - the first two require derivative evaluations but the last does not. Obviously these or other algorithms which maximise F(X) with respect to X, also solve the equation system F'(X) = 0.

The Newton-Raphson iteration for (36) where $\theta = \{D,\Sigma\}$ the parameter set is

$$\theta^{(k)} = \theta^{(k-1)} - L''(\theta^{(k-1)})^{-1}L'(\theta^{(k-1)}) \qquad (54)$$

$$L''(\theta^{(k-1)}) = \left[\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right]_{\theta=\theta^{(k-1)}}, \quad L'(\theta^{(k-1)}) = \left[\frac{\partial L}{\partial\theta}\right]_{\theta=\theta^{(k-1)}}$$

Newton-Raphson has the following convergence characteristics.

Theorem 22:

If the solution to (36) is θ^*, under the regularity conditions on $L(\cdot,\cdot)$ and if $L''(\theta^*)$ is nonsingular then $\exists$ a neighbourhood about θ^* so that for arbitrary starts within that neighbourhood, (54) converges only if $\rho(L''(\theta^*)) < 1$. Further it has quadratic superlinear convergence, i.e. $||L'(\theta^{(k)}) - L'(\theta^*)|| = O(\theta^{(k)} - \theta^*)^2$.

Proof:

Ortega and Rheinboldt (1970) Theorem 10.2.2. #

Although the method has very fast convergence when convergent, it may not converge at all. Further it involves the evaluation and inversion of the Hessian matrix at each iteration, involving considerable calculation. It is

therefore only recommended when the start is close to θ^* (e.g. from 3SLS estimates) or to replace slower algorithms already close to convergence. It is possible to reduce computation by approximating $L''(\theta^{(k)})$ by suitable differences[1] and by not updating it at every iteration. However the latter modification reduces the rate of convergence (the main benefit) and is therefore not considered (see Ortega & Rheinboldt (1970) Theorem 11.3.5).

The Davidon-Fletcher-Powell algorithm is specified as

$$\phi^{(k)}(\alpha) = - L\left(\theta^{(k)} - \alpha^{(k)} H^{(k)} L'(\theta^{(k)})\right) \tag{55}$$

where $\alpha^{(k)} > 0$ is chosen such that $|\phi^{(k)}(\alpha)| > L(\theta^{(k)})$ (56)

then $\theta^{(k+1)} = \theta^{(k)} - \alpha^{(k)} H^{(k)} L'(\theta^{(k)})$ and

$$H^{(k+1)} = H^{(k)} - (d^{(k)\prime} y^{(k)})^{-1} \left[d^{(k)} y^{(k)\prime} H^{(k)} + H^{(k)} y^{(k)} d^{(k)\prime} - \left[1 + \frac{y^{(k)\prime} H^{(k)} y^{(k)}}{y^{(k)\prime} d^{(k)}} \right] d^{(k)} d^{(k)\prime} \right]$$

where $y^{(k)} = L'(\theta^{(k)}) - L'(\theta^{(k-1)})$ and

$$d^{(k)} = - H^{(k)} L'(\theta^{(k)}) \tag{57}$$

In this form the "uphill" property, i.e. $L(\theta^{(k)}) \gtrsim L(\theta^{(k-1)})$, $k = 0,1, \ldots$. is guaranteed (Powell (1972)). Moreover the line search for $\alpha^{(k)}$ which minimises $\phi^{(k)}(\alpha)$ is not necessary, as precisely the same iterates $H^{(k)}$, $L'(\theta^{(k)})$ are generated as long as (56) is observed, (Powell (1972) Dixon (1973)). A streamlined version of (55)-(57) is based on the conjugate

---

1. As in Secant or Steffenson methods.

gradient method: let $t^{(1)} = -L'(\theta^{(0)})$ and $d^{(k)} = \alpha^{(k)}t^{(k)}$, $\alpha^{(k)}$ chosen so (56) with $t^{(k+1)}$ for

$$H^{(k)}L'(\theta^{(k)}) \tag{58}$$

holds; then

$$t^{(k+1)} = -L'(\theta^{(k)})^* + bt^{(k)} \tag{59}$$

in place of (58) where $b = \dfrac{[L'(\theta^{(k)})^*]'[L'(\theta^{(k)})^* - L'(\theta^{(k-1)})^*]}{t^{(k)'}[L'(\theta^{(k)})^* - L'(\theta^{(k-1)})^*]}$

and $b = 0$ if $k = n$; and $Z^{(k+1)} = Z^{(k)} + \dfrac{d^{(k)'}\left(L'(\theta^{(k)})\right)}{d^{(k)'}y^{(k)}}\, d^{(k)}$;

$$w^{(k+1)} = w^{(k-1)} + \frac{d^{(k)'}(L'(\theta^{(k)})}{d^{(k)'}y^{(k)}}\, y^{(k)} \text{ and } L'(\theta^{(k)})^* = L'(\theta^{(k)})-w^{(k)} \tag{60}$$

where $Z^{(0)} = 0 = w^{(0)}$ and $\theta^{(k)} = \theta^{(k-1)} + d^{(k)}$. Either method terminates where $L'(\theta^{(k)}) = 0$; $\theta^{(k)} = \theta^*$. This method also enjoys the property that line searches are unnecessary, with no deterioration in convergence characteristics (Dixon (1975)). It avoids the cumbersome calculations (58) which is valuable if n is large.

Lemma 1:

Method (58) - (60) has the "uphill" property.

Proof:

Follows conjugacy of directions $d^{(k)}$ $k = 1 \ldots$, Theorem 4.5 and corollary 1, Walsh (1975). The conjugacy is preserved if (59) is used (and extension to this method by Dixon (1975)). #

If we augment either method (55) - (57) or (58) - (60) with termination condition (i) $L(\theta^{(k)}) = 0$ and (ii) if $\exists$ no $\alpha_o > 0$ such that

$\phi_i^{*(k)}(\alpha) = -L(\theta^{(k)} - \alpha E_i)$, where E_i is the i^{th} unit coordinate vector,[1] implies

$$|\phi_i^{*(k)}(\alpha)| \gtrsim L(\theta^{(k)}) \qquad i = 1 \ldots n \tag{61}$$

for the whole open interval $\alpha\varepsilon(-\alpha, \alpha_o)$, then

Theorem 23:

The two methods are convergent

(a) for arbitrary start $\theta^{(0)}$ if $L''(\theta)$ is negative definite, all admissable θ

(b) for an arbitrary start $\theta^{(0)} \varepsilon N(\theta^*)$, $N(\theta^*) = \{\theta | L''(\theta)$ negative definite, all $\theta\}$, is a closed and bounded subset of the parameter space;

(c) they have a superlinear convergence rate of at least $2^{1/n}$.

Proof:

(a) Powell (1971) and Ortega and Rheinboldt (1970) Theorems 14.5.2, 14.5.4. Extensions to (55) - (57) and (58) - (60) by Dixon (1973,1975) respectively.

(b) The regularity conditions of $L(\cdot,\cdot)$ imply $N(\theta^*)$ is compact; this now follows from (a), and the "uphill" properties.

(c) Powell (1971), Daniel (1970). #

Lemma 2:

$\exists$ at least one $N(\theta^*) = \{\theta | L''(\theta)$ negative semidefinite, all $\theta\}$, closed and bounded subset of the parameter space.

Proof:

$L(\cdot,\cdot)$ of (36) is assumed twice differentiable with bounded differentials such that $\int_{-\infty}^{\infty} \ldots\ldots\ldots \int_{-\infty}^{\infty} L(\cdot,\cdot)\ d\theta_1 \ldots\ d\theta_n = 1$; logarithms being a monotonic transformation. #

---

1. Strictly this should be any set of vectors which are a basis for E^n.

Remarks:

(i) The methods are guaranteed convergence only to a local maximum under the conditions (a) or (b) of Theorem 23. Thus a global maximum may only be found by a search over possible starts, and stationary values of the likelihood function compared to indicate for the maximum.

(ii) The termination condition (ii) prevents the algorithm halting at the many points representing local minima, saddle points or points of inflection. The "uphill" property prevents returns to those points if, when (61) holds for index i, then $L'(\theta^{(k)}) = E_i$ is used and the algorithm proceeds as before from (56) or (58).

(iii) In view of Lemma 2 and the previous remark, either modified algorithm always converges if several starts are tried so that one $\theta^{(0)}$ must be in $N(\theta^*)$. For $\theta \notin N(\theta^*)$ they could fail if for some k, $\frac{\partial \phi^{(k)}(\alpha)}{\partial \alpha} > 0$ all $\alpha \geq 0$. But any $\theta^{(k)}$ is a start for $\theta^{(k+1)}$ .... so to use some $\alpha < 0$ satisfying (56) or (58) may be regarded as a new start and global convergence follows.

(iv) These algorithms may therefore fail at points in a horizontal hyperplane which may be part of a maximum, minimum, "point" of inflection or saddle "point". But in general cases of nonconvex surfaces $\theta^{(k)} \in N(\theta^*)$ will still result from $\theta^{(0)} \notin N(\theta^*)$, by (iii) above.

(v) The extensions to using α values not obtained by line searches may imply different local maxima are achieved, compared to when line searches are used.

(vi) Therefore, to solve

$$F'(X) = 0 \tag{62}$$

these methods will not always be appropriate, i.e. if it is specified independently of $L(\cdot,\cdot)$ we cannot even use Lemma 2. Generally they will only be successful if F'(X) is locally definite[1] at the solution and we start within a neighbourhood of it.

---

1. Choosing an appropriate sign at (55).

(vii) It is well known these mthods can be very slow for functions with awkward shapes (Goldfeld & Quandt (1972)).

Clearly (59) - (60) is superior computationally to (55) - (57) and will be referred to as the conjugate direction method. If $L(\cdot,\cdot)$ is quadratic then $\theta^* = \theta^{(k)}$ when $k \leq n + 1$, but for other nonlinear functions/no finite in particular (45) termination property has been proved. An algorithm for the general function (36) is detailed in Powell (1972) and Dixon (1973) respectively.

The Powell algorithm is summarised in Walsh (1975). It avoids the evaluation of any derivatives at a cost of performing (n + 1) line searches per iteration. Since the above methods require only vector derivatives which may be approximated by differences (and no inversions) this seems unlikely to be beneficial even if n is very small as the computation in accurate line searches is considerable. The method converges slower than (59) - (60) (Walsh (1975)) if $L(\cdot,\cdot)$ is quadratic, and although it is convergent in the general case under the conditions of Theorem 23 (see Zangwill (1967)), it will be slower. Moreover the method breaks down if any of the n marginal log likelihood functions becomes for some line searches unbounded or undefined. Since the searches cannot be restricted, convergence is then impossible. Convergence may be slower if a marginal function has several or indistinct maxima. No relaxation of the line searches has been proposed but if such were to prove convergent, it would be slower due to smaller changes in the list of search directions. Again a maximum is a local one - other stationary values are prevented by (61). For these reasons this algorithm will not be pursued further.

Only Newton-Raphson calculates standard errors of estimates automatically.

2.5.2 Restricted Estimation

Restrictions may be a priori knowledge of bounds on the parameters or (possibly stochastic) relations between them. In either case linear restrictions may easily be incorporated in (26), (27), (28) by the Theil-Goldberger

(1961) formulation. It is assumed this has been done if required.

For nonlinear restrictions, or if it is desired to obtain the maximum likelihood estimator, then the likelihood or distance functions (37) or (17) must be maximised subject to constraints in the usual way. The foregoing algorithms may be adapted directly (see Walsh (1975), Powell (1972) for details) when the restrictions are relationships. If they are bounds then an algorithm which may be used with any of the foregoing is contained in Fletcher and Jackson (1974) with no alteration to convergence properties.

Restrictions present no further difficulties.

2.5.3 Simplified FIML estimators:

Several have been suggested e.g. Chow (1968), Rothenburg and Leenders (1964), Bergstrom (1974), Dhrymes (1973), for when FIML computation is too heavy. However they are not simpler than 3SLS, whose asymptotic properties they share. It is not clear that they are superior approximations to FIML than (28) is to (41). Therefore it will be assumed that if more computation than (28) is undertaken it is with the objective of FIML estimates and therefore simplified versions are not of interest.

CHAPTER 3: THE OBJECTIVES AND PROCEDURE

3.1 Introduction

Turning now from the conventional estimators of (2.5) we review the analysis to be undertaken for the other estimators considered in this study.

We can enumerate the large number of alternative estimators that have been proposed in econometric literature and analyse their properties in a parallel manner to Chapter 2. This is done because almost without exception the statistical properties are not known, the relationships between estimators, conventional and unconventional, have not been investigated and certainly no analysis of their computational problems has been undertaken.

However our investigation of these estimators was motivated, in particular, by the requirement to separate economic and statistical maintained hypotheses in (2.5), and to relax the statistical as far as possible; by the necessity of developing estimators for models with small and inadequate samples; and by the desirability of a preference ordering over the estimators so that we can define a rational strategy for estimator selection. It is therefore necessary to state and justify selection criteria and justify the approach to answering these questions through asymptotic statistical properties and computational characteristics. The strategy of the analysis is therefore outlined in Section 2.

Chapter 1, and the several suggestions in Chapter 2, provide a potent theoretical justification, as well as motivation, for undertaking the study. It is also essential to provide evidence of empirical justification by differences in estimated versions of separate econometric models under various estimators. The results effectively dismiss the charge that different estimators rarely yield significantly different estimates or that the inevitable combination of mis-specifications and data problems make attempts to attain the desirable statistical properties pointless.

3.2 Selection criteria:

Criteria for estimator selection follow from our stated objectives in econometric theory. Naturally they will then automatically define the strategy of our analysis.

3.2.1

All the conventional estimators, with the exception of OLS, are unavailable in inadequate samples. It may also be argued that in small samples the small number of degrees of freedom render them unreliable, and probably undesirable, in terms of $\bar{R}^2$, the confidence interval of any estimate or forecast, and the accuracy of the approximation in making use of asymptotic statistical properties.[1] On the first count it is essential to consider alternative estimators that are available in the inadequate sample case, since it has proved so common in practice; and they should be estimators with as wide availability as possible.

Definition 1:

(i) The criterion of availability is that the estimator must be, at least, feasible whenever the OLS estimator of the same information class is. That is, covering the four classes, (2.13) for limited information estimators and (2.15) for full information, requiring $T > \max_i(r_i + s_i)$ and $T > \max_i(r_i + s_i, M)$ respectively. (ii) If one can be found with wider availability so much the better. On the second count the widest availability will be desirable so the criterion of availability will take this into account.

None of the conventional simultaneous estimators as they were presented in Chapter 2 satisfy the criterion of availability. In particular 2SLS, LIML are available only if $T - K > 0$ (for the existence of $(Z'Z)^{-1}$). 3SLS requires $T - \max(K,M) > 0$ if Σ^{-1} and $(Z'Z)^{-1}$ are to exist. Moreover both FIML and LIML[2] follow from the specification of a likelihood function whose existence

---

1. We cannot prove it in general but the conclusion of Chapter 16 is that finite sample bias and mean square error, where they exist, are likely to be approximately monotonic decreasing functions of T for consistent k-class estimators. This is certainly borne out in many Monte Carlo studies.

2. Recall footnote 2 to (2.42) that we do not consider LIML from $L(\beta,\Gamma,\Sigma) = \sum_{i=1}^{M} L_i(\beta_i\ \gamma_i\ \sigma_i)$.

depends[1] on T > K + M. FIML estimates obtained by direct maximisation of (2.35) require that condition. This does not preclude finding alternative algorithms to obtain the estimates requiring less stringent conditions - as in (2.41) where $T - \max_i(r_i + s_i, M) > 0$ is required for FIML equivalent estimates. There is also the possibility of finding modifications for 2SLS, 3SLS and LIML giving them wider availability. Although the estimators at this stage are neither available for inadequate samples nor satisfy the criterion of availability, they will be retained as benchmarks for the performance of other estimators.

3.2.2

Definition 2:

The criterion of optimality is that any estimator must produce estimates which minimise some well defined penalty function. In the case of maximum likelihood estimators the optimality rationale is obvious. For IV estimators, maximal correlation with the replaced variable as in (2.21) will be required (so Theorem 2.12 follows). For least squares estimators the requirement is it shall be a minimum distance estimator for the Euclidean norm on the stochastic components of an equation[2] or model. They will then also be minimax estimators. Many estimators may have more than one optimality property and, of course, any estimator that is equivalent to another with a particular optimality property is said to have that property also.

3.2.3

Estimators which survive the two criteria above will be selected on the desirability of their statistical properties. The first point is that a lexographic preference ordering placing desirable statistical properties above desirable computational characteristics follows immediately from the domination of the requirement to separate and relax the statistical maintained

---

1. Klein (1969) Sargan (1975)
2. As usual the Euclidean norm is arbitrary but convenient. Minimum distance estimators of other norms would be acceptable also.

hypothesis. We are assuming, thereby, extra computation has no disutility if an improved statistical property can be so achieved. By implication of the above arguments, we extend this assumption to the criteria of availability and optimality.

A problem arises in what statistical properties are to be considered. The candidates are bias, variance and mean square error in parameter estimates[1] representing location and spread. So far we have obtained only asymptotic statistical properties and the difficulty is to determine the sample size required to attain these properties, or at least an adequate approximation to them. In practice sample sizes will be quite small and far short of infinite. Therefore finite sample properties would be more appropriate for estimator selection, particularly as we have taken a special interest in inadequate samples. But the conclusion of Chapter 16 is that finite sample properties are not generally available. Indeed for a number of estimators of interest they do not even exist. Selection therefore must be performed on asymptotic properties.

Asymptotic properties must be interpreted carefully. If we deal with small or inadequate samples it implies if $T \to \infty$ then $T - K$ remains finite (from increasing model size, or just considering large T and K). It would appear there is no cause to seek desirable asymptotic properties, still less to claim them as an approximation and select estimators accordingly. However, (a) by Sargan (1975) Theorem 3 and the subsequent conjecture, we may still treat asymptotic properties as a good approximation in these cases; and (b) any estimator that satisfies the criterion of availability has asymptotic perperties a function of $T - \max_i(r_i + s_i)$ degrees of freedom[2] which of course becomes infinite if $T \to \infty$, inadequate sample or not. Thus asymptotic properties retain their conventional interpretations on either count. Additionally the sample size required to achieve a given accuracy of approximation to the

---

1. Degree of fit problems having been settled by the optimality criterion.
2. See Lemma 4.1 below for full information methods satisfying the condition, $T - \max_i (r_i + s_i, M)$ for finite, M, model size.

asymptotic properties is less if the availability criterion is satisfied whatever K and model size.

We should also require a weighting of those statistical properties so that we can derive a unique ordering of estimators. Traditionally a lexographic preference for unbiasedness over efficiency has been used - and criticised. Asymptotically the property of consistency avoids this debate since it implies zero mean square error. Since all the estimators selected hereafter are consistent, we prefer that with the lowest minimum variance bound (by Section 2.4). We return to this problem in the finite sampling theory in Chapter 16, where we fail to obtain a minimum mean square error estimator, or even find which of the estimators in this study has lowest mean square error (except in certain special cases). A weak justification for this lexographic ordering is thus provided, although it is hard to say if we actually use it in the finite sample case. It will be observed that our preferred estimators generally display lower biases, rather than necessarily lower variances and mean square errors; but this follows by the selection of asymptotic properties.

3.2.4

The combination of the three criteria above will clearly allow estimator selection independent of the economic maintained hypothesis. Further being forced to accept asymptotic statistical properties implies a certain relaxation of the statistical maintained hypothesis. Obviously a fourth criterion is that we prefer estimators which relax the statistical assumptions most.

3.2.5

The construction of selection criteria independent of the economic specification, may be argued to be less than optimal (as by Salmon & Eaton (1974)). It is inevitable that a model is not an exact representation of the system being modelled and thus there will be mis-specification biases. Robustness, or at

least sensitivity to mis-specifications, do not enter the selection process, except perhaps in the form of a few subsidiary considerations, because there are no general results available. Since we have no particular model in mind we must assume the general model (2.5) is correctly specified. The possibility that the simultaneity is weak and therefore OLS being more robust in the face of specification problems cannot be counted on to justify OLS as "near" consistent anymore than assuming β is block diagonal in general.

3.3 The empirical models

Nine popular and well known macroeconometric models were selected to demonstrate the empirical justification of this study. They were chosen to be representative of as widely differentiated model types as possible. They are retained to illustrate the results, and to demonstrate their impirical significance.

The models are (i) Summers (Summers (1965)) although actual parameter values were taken from model 2j in Mosbaek & Wold (1970); (ii) Klein I (Klein (1950)); (iii) Christ version 17 (Christ (1966)); (iv) Klein-Goldberger (Klein - Goldberger (1955) with the linearisation of Bergstrom (1974)); (v) Dutta-Su (Dutta & Su (1969)); (vi) Michigan (Suits (1962)); (vii) Hendry (Hendry (1974)); (viii) Wharton (Evans (1969)); (ix) Girschick-Haavelmo (Girschick and Haavelmo (1947)). In each case the specification may be found in the original expositions but a convenient summary, and the various estimated versions used here, of models (i) - (v) and (ix) are contained in Bergstrom (1974) and Agren (1972). For purposes of illustration the latter are taken as definative.

The nine models reflect a variety of different characteristics of typical econometric models. The size of M varies from 2 to 74, T from 10 to 68, K from 4 to 325, the spareseness of β from 0 to 66% a priori zero elements, the proportion of identities, (M - G)/M, from 0 to 60%; plus widely varying

degrees of overidentification, degrees of freedom positive and negative either as T - K or T - $\max_i(r_i + s_i)$, emphasis on dynamic structure, and predetermined variables per equation. The detailed differences in characteristics are left for the reader to summarise.

3.4 Empirical differneces due to estimation technique

Here we provide an empirical justification for this study. We use a subgroup of the models above, for which estimated versions are avilable for several estimators. The subgroup comprises Klein I, Christ, Klein-Goldberger, Dutta-Su and Hendry.

3.4.1 Structural form comparisons:

Several specimens of estimated parameters which vary substantially under different estimators in Tables 1 and 2. Further examples are contained in the references given.

Table 1

	Klein		Christ	Klein-Goldberger			
Estimator	β_{15}	β_{25}	β_{51}	$\beta_{14,1}$	β_{85}	γ_{33}	$\gamma_{22,9}$
OLS	.193	.48	.66	.565	-34.7	.419	-.624
TSLS	.017	.15	.68	.494	-42.9	.401	-.624
3SLS	.124	-.013	.73	-	-	-	-
LIML	-.222	.075	.73	.316	-46.5	.327	-.628
FIML	-.232	-.801	.52	.176	-72.7	.117	-.141
IIV	-.118	.107	.503	.342	-52.9	.412	-.625
FP	.022	.089	.37	.313	-28.8	.33	-.446
t	-	-	-	5.1*	7.2*	2.6*	-3.6

Table 2

Hendry

Estimator	$\beta_{8,3}$	β_{45}	γ_{11}	γ_{44}
2SLS	.19	.4	.69	.52
FIML	-.09	.49	.68	.4
ALS(1-4)	.58	.58	.26	.16
AML(1)	.02	.99	-.09	-.29
AML(4)	-.03	.28	.64	.35
t	5.2	2.2	4.5	2.7

(i) t is the asymptotic test statistic of Theorem A.2.5(i) (whose critical value is 1.96) for the two extreme values of that column.

* indicates the test is performed for the value of one coefficient fixed, because its standard error was not available.

(ii) ALS(1-4) is autoregressive least squares with up to fourth order serial correlation; ALM(1) and AML(4) are autoregressive maximum likelihood treating first and fourth order serial correlation respectively.[1]

(iii) The coefficients represent the following: Klein I, β_{15} the influence of profits on consumption, β_{25} profits on investment; Christ, β_{51} short run marginal propensity to consume; Klein-Goldberger $\beta_{14,1}$ is the short run marginal propensity to consume, β_{85} influence of the change in earnings on hours worked, γ_{33} the autoregressive parameter[2] in inventories, $\gamma_{22,9}$ the proportion of real net profits retained; Hendry, β_{83} is the accelerator, β_{45} the proportion of imports going to inventories, γ_{11} is the autoregressive

---

1. Details in Hendry (1974)

2. Autorgressive parameters contain a measure of the rate of adjustment to equilibrium.

term in consumer durables and γ_{44}, that in inventories. All of these are central concepts to any respectable econometric model.

(iv) Dutta-Su has no significant differences (apart from constants which are neglected throughout).

We note these differences in estimates are not only significant but substantial, and several involve changes of sign giving results contrary to well established economic theory. Secondly often the differences between maximum likelihood and iterative estimates is notable. It is shown below that theoretically LIML $\equiv$ IIV $\overset{a.}{\equiv}$ FP which suggests that they converge on separate solutions in these models. Plural solutions may therefore be expected to be common, a fact never emphasised in the literature to date.

Finally these significant differences may provide some empirical justification for the study in terms of structural inferences. Their presence or absence alone do not indicate significant differences to the performance of the model as a whole in the context of all other smaller differences.

3.4.2 Impact multiplier comparison:

Impact multipliers are the restricted reduced form coefficients. Table 3 contains some specimens, which show substantial differences.

Table 3

	OLS	2SLS	3SLS	LIML	FIML	IVV	FP	Multiplier
Christ	3.28	3.38	3.95	3.44	-	1.97	1.92	Government spending on GDP
Klein-Goldberger	4.36	3.24	-	1.73	.896	1.61	2.02	Import or general prices on consumption
"	.63	.7	-	.874	1.034	.835	.929	Balanced budget on GDP
"	1.79	1.65	-	1.5	1.39	1.5	1.53	Government spending on GDP
"	.55	.42	-	.234	.132	.25	.29	Government spending on consumption
"	-1.16	-.95	-	-.58	-.36	-.61	-.6	Real tax yield on GDP
"	-.92	-.74	-	-.41	-.21	-.45	-.43	Real tax yield on consumption

Many of the variations are large[1]; several of the multipliers vary by a factor of more than 4 for different estimators. The implications for economic policy analysis and planning exercises are both enormous and disastrous.

3.4.3 Comparison of forecasts:

Extensive forecasting experiments of these models under different estimation methods are described in Bergstrom (1974) and Agren (1972). Overall the results are very confused mainly because in no model is there a unique ordering of forecasting ability encompassing all endogenous variables. Average performance statistics tend to be very similar for each method and are not reproduced. Moreover such orderings that do emerge vary considerably from model to model.

A general synopsis of the results for the four models, excluding Hendry's is that iterative methods perform well together with FIML which tends to be somewhat more erratic. In contrast OLS is nearly always worst, often accompanied by 2SLS and 3SLS. These results support the theory of Section 16.2.

3.4.4. Final forms and stability:

The most awkward problem in assessing the performance of multiequation econometric models is to design an acceptable, but not arbitrary, criterion for the performance of the model as whole, such that an unambiguous measure results. Traditionally accuracy in "forecasting history" outside the sample period, and simulation of particular events have been used. One could also look for the highest sum of R^2 statistics, the likelihood function value, or maximum total explained sum of squares in endogenous variables as very similar indicators.

An alternative is to take the dynamic stability properties as a single measure of performance of a model for comparison purposes. Ideally the model

---

1. No standard errors can be calculated for restricted reduced forms.

should show the same cyclical properties as were observed in the actual economy of the sample period. We compare stability properties to show the differences between estimated versions of any model. The cyclical characteristics may be deduced from the eigenvalues of $\hat{\Pi}_1^*$ - estimated from the restricted reduced form. Note that an eigenvalue may be written a + bi = r(cos θ + i Sin θ) indicating a cycle of amplitude r = |a + bi| and period = 2Π/θ where cos θ = a/r. Some results for the models quoted are given in Table 4.

Table 4

Cyclical Characterisitics

Model version		Dominant eigenvalue	Amplitude	Period/years
Klein	OLS	.882	.882	∞
Klein	2SLS	.759 ± .351 i	.836	14.6
Hendry	2SLS	.9799 ± .0073 i	.97995	13.1
Hendry	FIML	.91887	.91887	∞
Hendry	ALS	.9843 ± .041 i	.9852	34.9
Hendry	AML1	1.0	1.0	∞
Hendry	AML4	.9522 ± .0162 i	.9523	60

Other versions of Klein (LIML, IIV, FIML, 3SLS and FP) plus Klein-Goldberger versions (OLS, TSLS, LIML, IIV, FIML and FP) and Dutta-Su versions (OLS, TSLS, IIV and FP) were investigated. All had spectral radii greater than or equal to unity.[1] Nowhere have any approximations to the standard errors of the spectral radii been calculated. On the basis of the point estimates all Dutta-Su versions and the first four Klein-Goldberger ones were weakly stable (i.e. $\rho(\hat{\Pi}_1^*) = 1$) indicating identical stability characteristics in each version. The remainder were dynamically unstable thus

1. Chow (1975) also notes that this characteristic is frequently observed. It may be attributed to the presence of lagged dependent variables in the identities.

suggesting inconsistent parameter and eigenvalue estimates.[1] Results beyond Table 4 were therefore not considered.

It can be seen from Table 4 that very different results can be obtained from different estimators - not only in spectral radii but more so in the period of cycles. The cyclical characteristics, despite rather similar looking dominant eigenvalues, are a very sensitive indicator of differences between estimated versions capable of producing a unique preference ordering. Table 4 does indicate that the choice of a suitable estimator can be very important, as well as the choice of the appropriate autoregressive scheme.

Equilibrium multipliers may be obtained from the final form matrix $(I - \hat{\Pi}_1^*)^{-1} \hat{\Pi}_2^*$ and this was calculated for the consistent verions of Hendry's model. Substantial differences were observed as might be expected from Table 4. Of the 40 equilibrium multipliers larger than 0.1 in absolute value, 21 showed sign changes and a further 11 showed differences in absolute size of greater than a factor of 10 (of which, 6 of the differences exceed a factor of 100).

3.5 Conclusions

It is clear from several points of view, in terms of the use of a macroeconometric model, that the choice of estimator may make a considerable difference, in empirical terms, to the estimates obtained. We can use the above argument to justify rather lengthy computations in estimators which have desirable asymptotic properties, as well as robustness and suitable finite sample properties where possible.

There some well known results of particular estimators being highly sensitive to certain problems. For example, maximum likelihood estimators are more sensitive to collinearity in the data (Klein (1969)) or the state of the

---

1. Therefore some elements of Tables 1 - 3 should be interpreted with caution as they are not strictly valid as comparisons. However it is felt that the differences shown are too large to be the result of inconsistent estimation.

data in the sense of Farebrother (1974). Further some simple results are possible on the sensitivity of estimators to small changes in the specification and mis-specification biases, but it is hard to generalise, and here we have taken the given model as correct. Without a set of data it is not possible to empirically demonstrate the sensitivity to these problems.

CHAPTER 4: ALTERNATIVE ESTIMATORS

4.1 Introduction

In this chapter we review alternative estimators for estimating the model (2.5). They are assessed by the selection criteria of Chapter 3, and those that survive pass on to subsequent analysis. The estimators are treated in groups with similar properties, differing only in their computational algorithms. Thus we make use of the lexographic preference over computational characteristics (which are discussed in Chapter 13). Although, obviously, not every possible variant of each estimator is included, it is thought the grouping of estimators is sufficiently general to cover very nearly all of them.

4.1.2 Recursive systems

Estimation methods that may successfully be used on Recursive or Block Recursive models have been mentioned in Chapter 2. Since such specifications have not been assumed, general analysis cannot proceed further. Algorithms for identifying any block recursive structure have been proposed by Steward (1962) although it seems unlikely that any useful results can be expected from most models, especially for those whose observations occur at short time intervals.[1] If a block recursive structure is present, we may treat the estimation problem as that within the largest block found.

4.2 Conditions for consistency

No estimator derived from a likelihood function can, by (2.42), satisfy the criterion of availability for (2.5). Without loss of generality therefore we may consider only repeated least squares (RLS) or IV estimators. Estimators which satisfy the criteria of availability, optimality and relaxation of

---

1. See for example Salmon and Eaton (1975) for a case in point.

assumptions, will be preferred if they are consistent and asymptotically efficient. It is therefore useful to obtain conditions for consistency and asymptotic efficiency, so that such estimators may be constructed.

Lemma 1:

Consistency follows in any least squares or IV estimator of (2.5) which satisfies the criterion of optimality iff Y_i alone in X_i $i = 1 \ldots M$ is replaced in the regressor or instrument sets respectively.

Proof:

Immediate from Theorem 2.1, with (2.17), for least squares; and (2.19), with (2.21), for IV estimates; that is for limited information estimators where

$$E(U_iU_i') = \sigma_i^2 I \text{ and } E(U'U) = \text{diag}(\sigma_{ii}) \quad i = 1 \ldots M \tag{1}$$

In any other case, prior application of the general sphericality transformation[1] matrix L^{-1} of Chapter 2, by premultiplication gives transformed variables $L^{-1}X = \hat{\hat{X}}$. (2.5) thus becomes

$$\hat{\hat{X}}D = \hat{\hat{U}} \tag{2}$$

and any least squares or IV estimator of (2) may then also be treated as a sequence of M single equation estimators under (1) with $\hat{\hat{U}}$ for U since $E(\text{vec}(\hat{\hat{U}})\text{vec}(\hat{\hat{U}})') = I_{MT}$. #

We consider the set of all possible replacements of Y_i $i = 1 \ldots M$ as the infinite sequence $Y_i^{(s)}$ $i = 1 \ldots M$, $S = 0, 1, 2, \ldots$

Theorem 1:

Under the given assumptions on (2.5) an IV estimator with $Y_i^{(s)} = ZP_i^{(s)}$ $i = 1 \ldots M$ in the instrument set is consistent iff $P_i^{(s)}$ has rank, at least, r_i.

---

1. i.e. LI methods have L = I, FI methods L = R* ⊗ I, AR methods L=diag(R_i) in the notation of Section 2.5 and AR - FI have L = (r* ⊗ I) diag (R_i) as in Chapter 7 below.

A modified 2SLS estimator using any linear function in the restricted reduced form is consistent iff $ZP_i^{(s)}$ contains a basis for the space spanned by Z_i.

Proof:

By Lemma 1 we replace only $Y_i^{(s)}$ in instrument or regressor sets. For any information class rewrite (2.5) as (2) above. If $Y_i = YS_{1i}$ and $Z_i = ZS_{2i}$ defines S_{1i} S_{2i} and thus S of Theorem 2.11, then (2) becomes, including all specification restrictions,

$$y = (I \otimes \hat{\hat{X}})Sd + U^* \tag{3}$$

where $d = \text{vec}\begin{pmatrix} I & -\beta \\ & \Gamma \end{pmatrix}$ and $U^* = \text{vec}\,(\hat{\hat{U}})$. Then by (2.19) an IV estimator is consistent iff

$$\text{plim}\ T^{-1}\ S'\ (I \otimes \hat{\hat{X}}^{(s)'})U^* = 0 \tag{4}$$

when $\hat{\hat{X}}^{(s)} = (\hat{\hat{Y}}^{(s)} : Z)$ in obvious notation. The RLS estimator, here, is consistent iff

$$\text{plim}\ \hat{d} = \text{plim}\ \{[S'(I\otimes\hat{\hat{X}}^{(s)'}\hat{\hat{X}}^{(s)})S]^{-1}\ S'(I\otimes\hat{\hat{X}}^{(s)'}\hat{\hat{X}})S\}d$$
$$+ \text{plim}\ \{[S'(I\otimes\hat{\hat{X}}^{(s)'}\hat{\hat{X}}^{(s)})S]^{-1}\ .S'(I\otimes\hat{\hat{X}}^{(s)})'U^*\} = d \tag{5}$$

Necessary and sufficient conditions are clearly

$$\text{plim}\ T^{-\frac{1}{2}}\{S'(I \otimes \hat{\hat{X}}^{(s)'}\ (\hat{\hat{V}}^{(s)}\beta + \hat{\hat{U}}))S\} = 0 \tag{6}$$

$$\text{i.e. plim}\ T^{-\frac{1}{2}}\ \{\hat{\hat{X}}_i^{(s)'}\ (\hat{\hat{V}}_i^{(s)}\ \beta_i + \hat{\hat{u}}_i) = 0 \quad i = 1 \ldots M \tag{7}$$

by using the structure of S. Similarly (4) becomes

$$\text{plim } T^{-\frac{1}{2}}\hat{\hat{X}}_i(s)'\hat{\hat{u}}_i = 0 \qquad i = 1 \ldots M \tag{8}$$

By theorem[1] 4.7 Dhrymes (1970), (7) and (8) hold iff

$$\text{plim } T^{-\frac{1}{2}}X_i(s)'(V_i(s)\beta_i + u_i) = 0 \tag{9a}$$

and

$$\text{plim } T^{-\frac{1}{2}}X_i(s)'u_i = 0 \qquad i = 1 \ldots M \tag{9b}$$

Provided $r(P_i(s)) = r_i$ to ensure identification, IV estimators are consistent since assumption (e) of Section 2.2 is sufficient for (9b).

Now (9a) holds iff

$$\text{plim } T^{-\frac{1}{2}}X_i(s)'V_i(s)\,\beta_i = 0 \qquad i = 1 \ldots M \tag{10}$$

However

$$V_i(s) = Y_i - Y_i(s) = (I - ZP_i(s)(P_i(s)'Z'ZP_i(s))^{-1}P_i(s)'Z')Y_i$$

hence

$Y_i(s)'V_i(s) \equiv 0$, and[2] $Z_i'V_i(s) = 0$ iff $ZP_i(s)$ contains a basis for Z_i (say $T_i \neq 0$ exists so that $Z_i = ZP_i(s)T_i$). #

Lemma 2:

Any modified 2SLS estimator that is consistent is identically the IV estimator with instrument $Y_i(s)$. Any RLS estimator is asymptotically equivalent to the same IV estimator iff (10) holds. #

Lemma 3:

A sufficient condition (not necessary as implied by Brundy and Jorgenson (1974a)) for the consistency of any RLS estimator is plim $P_i(s) = \Pi_i$. #

---

1. or Theorems 2.9, 2.10 etc.

2. A special case where this is not true is if plim $T^{-\frac{1}{2}}X_i(s)'V_i(s)\beta_i = 0$ but plim $T^{-\frac{1}{2}}X_i(s)'V_i(s) \neq 0$. To select such a $X_i(s)$ requires a priori knowledge of β_i and is not feasible for a general model. Moreover it has a negligable

Proof:

(10) may be written

$$\text{plim } T^{-\frac{1}{2}}\begin{bmatrix} P_i^{(s)'}Z' \\ Z_i' \end{bmatrix}[Z(\Pi_i - P_i^{(s)}) + Ub_i]\beta_i = 0 \qquad (11)$$

if $b_i = \beta^{-1}S_{1i}$. (11) vanishes if plim $P_i^{(s)} = \Pi_i$ but an infinite number of vectors orthogonal to $\begin{bmatrix} P_i^{(s)'}Z' \\ Z_i' \end{bmatrix}Z$ may be constructed if Π_i were known. #

Comment:

We have considered only replacements $Y_i^{(s)}$ as linear combinations of Z; linear combinations of other variables of the model would be possible but would require a prior knowledge of some parameters and thus is not feasible.

Theorem 2:

IV[1] and modified 2SLS estimators are asymptotically efficient iff Theorem 1 holds and plim $P_i^{(s)} = \Pi_i$.

Proof:

(a) IV estimators; by Theorems 1 and 2.12 (for necessity and sufficiency).

(b) 2SLS estimators; sufficiency by Theorem 1 and part (a) above since they are identically IV estimators if they are consistent; necessity, since otherwise the Cramer-Rao lower bound (by Lemma A2.2, etc.) is not attained and yet it is attained by the IV estimator of part (a). (Recall definition A2.2). #

Corollary 1:

Any consistent IV estimator under the given assumptions is made asymptotically efficient with $Y^{(s+1)} = -Z\Gamma^{(s)}\beta^{(s)-1}$, i.e. $X_i^{(s+1)} = X^{(s+1)}S_i$, as an instrument in one extra step. #

---

1. This theorem, resulting from Theorem 2.12 and (2.19), is a simpler statement of Brundy and Jorgenson (1971) theorems, and those in Brundy and Jorgenson (1974a), although they develop it no further as in Corollaries 1 - 4 here.

Corollary 2:

Assumption (e) of Section 2.2 is evidently only sufficient for Theorem 1. For IV estimators the necessary and sufficient assumption is that (9b) itself holds. However asymptotic efficiency still follows iff Theorem 2 or Coarollary 1 apply. #

Corollary 3:

Assumption (e) of Section 2.2 is a sufficient condition for Theorem 1 in the least squares context. Any RLS estimator using $Y_i^{(s)}$ will be consistent under the necessary and sufficient condition (9a), if this assumption is abandoned. #

Corollary 4:

RLS estimators using $Y^{(s+1)} = -Z\Gamma^{(s)}\beta^{(s)-1}$ in the regressor set (where $\Gamma^{(s)}$, $\beta^{(s)}$ are any consistent estimates of Γ,β) are consistent, asymptotically equivalent to the IV estimator with instrument $Y^{(s+1)}$ and asymptotically efficient.

Proof:

plim $\Gamma^{(s)}$.plim $\beta^{(s)-1} = \Pi$ hence by (11) and Lemma 2 under the assumptions of Section 2.2 the first two properties hold. The third then follows by Theorem 2. #

Comments:

(a) It is in the nature of the estimation problem that the suitable $P_i^{(s)}$, under corollaries 2 and 3 above, must be a priori unknown. Thus it is logically impossible to use (9a) or (9b) as an a priori assumption except as part of the statistical maintained hypothesis. Such an assumption would be with ζ_i, say, replacing $P_i^{(s)}$ and then we must select any $P_i^{(s)}$ such that plim $P_i^{(s)} = \zeta_i$.

(b) Clearly the traditional assumption given in Section 2.2 for the model (2.5), or indeed a single equation regression model, are merely sufficient for consistency. This result is investigated further in Chapter 6.[1]

---

1. Corollaries 2 and 3 contain the GEID and symmetric GEID conditions. This result can also be inferred from Theorem 2.1.

(c) If, with ζ_i of comment (a), the conditions of corollary 2 hold for $\zeta_i = \Pi_i^+$ then 2SLS is also consistent but asymptotically inefficient. It is then also consistent under Corollary 3.

(d) The necessary and sufficient conditions of Corollary 3 on RLS estimators imply a slightly wider class of models can be handled than that indicated by the standard assumption set. But, of that slight generalisation, corollaries 2 and 3 imply only one such specification may be consistently and efficiently estimated. Additionally in all cases these assumptions of zero asymptotic correlations are highly specialised and involve prior knowledge of the parameters to be estimated.

(e) In contrast, provided plim $T^{-\frac{1}{2}}Z_i'u_i = 0$, any model without assumption (e) Section 2.2 can in principle be consistently and efficiently estimated by suitable IV methods. In practice, without prior knowledge of Π_i, it may not be easy because of the problem of selecting $P_i^{(s)}$ orthogonal to $Z'u_i$; unless r_i arbitrary instruments replace Y_i in a preliminary step to give consistent estimates $\hat{u}_i$, and thus $\hat{P}_i^{(s)}$, in a highly arbitrary manner. It is a very much more general situation than in part (c) above.

(f) Theorem 2 implies, for any consistent IV estimator satisfying $Y_i^{(s)} = ZP_i^{(s)}$, plim $P_i^{(s)} = \Pi_i$ is optimal (Theorem 2.12). Thus modified 2SLS estimators are not optimal, unless Theorem 2 happens to hold, but RLS estimators as in Corollary 4 are asymptotically optimal. These conditions are necessary and sufficient by the respective IV interpretations.

4.3 Generalised inverses

Existence problems in 2SLS, 3SLS, LIML, when $K \geq T$, might be overcome by employing a generalised inverse for $(Z'Z)^{-1}$ (see Rao and Mitra (1971)). It is shown by Swamy and Holmes (1971) that $V_i \equiv 0$, $i = 1 \ldots M$, in any k-class estimator (and this proof extends to the systems k-class estimators of (2.48)) in this case. The result poses a problem of interpretation. OLS is inconsistent and therefore unacceptable on statistical criteria.

Equivalence with OLS results not from using generalised inverses but from the sample size. If T > K, any generalised inverse gives BLU estimates of Π (Theorem 7.2.4 Rao and Mitra (1971)). Specifically we seek the solution Π to $Z'Z\Pi = Z'Y$ which, if T < K, has an infinite set of exact solutions (the least squares criterion here is inappropriate). Some other criterion must be employed to select a desirable solution; each one of these is associated with a particular generalised inverse $(Z'Z)^-$ in $\Pi_i^{(s)} = (Z'Z)^-Z'Y_i$. The estimators following are then arbitrary depending on the particular inverse employed. The inverse will imply some prior "knowledge" on Π values - e.g. minimum norm inverses (Rao & Mitra (1971)) pick $\min_{(Z'Z)^-} ||\Pi_i^{(s)}||$ for a given norm - and the estimates have BLU properties conditional on this arbitrary "knowledge", which was not assumed in the model specification. By hypothesis we have no such information since the model is specified in structural form. If it is incorrect, which generally it will be, the estimates will be biased (but of finite variance). This point is raised again in Section 6.

Thus the final estimates will be inefficient and suboptimal in either an RLS or IV framework; and generalised inverses in a k-class estimator are unacceptable. Moreover to use OLS in inadequate samples is inappropriate as it displays no intent to switch to consistent and efficient k-class estimators as soon as T > K. This intent is, as Theil (1973) argues, essential because, although an identical procedure with $Y_i^{(s)} = Z(Z'Z)^-Z'Y_i$ might be said to be consistent,[1] the equivalence of OLS and k-class rests on the asymptotic zero correlation of X_j and u_j. OLS makes no allowance for this nonzero correlation in small samples, whereas other methods do so in finite samples and are therefore preferred.

4.4 Fisher's SOIV method

The SOIV method was proposed by Fisher (1965) and improved by Mitchell and Fisher (1970). They give an algorithm which uses the economic structure

---

1. It is not clear that every generalised inverse contains a basis for Z_i.

of (2.5) to select predetermined variables as instruments, which have the closest causal relationship with the replaced variable, to replace each element of Y_i. A major disadvantage of the method is the massive computational burden since the algorithm requires $\Sigma r_i(\max(K,K-T))$ regressions and significance tests, each with up to max (K,K-T) regressors. However, the selected "instrument" set Z_{oi} is used in the estimator

$$\hat{\delta}_i = (X_{oi}'X_{oi})^{-1}X_{oi}'\, y_i \tag{12}$$

where $X_{oi} = Z_{oi}(Z_{oi}'\, Z_{oi})^{-1}Z_{oi}'\, X_i$. Clearly $\hat{\delta}_i$ of (12) is consistent. But generally it is inefficient (Theorem 2). Although the criterion of availability is satisfied, that of optimality is not since Z_i is replaced (Lemma 1). The same holds if X_{oi} is used as an instrument.

If $Y_i^{(s)} = Z_{oi}(Z_{oi}'\, Z_{oi})^{-1}\, Z_{oi}'\, Y_i$ is the only replacement in the regressor set, (12) is inconsistent (Theorem 1). If it is the only replacement in an instrument set, the estimator is consistent, but generally inefficient and suboptimal. The last sentence also applies to the modification of SOIV that constructs Z_{oi} and Z_i with additional elements selected by the Mitchell-Fisher algorithm and estimates by (12).

Any attempt to select regressor sets for y_j and y_k, where $\beta_{ji} \neq 0$, $\beta_{ki} \neq 0$, which are not identical (as in the Fisher (1965) version) introduces inconsistency in (12) since Theorem 1 implies that the replacement of each element of Y_i must be from identical regressors. This, and any of the modifications above, weakens the causality interpretation in the selection of Z_{oi}. Further y_j in positions $\beta_{ji} \neq 0, \beta_{jk} \neq 0$ will have $Z_{oi} \neq Z_{ok}$ both claiming a close causal relationship. Finally the structural ordering vectors used in the Mitchell-Fisher algorithm reflect closeness of causality in the sense of minimal increments of time separating the reaction of the endogenous variable in a causal chain structure from the movements in a predetermined variable. They do not necessarily also reflect strength of reaction

(explanitory power) because the order, in which predetermined variables are investigated in the stepwise regressions, may influence the final selection Z_{oi}; especially for collinear variables. Any effort expended in seeking the best ordering of equations may influence the block recursive structure and hence the list of available instruments and regressors. In any event the Z_{oi} set selected will have a degree of arbitrariness and only weak causality, and suboptimality.

The SOIV method therefore fails to satisfy at least two of the required criteria, and is not considered further. However Theorem 2 Corollary 1,may be applied to any of its consistent versions. But for the massive computations, it could be used as a start for iterative methods, so from above it is preferable to use an IV estimating version.

4.5 Principle components

Four methods are associated with Kloek and Mennes (1960) and another method has been applied by Evans and Klein (1968). The aim is to select some principle components of Z and use them as regressors on each Y_i, the systematic part of which replaces Y_i in a second step. The replacement may be in a regressor or instrument set.

The methods of Kloek and Mennes select first stage regressor sets as follows: define $\bar{Z}_i$ as the complement of Z_i in Z,

(i) Z_i and the first g principle components of $\bar{Z}_i$.

(ii) Z_i and the first g principle components of the residuals of the regressions of $\bar{Z}_i$ on Z_i.

(iii) Z_i and the g principle components of $\bar{Z}_i$ which have lowest $\bar{R}^2$ when regressed on Z_i.

(iv) Z_i and the g principle components of Z selected as in (iii).

All four methods are consistent by Theorem 1 if $Y_i^{(s)}$ is the systematic part of the regressions of Y_i on the selected regressor sets; as will any method selecting Z_i and other components as regressors for each element of Y_i. By Theorem 2 none of these methods will be efficient; although they satisfy the criterion of availability, they are therefore not optimal.

The Evans-Klein estimator (2SPC), however, uses g principle components of Z as regressors in the first stage, and the resulting $Y_i^{(s)}$ as a regressor in a second stage. By Theorem 1 this procedure is inconsistent if $g < k$. Had $Y_i^{(s)}$ been used as an instrument at the second stage; the estimates (PCIV) would be consistent but inefficient and suboptimal.

By Corollary 1 efficiency, and hence optimality, is attained (considering any consistent method above) in one extra step. However, the methods have major disadvantages. Method (i) is likely to suffer from severe collinearity if only from the first component of $\bar{Z}_i$ being the inevitable trend in economic data which will also be present in Z_i. If so, method (ii) will tend to have residuals of small variance and their components be of little explanitory power for Y_i. Methods (iii) and (iv) involve massive computation strictly the g components should be selected by "optimal" regression described by Coen, Gomme and Kendall (1969); however quicker is stepwise regression where the components are ordered by decreasing $\theta_j = \nu_j(1 - \bar{R}_j^2)$ where ν_j is the eigenvalue in the j^{th} component $\bar{R}_j^2$ comes from the regression of that component on Z_i. The first g elements of that ordering are taken. That implies MK regressions, and principle component estimation, before the estimator is used. Simpler by far is PCIV, where components are ordered by decreasing[1] ν_j, yielding asymptotically similar statistical properties.

---

1. To retain the most explanatory power in Z; ν_j = variance of j^{th} component; Anderson (1958).

4.5.2

The most obvious difficulty is the choice of g; given $r_i \leq g < \min(T,K)$. No rule for choosing a value has been proposed, but unless g = K (when any k-class estimator would be used) the dropping of any particular nonzero components must be arbitrary, as well as the decision on how many to drop. If $Z \sim N(\bar{Z}, \Sigma_Z)$ then a sequence likelihood ratio tests that $\nu_j = 0$, $j > g$, is easy to construct (Theorem 9.3.4 Press (1972)). But we have no such knowledge, and the degree of arbitrariness cannot be removed in any of these methods. Nor can we justify the methods as having negligable arbitrariness when components corresponding to the smallest K-g eigenvalues are ignored, because it is correlations[1] with Y_i, not individual variances, which are really of interest. Y_i may be highly sensitive to a component of very small absolute size or only small variation, but it is exactly that component which will be dropped. Note also consistent methods above have sampling errors proportional to $[(X_i - (V_i^{(s)} : 0))'X_i]^{-1}$ which expodes as $V_i^{(s)} \to Y_i$. Thus, not only are such methods inefficient but their inefficiency increases as $V_i^{(s)}$ increases (i.e. as more components are ignored at the first stage) which may become acute in inadequate samples.

The arbitrary nature of principle component methods, as detailed above, is well known. However they are arbitrary in another sense.

Theorem 3:

If g < K, the g principle components are formally equivalent to g orthonormal factors in the factor analysis of Z, or ζ_i, having maximal individual, and hence generalised, variance.

Proof:

Section 10.7 Press (1972) together with Theorem 11.2.1 Anderson (1958). #

Theorem 3 is in contrast to the use of principle components in place of collinear variables which may be shown to be covariance transforms (Hughes Hallett (1973)). By (2.8) we have

---

1. Any consistent method has an IV interpretation; this remark is implied by (2.21).

$$Y_i = Y_i^* + V_i = (\Lambda F)\ \Pi_i^+ + e + V_i \qquad \text{so } Z = \Lambda F + e \tag{13}$$

which is the standard factor analysis model if Λ is a T x g factor loading matrix and F is a g x r_i matrix of factors with variables measured as deviations from their means. By theorem 3, Λ may be taken as the first g eigenvectors of Z and $F'F = I$. Hence $\Lambda'\Lambda = I_g$. Let Var $(e + V_i) = D_\psi + \text{Var}(V_i)$; so $D_\psi = \Sigma_Z - \Lambda\Lambda'$. The g principle components have maximal variance and hence the remaining K - g minimal variance. In this sense of all orthonormal factors the minimum of information is discarded. But,

Lemma 4:

Any factors, including g principle components, have generalised variance, and sum of component variances, invariant under an orthogonal transform.

Proof:

Anderson (1958) Theorem 11.2.2. Factors are unique only up to an orthogonal transform in (13), and D_ψ is invariant to such transforms. #

Thus principle components is one way of obtaining g unique factors, their variances being strictly ordered. Another way is the varimax principle (e.g. Press (1972)) which chooses that transformation which maximises the sum of the column variances of Λ to make the factors as distinct as possible. This may be preferable as no information is lost and $Y_i^{(s)}$ $i = 1 \ldots M$ will not be dominated by one major component of Z (the trend) to the same extent as when principle components are used. Indeed as we must use some arbitrary rule for discarding components such as sufficiently small θ_j or ν_j there is a real danger $Y_i^{(s)}$ will be sensitive only to the trend and one or two other small movements in Z but independent of other influences on Y_i. Clearly this method does not satisfy the requirement that $Y_i^{(s)}$ be as closely correlated with Y_i as possible. It may be these smaller influences are the important ones to catch. In short there is no economic rationale to the method. Varimax rotations of the components will overcome this difficulty to some extent, and things may be further improved if oblique factors are used. In

any event the principle components are arbitrarily selected in ordering the variances and their arbitrary construction of communalities in the factors since there was no necessity for either. The final estimates are then arbitrary.

The methods considered so far carry the implication that D_ψ shall be minimised as a residual variance. This seems intuitively plausible but is not necessarily optimal if $D_\psi \neq 0$ as Theorem 4 shows. We want factors such that $r_{(Y_i,\ \Lambda FP_i^{(s)})}$ is a maximum in the notation of (2.21).

Theorem 4:

The optimal orthonormal factors are those for which Λ maximises $|Y_i'\Lambda\Lambda'Y_i|$ given $D_\psi = 0$ is not feasible (otherwise $D_\psi = 0$ is optimal). That choice of Λ is also not unique.

Proof:

Dhrymes (1970), Chapter 6, gives the formula for $r_{(Y_i,\ \Lambda FP_i^{(s)})}$ which in this case equals

$$|Y_i'\Lambda FP_i^{(s)}(P_i^{(s)}{}'P_i^{(s)})^{-1}P_i^{(s)}{}'F'\Lambda'Y_i||Y_i'Y_i|^{-1} \tag{14}$$

Also $P_i^{(s)} = F'\Lambda'Y_i$ giving (14) as

$$|Y_i'\Lambda\Lambda'Y_i||Y_i'Y_i|^{-1} \tag{15}$$

Since $\Lambda\Lambda'$, Σ_Z, D_ψ are positive definite (15) implies

$$\max_{D_\psi} |Y_i'\Sigma_Z Y_i - Y'_i D_\psi Y_i| > 0 \tag{16}$$

is required. (16) may be written as

$$\max_{D_\psi} \prod_{j=1}^{r_i} (1 - \mu_j) \tag{17}$$

where $0 \leq \mu_j < 1$ are the eigenvalues of $Y_i'D_\psi Y_i$ in the metric of $Y_i'\Sigma_Z Y_i$. Ideally $\mu_j = 0$ all j but this is possible only if $D_\psi = 0$ (i.e. g = K). If $D_\psi \neq 0$, there is an infinite set of μ_j, and hence D_ψ or Λ, to give any one (including the maximum) value of (16). #

Comment:

An expression for Λ^* to maximise $|Y_i'\Lambda\Lambda'Y_i|$ is not derived. The (ℓ,m) element is $\sum_{k=1}^{T}\sum_{i=1}^{T}\sum_{j=1}^{g}\lambda_{kj}\lambda_{ij}y_{i\ell}y_{km}$ so Λ^* is a function of Y_i. Λ^* should be selected by direct comparisons of r_i - order determinants.

Estimators based on factor analysis are therefore arbitrary in the number of factors employed, also the chosen method (i) - (iv), also the particular factoring method and perhaps the suboptimal criterion selecting the factor set. But for the massive computation, they are useful as starts for iterative estimators. Variables must be standardised first as factors are not otherwise independent of units.

4.6 Prior knowledge methods:

4.6.1

An obvious way to avoid the singularity of Z'Z, when T < K, is to introduce some prior knowledge, or even fictitious data, to supplement Z. We argued earlier it was very unlikely that we should have such information and that the statistical properties of such estimators were conditional on that information. However the approach is useful as some estimators implicitly carry prior knowledge, although none was assumed at the application of the estimator. Prior knowledge may be stochastic restrictions

$$r_i = R_i\delta_i + e_i \qquad (18)$$

and for illustration only we take $E(V_iV_i') = I$ and $E(e_ie_i') = \psi_i$. The best known estimator combining (2.6) and (18) is that of Theil and Goldberger (1960).

$$\hat{\delta}_i = [Z_i^{(s)\prime} Z_i^{(s)} + R_i' \psi_i^{-1} R_i]^{-1} (Z_i^{(s)\prime} y_i + R_i' \psi_i^{-1} r_i) \qquad (19)$$

It is assumed the information (18) is orthogonal to (2.6). The general analysis of these estimators is by Chipman (1964), where in this notation

$$\hat{\delta}_i = [Z_i^{(s)\prime} Z_i^{(s)} + \Omega_i^{-1}]^{-1} [Z_i^{(s)\prime} y_i + \Omega_i^{-1} \bar{\delta}_i] \qquad (20)$$

is shown to have minimum mean square error among all linear estimators, with prior knowledge in the form

$$E(\delta_i) = \bar{\delta}_i, \; E(\delta_i - \bar{\delta}_i)(\delta_i - \bar{\delta}_i)' = \Omega_i, \qquad (20)$$

is asymptotically BLU conditional on the prior knowledge- any bias depends on δ_i and the extent that the knowledge is wrong. We have a range of estimators[1] depending on the weight given to the prior knwoeldge - ignoring it implies $\Omega_i \to \infty$. The prior knowledge is equivalent in (19) and (20) whenever $R_i' \psi_i^{-1} R_i = \Omega_i^{-1}$ and $R_i' \psi_i^{-1} r_i = \Omega_i^{-1} \bar{\delta}_i$.

The problem is to obtain $Z_j^{(s)}$ for (19) or (20). Using generalised inverses for $P_i^{(s)} = (Z'Z)^{-} Z'Y_i$ implies prior knowledge on Π_i. Here $\Omega_i^{-1} = [T\Lambda^{-1}T' - D]^{-1} - Z'Z$ and $([T\Lambda^{-1}T' - D]^{-1} - Z'Z)\bar{\delta} = 0$ where $T\Lambda T'$ is the similarity transform on $Z'Z$, and $D + (Z'Z)^{-} = T\Lambda^{-1}T'$. In our case $\lambda_i = 0 \quad i = T + 1, \ldots, K$ so $D = 0$. Here the generalised inverse confines estimates to the linear subspace of dimension T about the origin (Theorem 9 Marquardt (1970)), i.e. they are constrained to include $\bar{\Pi}_{ij} = 0 \quad j = T+1, \ldots K$ and $\Omega_i = \begin{pmatrix} \infty & 0 \\ 0 & 0 \end{pmatrix}$. This prior knowledge contradicts the model specification

---

1. Ridge regression belongs to this family, taking $k_i I = \Omega_i^{-1}$ for some scalar $k_i \neq 0$, and thus $\bar{\delta}_i = 0$. In the face of such prior knowledge the decision to estimate seems unjustified. Any nonzero results must be biased. It is therefore ignored.

that all K variables Z occur in the model, and therefore implies biased estimates when the model is correctly specified.[1] Hence the comments of Section 3.

4.6.2

An ingeneous method is presented by Theil (1973). Premultiplying (2.6) by Z_i' and $\bar{Z}_i'$

$$Z_i'y_i = Z_i'X_i\,\delta_i + Z_i'u_i \tag{21a}$$

$$\bar{Z}_i'y_i = \bar{Z}_i'X_i\delta_i + \bar{Z}_i'u_i \tag{21b}$$

2SLSis obtained by applying GLS to (21a-b) and this is possible iff $T > K - s_i$. If not, however, 2SLS normal equations contain the relation (the second relation in (2.4)).

$$Z_i'y_i = Z_i'X_i\hat{\delta}_i \tag{22}$$

so 2SLS.could also be approximated by application of GLS to (21b) under the restriction (22). If λ_i are s_i lagrangian multipliers, estimates are given by

$$\begin{pmatrix} X_i'C_iX_i & X_i'Z_i \\ Z_i'X_i & 0 \end{pmatrix} \begin{pmatrix} \hat{\delta}_i \\ \lambda_i \end{pmatrix} = \begin{pmatrix} X_i'C_iy_i \\ Z_i'y_i \end{pmatrix} \tag{23}$$

where $C_i = \bar{Z}_i(\bar{Z}_i'\bar{Z}_i)^{-1}\bar{Z}_i'$. Note (23) does not yield 2SLS estimates since (23) compared with (2.24) indicates $P_i^{(s)} = \Pi_{2i}^{+}$ where $\Pi_i^{+'} = (\Pi_{1i}^{+'} : \Pi_{2i}^{+'})$ is conformable with Z. Thus the procedure is consistent in an IV context (indeed (22) constrains it to be so; Theorem 1 does not apply) but inefficient

---

1. In contrast to using generalised inverses when Z contains near perfect collinearity. For K - s collinear variables $D = T_{K-s}\Lambda^{-1}_{K-s}\,T'_{K-s} \neq 0$ and prior information is induced, similarly to above, to counteract the unsatisfactory information. The estimates are biased but probably with reduced mean

and suboptimal (Theorems 1,2). Theil proves that, whenever the equation is identified and $T > r_i + 2s_i$, a solution will exist. Naturally when $T < K - s_i$, $\bar{Z}_i'\bar{Z}_i$ is singular, but replacing that with any arbitrary nonsingular D_i permits estimation with no change in the properties. Any approximation will suffice, but $D_i = \text{diag} \sum_t (\bar{Z}_{tj})_i^2$ is particularly simple.

Therefore we have consistent but inefficient and suboptimal estimates, which nearly satisfy the availability criterion, but with two elements of arbitrariness:

(a) the selection of D_i,

(b) the approxmation of constrained GLS on (21b) alone.

Solving for $\hat{\delta}_i$ yields

$$\hat{\delta}_i = (X_i'C_iX_i)^{-1}X_i'\{C_i+Z_i[Z_i'X_i(X_i'C_iX_i)^{-1}X_i'Z_i]^{-1}Z_i'[I-X_i(X_i'C_iX_i)^{-1}X_i'C_i]\}y_i \tag{24}$$

which is identical to (20) with implicit prior information

$$\bar{\delta}_i = (X_i'C_iX_i)^{-1}X_i'C_iy_i$$

and

$$\Omega_i = \left[X_i'Z \begin{pmatrix} A_i^{-1}Z_i'Z_iA_i^{-1} & 0 \\ 0 & -D_i^{-1} \end{pmatrix} Z'X_i\right]^{-1}$$

$$\text{if} \quad A_i = Z_i'X_i(X_i'C_iX_i)^{-1}X_i'Z_i \tag{25}$$

Note that consistency holds since $E(\bar{\delta}_i) = \delta_i$. The two sources of arbitrariness are clear in (25) i.e. D_i, and the use of prior "knowledge" variances arbitrarily selected and imposed. The legitimacy of the procedure depends on the selection of D_i to give a plausible Ω_i, which indicates how difficult it is to use the method sensibly. (23) now has three interpretations:

(i) a restricted 2SLS approximation,

(ii) RLS incorporating prior "knowledge" (25) and $Y_i^{(s)}=Z_iA_i^{-1}Z_i'X_i$(hence

consistency but inefficiency by Theorems 1 and 2),

(iii) an IV estimator with instrument set $N_i'=(X_i'C_i,Z_i')$ where any set of s_i irrelevant exogenous variables orthogonal to Z_i are included in specification (2.6) - hence the availability constraint.

4.6.3

A much simpler estimator with one less source of arbitrariness and which exactly satisfies the availability criterion may be obtained from (21). A Theil-Goldberger estimate treating (21a) as the model, and (21b) as prior knowledge gives 2SLS. So using D_i exactly as above we get

$$\hat{\delta}_i = \left[X_i'Z\begin{pmatrix}(Z_i'Z_i)^{-1} & 0\\ 0 & D_i^{-1}\end{pmatrix}Z'X_i\right]^{-1}\left[X_i'Z\begin{pmatrix}(Z_i'Z_i)^{-1} & 0\\ 0 & D_i^{-1}\end{pmatrix}Z'y_i\right] \qquad (26)$$

(26) is preferable to (23). It is markedly simpler, and quicker to calculate. It involves no approximation to 2SLS and hence is only arbitrary at one point; in the selection of $D_i \neq \bar{Z}_i'\bar{Z}_i$. It is clearly unconditionally consistent in view of (21b) as prior knowledge, but will be efficient and optimal only if

$$\text{plim } Y_i'\bar{Z}_iD_i^{-1} = \Pi_{2j}' \qquad (27)$$

dictates the choice of D_i. Evidently, although it is much easier to select D_i to give plausible ψ_i or Ω_i, and with a higher chance of obtaining at least near optimal estimates, it is unlikely that (27) holds exactly. (26), as well as (23), may of course be used as a start for iterative estimators. Efficiency follows in one extra step (Corollary 1). (26) therefore recommends itself by its simplicity.

4.7 Iterative Estimation:

4.7.1

Iterative estimators for (2.5) were introduced by Wold (1965), Mosbaek and

Wold (1970) and Lyttkens (1973). The simple motivation, and the only one used to date, is to observe that in (2.30) if Y* were known a priori then, by Theorem (2.1), using $Y_i^{(s)} = Y_i^*$ implies consistent estimates are obtainable from

$$\hat{\delta}_i = (X_i^{*\prime}X_i^*)^{-1}X_i^{*\prime}y_i \text{ and } \hat{\delta} = (\hat{X}^{*\prime}\hat{X}^*)^{-1}\hat{X}^{*\prime}y \qquad (27)$$

Theorem 5:

(27) yields consistent and asymptotically efficient estimates satisfying the availability and optimality criteria.

Proof:

Either

(a) Since (2.30) implies a predictor specification $E(Y|X) = Y^*\beta + Z\Gamma$ and given $Y^* = -Z\Gamma\beta^{-1}$, by our assumptions plim T^{-1} $X^{*\prime}V = 0$ so ergodicity and part (c) of Theorem 2.1 hold. Moreover estimates using Y* are efficient by Corollary 4. Thus optimality follows (clearly also availability),

or

(b) Consistency by (10) and (11) since Y* is used; thus efficiency and optimality by Lemma 2 and Theorem 2. #

In any event, the assumptions of Section 2.2 may be substantially relaxed in that X_i replaces (X_i,Z), i = 1 .... M, in assumption (g); and plim $T^{-1}X_i^{*\prime}(V_i\beta_i + U_i) = 0$ i = 1 .... M can replace assumption (e) with no damage to the properties quoted.

Consistent estimates may also be obtained from (2.5) and (2.6) directly, without the REID specification, if Y* is known:

$$\tilde{\delta}_i = (X_i^{*\prime}X_i)^{-1}X_i^{*\prime}y_i \text{ and } \tilde{\delta} = (\hat{X}^{*\prime}\hat{X})^{-1}\hat{X}^{*\prime}y \qquad (28)$$

Theorem 6:

Theorem 5 holds for (28) also.

Proof:

Consistency by Theorem 1; efficiency by Theorem 2; optimality by Theorem 2.12. Availability is obvious. #

It is sometimes suggested that not only should Y_i be replaced by Y_i^* in the regressor or instrumental sets but also any lagged values of Y by corresponding lagged values of Y* (Fisk (1973)). However this should not be done since it induces a composite error term in (2.30) of $V_i\beta_i + V_{-1}\Gamma_i + u_i$ where Γ_i is the i^{th} column of Γ. This error term is certainly serially correlated if U components are not. It is only not serially correlated in the most unlikely circumstances and in view of the lagged dependent variables, this procedure enormously complicates the estimation problem for no good reason. In the IIV case this replacement is suboptimal by Lemma 1.

By corollary 2 we can relax the assumptions to plim $T^{-1}X_i^{*\prime}u_i = 0$ i=1....M for assumption (e) of Section 2.2. Although (28) unlike (27) does not require the model to be respecified in the REID form, they both require Y* which is a T x M matrix satisfying

$$Y^*\beta + Z\Gamma = 0 \tag{29}$$

The problem therefore is to specify algorithms which estimate Y*, which will in fact not be known, together with $\hat{\delta}$ or $\tilde{\delta}$. The solution of an equation system, highly nonlinear in unknowns, is therefore required.

4.7.2

Equation (29) suggests that the estimates (27) or (28) be obtained by the limit of the sequences, either

$$\hat{\delta}^{(s+1)} = (\hat{X}^{(s)\prime}\hat{X}^{(s)})^{-1}\hat{X}^{(s)\prime}y \text{ where } Y^{(s)} = Y^{(s-1)}(I-\hat{\beta}^{(s)}) - Z\hat{\Gamma}^{(s)} \tag{30a-b}$$

$$\text{or } \tilde{\delta}^{(s+1)} = (\hat{X}^{(s)\prime}\hat{X})^{-1}\hat{X}^{(s)\prime}y \text{ and } Y^{(s)} = Y^{(s-1)}(I-\tilde{\beta}^{(s)}) - Z\tilde{\Gamma}^{(s)};\ S=0,1,.... \tag{31a-b}$$

If the sequence $Y^{(s)}$ converges under both (30) and (31) such that $\lim_{s\to\infty} Y^{(s)}=Y^*$ then $\lim_{s\to\infty} \hat{\delta}^{(s+1)} = \hat{\delta}$ and $\lim_{s\to\infty} \tilde{\delta}^{(s+1)} = \tilde{\delta}$. In either case a start $Y^{(o)}$ must be specified[1]; and in practice at the final stage interaction stops when

$$\left|\frac{\hat{\delta}_{ij}^{(s+1)} - \hat{\delta}_{ij}^{(s)}}{\hat{\delta}_{ij}^{(s)}}\right| < \tau_1 \text{ or } \left|\frac{\tilde{\delta}_{ij}^{(s+1)} - \tilde{\delta}_{ij}^{(s)}}{\tilde{\delta}_{ij}^{(s)}}\right| < \tau_1 \qquad (32)$$

$$\text{all } j = 1 \ldots r_i+s_i,\ i = 1 \ldots M$$

for some arbitrarily small tolerance τ_1 usually taken in the interval $[10^{-6}, 10^{-3}]$. Equivalently (see Chapter 12) we may consider

$$\left|\frac{y_{ij}^{(s+1)} - y_{ij}^{(s)}}{y_{ij}^{(s)}}\right| < \tau_1 \text{ all } j = 1 \ldots r_i,\ i = 1 \ldots M \qquad (33)$$

Both (30) and (31) represent infinite iterations towards the fixed point (29) from an arbitrary start $Y^{(o)}$. The iterations become finite by accepting the arbitrarily good approximation to (27) or (28) specified by τ_1. To this extent any statistical properties quoted below, and optimality in the sense of RLS or IV estimators, may only be arbitrarily well aproximated. We therefore distinguish two possibilities:

Definition 1:

A virtual statistical property[2] is said to hold, if that property holds exactly for (27) or (28), i.e. only holds exactly for $\lim_{\tau_1\to 0} \hat{\delta}^{(s+1)}$, or $\lim_{\tau_1\to 0} \tilde{\delta}^{(s+1)}$ but not for any arbitrarily small $\tau_1 > 0$.

The distinction between virtual and actual statistical properties turns out largely to be a matter of the arbitrary start $Y^{(o)}$ and is examined in Chapter 11. The sensitivity of the results to choosing different suitable $Y^{(o)}$ values is of some importance.

---

1. Evidently the remarks above on replacing the lagged dependent variables by $Y_{-1}^{(s)}$ apply word for word for all previous iterates.

2. e.g. Consistency, efficiency, optimality, equivalence with other estimators, independence of starts $Y^{(o)}$, etc.

Theorem 7:

If Y* is unique, then $\hat{\delta}^{(s+1)}$ and $\tilde{\delta}^{(s+1)}$ are invariant to arbitrary $Y^{(o)}$ only at convergence. Otherwise all $\hat{\delta}^{(s+1)}$ and $\tilde{\delta}^{(s+1)}$ are functions of arbitrary starts.

Proof:

in (30b), (31b) noting $Y^{(s)} = Z\Pi^{(s)}$; $\Pi^{(s)} = \Gamma^{(s)} \beta^{(s)-1}$; $\Gamma^{(s)}$; $\beta^{(s)}$ from either[1] $\hat{\delta}^{(s)}$ or $\tilde{\delta}^{(s)}$,

$$Z[\Pi^{(s)} - \Pi^{(s-1)}(I - \beta^{(s)}) - \Gamma^{(s)}] = 0 \tag{34}$$

By (30a) or (31a) we can then write

$$\Pi^{(s)} = g(\Pi^{(s-1)}) \tag{35}$$

where $g(\cdot)$ is the implied nonlinear functional. The fixed point is reached either as (29) or where

$$\Pi^{(k)} = g(\Pi^{(k)}) \tag{35a}$$

for some s = k. Hence the theorem follows if Y* is unique and convergence is observed since the contraction mapping theorem of Wold (1965) is sufficient to guarantee that if convergence is observed it is to some Y*, with obvious consequences if Y* is unique. In any other case with or without convergence, estimates remain arbitrary in the sense of $Y^{(o)}$. #

Corollary 5:

The uniqueness of Y* may be tested, or its various possible values estimated, by letting $Y^{(o)}$ range over the feasible endogenous variable space. #

---

1. The absence of "^" or "~" hereafter indicates either may be considered.

What is "feasible" is a matter of judgment, but it must include all values around Y, but not too distant, since if the estimates of β are to be of any significance the noise to signal ratio in Y should be relatively small.

Corollary 6:

(35) may have more than one solution with real $\Pi^{(k)}$ (complex values may be ignored in linear econometric models but not necessarily in other contexts). Hence the uniqueness of Y* and $\hat{\delta}^{(k)}$, $\tilde{\delta}^{(k)}$ are not guaranteed as it was in some previous estimators. #

Corollary 7:

Convergence for (35) is also not guaranteed since by Theorem 12.1 it is only locally convergent if $\rho\left(g'(\Pi^{(k)})\right) < 1$ for an arbitrary start $\Pi^{(o)}$ within a neighbourhood of $\Pi^{(k)}$, the fixed point; where $g'(\cdot) = \partial\Pi^{(s)}/\partial\Pi^{(s-1)}$. $g'(\cdot)$ depends only on the true parameters β,Γ at a fixed point Y*. #

Comments:

(a) Theorem 5 is a virtual property when $\tau_1 \neq 0$.

(b) If Y* is not unique, but convergence is observed we may say estimates are locally independent of $Y^{(o)}$ within some neighbourhood about Y*.

(c) If $\beta^{(s)}$, $\Gamma^{(s)}$ were fixed with respect to (s) then (30b), (31b) and (34) reduce to Jacobi's method for solving equation systems.

(d) There is no consideration of the speed of convergence.

4.7.3

(30b) and (31b) were chosen by similarity to (29) for estimating Y*. They are not the only possibilities. However in selecting other methods it is important not to violate the properties of (27) or (28) when Y* has been obtained, nor lose the availability criterion en route which is preserved by (30b), (31b). (30a-b) is known as the Fixed Point (FP) estimator and (31a-b) the Iterative Instrumental Variable (IIV) estimator. We designate any estimator that preserves (30a) as a member of the FP family, and any that uses (31a) as a member of the IIV family. The following have been proposed.

By comment (c) above, we could solve for $Y^{(s)}$ by direct elimination (Lyttkens (1973)). That is replace (30b), (31b) by

$$Y^{(s)} = -Z\Gamma^{(s)} \beta^{(s)-1} \tag{36}$$

Applying (36) with (30a) and (31a) we get reduced form methods[1] (RFFP, RFIIV) or we could take the iteration in the bracketed term in (34) to be zero, rather than orthogonal to Z giving parametric methods (PFP, PIIV, RFPFP, RFPIIV) (Lyttkens 1973, Bergstrom 1974) according to whether Jacobi or direct elimination methods are used.

It is well known iterations may be accelerated by obtaining $Y^{(s)}$ by a suitable method and then calculating $Y_\alpha^{(s)} = \alpha Y^{(s)} + (1 - \alpha) Y_\alpha^{(s-1)}$. If $Y_\alpha^{(s)}$ is used for $Y^{(s)}$ in (30a-b) we have fractional FP or in (31a-b) fractional IIV(FFP,FIIV), with (Agren 1972)

$$Y_\alpha^{(s)} = \alpha(Y^{(s-1)}(I - \beta^{(s)}) - Z\Gamma^{(s)}) + (1 - \alpha) Y_\alpha^{(s-1)} \tag{37}$$

Another acceleration is

$$y_{i,\alpha}^{(s)} = \alpha \sum_{j=1}^{i-1} y_{j,\alpha}^{(s)} \beta_{ji}^{(s+1)} + \alpha \sum_{j=i+1}^{M} y_{j,\alpha}^{(s-1)} \beta_{ji}^{(s)} + (1-\alpha) y_{i,\alpha}^{(s-1)} + \alpha Z_i \gamma_i^{(s)} \tag{38}$$

which gives recursive methods (RFP, RIIV in (30a) or (30b)). (38) implies reestimating $\beta_{ji}^{(s)}$ and $\gamma_i^{(s)}$, (with y_{ji} replaced using $y_{ji,\alpha}^{(s)}$ if $j < i$ but $y_{ji,\alpha}^{(s-1)}$ if $j > i$) as soon as $y_{i,\alpha}^{(s)}$ is available but before $y_{i+1,\alpha}^{(s)}$ is. If this is not done, substituting $\hat{\beta}_{ji}^{(s)}$ in the first term gives R_1FP (Bodin 1974). If $\beta^{(s)}, \Gamma^{(s)}$ were constant (37) is the JOR method, and (38) the SOR method, of equation system solution (Chapter 9); so $\alpha = 1$ in (37) gives Jacobi's method but in (38) gives Gauss-Seidel. Such improvements and many other possibilities are considered in Chapter 11.

---

1. There is much confusion on naming iterative estimators. Therefore in Chapter 11 I establish a convention. Comparison with Bergstrom (1974) reveal RFIIV identical to his IIV, FIIV to his SIIV and RIIV to his RSIIV.

4.7.4

The unsatisfactory nature of the iterative theory is obvious because of its incompleteness and Corollaries 6 and 7. The properties of FP and IIV depend on the assumption that $\lim_{s\to\infty} Y^{(s)} = Y^*$. Properties of RLS estimators not defined by Corollary 4 are unknown. But in addition to Theorems 5 and 6 we have

Theorem 8:

(a) $\hat{\delta}^{(s)}$ of (27) with consistent[1] $Y^{(o)}$ starts, is consistent.

(b) RFFP, not FP, is asymptotically efficient, optimal and available if $T > \max_i (r_i + s_i)$ all finite $S \geq 2$.

Proof:

(a) immediate from (11).

(b) Theorem 1 and Corollary 4. #

Theorem 9:

All IIV estimates are consistent for consistent $Y^{(o)}$ all $S \geq 1$. RFIIV, not IIV, is efficient and optimal all finite $S \geq 2$.

Proof:

(a) Regard each $Y_i^{(s)}$ from any algorithm in the IIV family as a separate instrument set in an IV estimator; then Theorem 1 for a subsequence of $Y_i^{(s)}$. #

(b) Theorem 2, Corollary 1. #

We need to investigate the properties of all iterative estimators when $Y^{(o)}$ is inconsistent or Y^* not unique. We have no knowledge of the efficiency of earlier iterates of FP and IIV even if $Y^{(o)}$ is consistent and Y^* unique; nor any properties of earlier iterates of FFP, RFP, etc. The properties of any iterative estimator under relaxed assumptions in Corollaries 2 and 3 are investigated in Chapter 6. These results are important since in practice all results are approximate because $\tau_1 \neq 0$. It is important to know how often virtual properties apply (and how much they differ from the actual). Tests for nonconvergence (and properties in that case) would be useful.

---

1. i.e. $Y^{(o)}$ in (30a) such that plim $\hat{\delta}^{(1)} = \delta$; in IIV estimators $Y^{(o)}$ in (31a) such that plim $\tilde{\delta}^{(1)} = \delta$.

If a consistent start[1] $Y^{(o)}$ was used iteration to convergence appears only to be justified by removing any arbitrariness since consistency, and with (36) efficiency and optimality follow (Theorem 1 and Corollary 1) for $S \geq 2$ for (30a) and (31a). Since a consistent but inefficient $Y^{(o)}$ is much easier to find for (31a) than (30a) (for example various methods of sections 4 - 6 above, or even a set of dummy variables) this has led to the 2 and 3 step IV estimators of Lyttkens (1970), Dhrymes (1971) and Brundy and Jorgenson (1971) all of which follow the pattern of (31a) and (36).

We need to know if plural solutions[2] imply different properties and how to select the most desirable solution. It is not clear what relationship iterative estimates have to conventional estimators, nor what formula should be used for their standard errors, nor if different algorithms for estimating Y* affect the statistical properties. It would be useful to guarantee convergence for any $Y^{(o)}$ and also to minimise the computational burden.

4.8 Summary:

In Sections 4 - 6, we obtained several estimators which are unconditionally consistent but inefficient and satisfy the criterion of availability. Many of them require substantial computation; the exception is (26). Any unconditionally consistent estimator may be made asymptotically efficient and optimal in one extra step. That extra step on (26) appears a very simple procedure to satisfy all the tests we proposed in Chapter 3.

In Section 7 we noted some iterative estimators. As Y* has to be estimated iteratively there could be substantial computation. The statistical properties of iterative estimators are only valid from their very tenuous connection with (27) and (28) by virtue of Wolds (1965) contraction mapping theorem if also (a) $\lim_{s\to\infty} Y^{(s)} = Y^*$ and (b) Y* is unique. Thus all the statistical properties must be reworked for a general case and finite iterations. An advantage however is the possible relaxation of assumptions.

---

1. i.e. $Y^{(o)}$ in (30a) or (31a) such that plim $\hat{\delta}^{(1)} = \delta$ or plim $\tilde{\delta}^{(1)} = \delta$.
2. Plural solutions imply two distinct values satisfy (29); multiple solutions imply repeated roots to (35a)

Estimators corresponding to those of this chapter but which are efficient and optimal in the three information classes which dominate limited information are presented in Chapter 7. Estimators of Section 6 have an obvious connection with Bayesian estimation theory.

CHAPTER 5: A MAXIMUM LIKELIHOOD INTERPRETATION OF ITERATIVE ESTIMATORS

5.1 LIML as an Instrumental Variable Estimator

The equation (2.41) shows that the FIML estimator may be thought of as an instrumental variable estimator which is identical to the full information version of IIV, given by (4.28).

Equation (2.41) is nonlinear in β,Γ and would be solved by iteration. At the k^{th} iteration

$$\hat{\delta}_{(k+1)} = (N'_{(k)}\hat{X})^{-1}N'_{(k)}y; \quad N_{(k)} = [(\Sigma_{(k)} \otimes I)^{-1}\hat{X}_{(k)}];$$

$$\Sigma_{(k)} = T^{-1}(Y\beta_{(k)} + Z\Gamma_{(k)})'(Y\beta_{(k)} + Z\Gamma_{(k)}) \qquad (1)$$

For FIML the log likelihood function was written (assuming throughout no restrictions on Σ) as (2.35). We maximised the likelihood function with respect to Σ and the unknown elements of β, Γ, substituted Σ out from the three necessary conditions obtained, and obtained (2.41) from which estimates of β,Γ,Σ may be obtained. It may also be shown that LIML is an instrumental variable estimator.

LIML estimates are obtained from the same likelihood function transformed so the parameters of an equation of interest are expressed in terms of those of other equations. The likelihood function is concentrated with respect to those not of interest and then maximised with respect to those that are.

We follow this procedure with exactly the same steps used to obtain (1). Upon implementing the transformation in the necessary conditions the estimating equation for the parameters of the equation of interest show a result of the form of (2.40). The same transformation matrix as in text book expositions of LIML (here in Dhrymes (1970) notation) is used. Rewriting (2.35) in transformed parameters

$$L^*(D,H,\Sigma) = C + {}^{T}/_{2} \log \det (H'\Sigma H)^{-1} + T \log \left| \det (\beta H) \right|$$

$$- {}^{T}/_{2} \operatorname{tr} T^{-1}(H'\Sigma H)^{-1}(XDH)'(XDH) \qquad (2)$$

where $H = \begin{bmatrix} I & -\Sigma_{11}^{-1}\Sigma_{12} H_{22} \\ 0 & H_{22} \end{bmatrix}$, (3)

the transformation matrix used by Dhrymes. Subscript 1 refers to parameters of interest (say the first equation or subset of equations) and subscript 2 to those not of interest. Thus

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad \text{and} \quad DH = \begin{bmatrix} \beta_1 & \beta_2 \\ \Gamma_1 & \Gamma_2 \end{bmatrix} \quad \text{etc.} \qquad (4)$$

In obtaining the likelihood function of the transformed system (2.5), we transformed the errors such that Cov $(U_{.t}H, U_{.S}H) = H'E(U_{.t}'U_{.S})H = H'\Sigma H$. Using (3) and the symmetry of Σ we have

$$H'\Sigma H = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & Q \end{bmatrix} \quad \text{where } Q = H_{22}'[\Sigma_{12}'\Sigma_{12} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}] H_{22} + I \qquad (5)$$

Rewriting (2) again with $\Sigma^* = H'\Sigma H$, $\beta^* = \beta H$, $\Gamma^* = \Gamma H$, $D^* = DH$, maximising the result with respect to all parameters and substituting for Σ^* we get necessary conditions

$$\begin{pmatrix} -Z^{*\prime} \\ (\beta^{*\prime})^{-1}\ \Gamma^{*\prime}Z' \end{pmatrix} \quad XD^*\Sigma^{*-1} = 0 \qquad (6)$$

where $Z^* = ZH$ and H^{-1} exists, again following Dhrymes.

Upon sorting out, using the structure of H and since Z is conformable with Γ, we get

$$\begin{bmatrix} -Z'_1 & -Z'_2 \\ [(\beta')^{-1}\Gamma']_1 Z' & [(\beta')^{-1}\Gamma']_2 Z' \end{bmatrix} \begin{bmatrix} Y\beta_1 \Sigma_{11}^{-1} & + & Z\Gamma_1 \Sigma_{11}^{-1} \\ Y\beta_2 Q^{-1} & + & Z\Gamma_2 Q^{-1} \end{bmatrix} = 0 \quad (7)$$

Discarding the second equation, since the first involves only Γ_1 and β_1, we have necessary conditions from which parameter estimates may be obtained. Now β_1 Γ_1 contain the zero restrictions within them so, if we are concerned only with the first equation, we have

$$\begin{bmatrix} -Z_1' \\ [(\beta')^{-1}\Gamma']_1 Z' \end{bmatrix} (Y_1\beta_1 + Z_1\gamma_1)\, \sigma_{11}^{-1} = 0 \quad (8)$$

Normalising, we may put this in instrumental variable estimator form

$$\hat{\delta}_1 = (N_1'X_1)^{-1} N_1'y_1 \quad (9)$$

where $N_1 = [Z(\Gamma\beta^{-1})_1 \vdots Z_1]$ and $X_1 = [Y_1 \vdots Z_1]$

(9) is nonlinear in the parameters β_1 and Γ_1 and could therefore be solved iteratively - perhaps as Durbin's (1963) method if no knowledge of β or Γ is obtained from elsewhere. Alternatively iterations may be performed with starts from elsewhere (e.g. OLS or 2SLS). $(\Gamma\beta^{-1})_1$ are the restricted reduced form coefficients of the RHS endogenous variables since $\Gamma\beta^{-1}$ was partitioned conformably with Z_1, Z_2. The whole system would be obtained from

$$\hat{\delta}_{LIML} = (\hat{N}'\hat{X})^{-1}\hat{N}'y \quad (10)$$

where $\hat{N} = \text{diag}\ (N_i)$.

(10), as an alternative to the representation of LIML involving M minimal eigenvalues and associated eigenvectors, may be simpler analytically

and computationally. Further, the LIML estimator is independent of the structural error variance as well as their covariances.

The log likelihood functions (5) or (7) require the existence of

$$(D'X'XD)^{-1} = T\Sigma^{-1} \qquad (11)$$

i.e. $T > K + M$ is required for the existence of both FIML and LIML. However (2.41) and (10) are available provided only $T > \max \{\max_i (r_i + s_i), M\}$ and $T > \max_i (r_i + s_i)$ respectively.

5.2 The asymptotic equivalence of iterative and maximum likelihood estimators

It is obvious that any simultaneous estimator may now be ranked asymptotically equivalent to LIML if (a) it may be shown to have an instrumental variable interpretation and (b) that instrument set is asymptotically equivalent to $\hat{N}$ or (10). Equally an estimator will be asymptotically equivalent to FIML if it has an instrumental variable interpretation with instruments asymptotically equal to N of (2.41).[1]

Expressing estimators in terms of instruments containing explicit estimates of reduced form parameters makes the proof of asymptotic relations between estimators particularly straightforward. The ranking of estimators by asymptotic efficiency or as approximations to FIML is correspondingly clearer in this scheme than those of the past. Although recent surveys have considered the classification of a number of estimators (Hausman (1975), Brundy and Jorgenson (1974) and Hendry (1976)) there has been no consideration of limited information methods nor of repeated least squares, nor of FP family estimators, (see section 7 below).

Fixed point and its variants fit directly into this classification. Under the assumption here, plim $T^{-1}(Z'U) = 0$, FP has an IV interpretation, and is asymptotically equivalent to LIML. In probability limits

---

[1] We refer of course to theorem 2.12 i.e. that plim $P_i = (\Gamma\beta^{-1})_i$ $i = 1 \ldots M$ and (1) and (10).

$$\hat{\delta}_{FP} = (\hat{X}^{*\prime}\hat{X}^{*})^{-1}\hat{X}^{*\prime}y = (\hat{X}^{*\prime}\hat{X})^{-1}\hat{X}^{*\prime}y = \hat{\delta}_{IIV} \tag{12}$$

because $Y = -Z\Gamma\beta^{-1} + U\beta^{-1} = Y^{*} + V$ at the final stage, so regressors $X_i^{*} = [Y_i^{*} : Z_i]$ imply $X_i^{*\prime}X_i^{*} = X_i^{*\prime}X_i - X_i^{*\prime}U\beta_i^{-1} = X_i^{*\prime}X_i$ in probability limits. Nonlinear iteration is required to estimate both $\hat{\delta}_{FP}$ and the instrument $\hat{X}^{*} = \text{diag}\,[-Z(\Gamma\beta)^{-1}_i : Z_i]$ in (12). Since plim $\pi^{+} = \pi = -$ plim $\hat{\Gamma}_{FP}$plim $\hat{\beta}_{FP}^{-1}$, FP is asymptotically equivalent to 2SLS and further both may be ranked equal to LIML for asymptotic efficiency by theorem 2.12 and (10) above. (The correct expression for asymptotic variance of FP estimates if that of 2SLS).

A full information version of FP, FIFP, may be constructed. One might carry FP to convergence in the form

$$\hat{\delta}_{FIFP} = (\hat{X}^{*\prime}\hat{\phi}^{-1}\hat{X}^{*})^{-1}\hat{X}^{*\prime}\hat{\phi}^{-1}y \quad ; \quad \hat{\phi} = \hat{\Sigma}\otimes I \tag{13}$$

$$\hat{\Sigma} = T^{-1}(Y\hat{\beta} + Z\hat{\Gamma})'(Y\hat{\beta} + Z\hat{\Gamma}).$$

Clearly, as with (12)

$$(\hat{X}^{*\prime}\hat{\phi}^{-1}\hat{X}^{*}) = (\hat{X}^{*\prime}\hat{\phi}^{-1}\hat{X}) \tag{14}$$

in probability limits. So FIFP is an IV estimator with instrument set $(\hat{\Sigma} \otimes I)^{-1}\hat{X}$, where $\hat{X}^{*} = \text{diag}\,[-Z(\Gamma\beta^{-1})_i : Z_i]$. This establishes consistency and efficiency of $\hat{\delta}_{FIFP}$. FIFP is asymptotically equivalent to, and may be ranked with 3SLS as well as with FIML and (2.41) for asymptotic efficiency.

FP is but one family of iterative estimators. In view of (12) or (13) the IIV estimator is asymptotically equivalent to FP and of course IIV family members may also be classified by the above analysis.

5.3 An identity between iterative and maximum likelihood estimators

FP and FIFP are therefore asymptotically equivalent to LIML, FIML

respectively. However, IIV estimates are the LIML, FIML estimates themselves rather than approximations of them. Instruments and regressors in an iterative estimator may be formed by

$$Y_{(k+1)} = Y_{(k)}\hat{\beta}_{(k)} + Z\hat{\Gamma}_{(k)} \tag{15}$$

or by

$$Y_{(k+1)} = -Z\hat{\Gamma}_{(k)}\hat{\beta}_{(k)}^{-1} \tag{16}$$

giving IIV or FP, or the reduced form versions of IIV or FP respectively according to the use of $Y_{(k+1)}$ in an instrument or regressor set. Evidently, numerical errors apart, at convergence (15) and (16) are identical. Thus neither (1) nor (10) indicate how instruments should be formed, although they suggest (16) for LIML, and (16) with $\Sigma_{(k)} = T^{-1}(Y\hat{\beta}_{(k)}+Z\hat{\Gamma}_{(k)})'(Y\hat{\beta}_{(k)}+Z\hat{\Gamma}_{(k)})$ giving

$$\hat{\delta}_{FIML}^{(k+1)} = (N_{(k)}'\hat{X})^{-1} N_{(k)}' y \text{ with } N_{(k)}' = \hat{X}_{(k)}' (\Sigma_{(k)} \textcircled{x} I)^{-1} \tag{17}$$

and $\hat{X}_{(k)} = \text{diag}\,(Y_{i(k)} \vdots Z_i)$ and arbitrary starts $\hat{Y}_{(1)}$ and $\Sigma_{(1)}$. In this case RFIIV is identical to LIML, and the full information RFIIV with FIML iterates at every step including convergence, for identical starts. Alternatively if the maximum likelihood estimators replace (16) by (15) then IIV and FIIV are identical at every step. In any event IIV with RFIIV and FIIV with full information RFIIV are identical with LIML and FIML respectively at convergence given identical starts or a unique convergence point.

One might try speeding convergence by calculating new instruments $N_{(k)}$ or $\hat{X}_{(k)}$

$$\text{from } \hat{Y}_{(k)} = \alpha\hat{Y}_{(k)}+(1-\alpha)\hat{Y}_{(k-1)} \quad \text{or} \quad \hat{Y}_{(k)}=(I-\alpha L_{(k)})^{-1}(\alpha U_{(k)}+(1-\alpha)I)\hat{Y}_{(k-1)} \tag{18}$$

for FIML and LIML with α a suitable positive real scalar,

$$\left.\begin{array}{ll} \text{(where } L_{ij(k)} = \beta_{ij(k)},\ U_{ij(k)} = 0 & \text{if } i < j \text{ and zero diagonals)} \\ = 0 \qquad\qquad = \beta_{ij(k)} & \text{if } i > j \end{array}\right\} \tag{19}$$

which is merely speeding the calculations i.e. solving

$$Y_{(k)}\,\hat{\beta}_{(k)} + Z\hat{\Gamma}_{(k)} = 0 \tag{20}$$

for $Y_{(k)}$ by methods (4.37) and (4.38). As is shown in Mosbaek and Wold (1970) and Chapter 4 such speeding up of calculations (18) gives the Fractional and Recursive IIV methods respectively so we obtain the same estimates from them as from accelerated FIML and FIML.

Thus IIV and FIIV have precisely the same statistical properties, both in small samples and asymptotically as LIML and FIML respectively. Further in view of (12) and (14) all the above remarks hold equally for FP methods in an asymptotic sense.

It can be seen that (1) and (10) may be computed in various ways, with or without accelerations, for either OLS or IV at the estimation phase, since the instrument formation phase may be altered. For example in (10) we could use either

$$X_{(k)} = \{X_{(k-1)}\,\hat{D}_{(k)i} \vdots Z\} \quad \text{where } X_{(k)} = [Y_{(k)} \vdots Z] \tag{21}$$

or

$$X_{(k)} = \{-Z\left(\hat{\Gamma}_{(k)}\,\hat{\beta}_{(k)}^{-1}\right)_i \vdots Z\} \tag{22}$$

or

$$X_{(k)} = \{Z[\hat{\pi}_{(k-1)i}(I-\hat{\beta}_{(k)})_i - \hat{\Gamma}_{(k)i}] \vdots Z\} \tag{23}$$

for i = FP, IIV where $\hat{\pi}_{(k)}$ is k^{th} iteration estimate of restricted reduced form parameters. Thus IIV by (21), RFIIV by (22) and PIIV by (23) respectively in (10) are identical to LIML at convergence, and at every step if the same instrument formation process is used. Corresponding FP family members are

asymptotically equivalent. The same goes for the full information case (1) for FIML. There are many other possibilities.

5.4 Two illustrations

An application of these results is to a suggestion by Hausman (1974). FIIV and IIV, and their variants, do not have the "up hill" property, i.e. that the likelihood function is strictly increased at each iteration. The "up hill" property is implied by the iteration as specified. This property requires the moment matrix of (1) or (10) to be positive definite. It is, in fact, only necessarily the case asymptotically with IIV estimators although it may be highly likely in small samples. $\hat{X}^{*\prime}\hat{\phi}^{-1}\hat{X}^{*}$

Hausman proposes a modification in replacing $N'\hat{X}$ by / as the moment matrix of (1) (or $\hat{N}'\hat{X}$ by / $\hat{X}^{*\prime}\hat{X}^{*}$ in (10)) which is precisely that of FIFP in (13) (or equivalently that of FP). The modified moment matrix is positive definite and guarantees the "up hill" property as claimed.

Another application is 2SPC (two stage principle components). If not all the components were used in a truncated reduced form, 2SPC may be consistent but not efficient since it cannot satisfy the criterion plim $\hat{\pi}$ = plim Γ.plim β^{-1}. If this estimator were extended by an additional iteration through the restricted reduced form, it would be consistent and efficient at least in an IV context. On convergence it would be IIV (or FP) with PC starts - and the perennial problem of the arbitrary selection of a number of the components is no longer relevant(c.f. Theorem 4.5 and Section 4.5). This example is singled out as the efficiency gains in the studies of Brundy and Jorgenson (1974) appear empirically to be of more significance for this estimator than for others (hence the significance of Corollary 4.1).

5.5 A Summary of Results

We now draw together the main implications of the theory developed in this chapter.

Theorem 1:

The following estimators are asymptotically equivalent, the first two mentioned being identical:

(a) LIML, IIV and FP family members, 2SLS.

(b) FIML, full information IIV and FIFP family members, 3SLS.

Proof:

Considering iterative estimators at convergence - if plural convergence points exist we consider only that point which corresponds to the maximum of (2.35) (see Chapter 14 for methods for selecting that point). (a) by (10) and (12) for LIML, IIV and FP family members. 2SLS may be written as (2.26) and plim π^+ = plim $\hat{\Gamma}$. plim $\hat{\beta}^{-1}$ = π. (b) for any consistent estimator plim $\hat{\Sigma}$ = Σ; the result then follows by (2.41) and (14) and plim π^+ = plim $\hat{\Gamma}$ plim $\hat{\beta}^{-1}$ = π as above (consistency theorems are in Chapters 2 and 4). #

Remark:

Theorem 1 considers only the global maximum solution for LIML, FIML and therefore its equivalent from IIV and FP families. It does not hold for other solutions or convergence points (c.f. Theorem 2.15).

Corollary 1:

Any RLS or IV estimator is asymptotically equivalent to the maximum likelihood estimator of its information class if it is consistent and asymptotically efficient.

Proof:

IV estimators by (10), (2.41) and theorem 4.2.

RLS estimators by (4.11), (10), (2.41) and theorem 4.2. #

Corollary 2:

RFFP, RFIIV and any accelerations of that instrument formation are asymptotically equivalent to the maximum likelihood estimator of their information class, all[1] $S \geq 3$ and consistent $Y^{(o)}$.

Proof:

By Corollary 1 and theorems 4.8 and 4.9. #

---

1. $S \geq 2$ for limited information methods.

Comments:

(i) Theorem 1 for 2SLS, LIML, and 3SLS and FIML is well known but this proof is very much simplified compared with any other of which I am aware. Corollary 1 represents a new extension.

(ii) It is not generally possible to extend the asymptotic equivalences to finite samples, even for FP and IIV family members.

(iii) All the results above are virtual properties.

Theorem 2:

FFP, RFP (R_1FP); FIIV, RIIV (R_1IIV) are consistent, optimal and asymptotically efficient. Any iterative estimator of the FP or IIV family obtained by any convergent acceleration of (15) or (16) shares the asymptotic statistical properties of the unaccelerated iterative estimator using (15) or (16) if Y* is unique.

Proof:

By (18) in the maximum likelihood estimator for IIV family members. Corresponding FP family members are asymptotically equivalent by Theorem 1. #

Theorem 3:

Any FP or IIV family estimator, using an inconsistent start $Y^{(o)}$, is consistent, asymptotically efficient and optimal if it is convergent and Y* unique.

Proof:

If Y* is unique then by (2.41), (10) etc. the iterative estimator coincides with the maximum likelihood one which is derived from the global maximum of (2.35). By Theorem 2.15 that estimate set is consistent, asymptotically efficient and optimal. #

Corollary 3:

If Y* is not unique, plural solutions to iterative estimators can and will be observed corresponding to the local maxima of (2.35). By Theorem 2.15 only that corresponding to the global maximum has the desirable properties of

Theorem 3 whatever the start. As with all maximum likelihood estimation there is no easy check among the plural solutions. By Theorem 4.5 and Corollary 4.5 local solutions are functions of the start values and experimenting with different starts is equivalent to searching the likelihood space for the global maximum. #

Corollary 4:

The consistent, asymptotically efficient and optimal estimates obtained at the global maximum are independent of the starts.

Proof:

Theorem 4.7. #

Remark:

Methods for selecting the desirable estimate set (corresponding to the global maximum of (2.35)) when the likelihood function is not explicitly considered are given in Chapter 14.

Theorem 4:

The asymptotic estimator variance is given by (2.26a) and (2.28a) for limited and full information iterative estimators respectively.

Proof:

Theorem 1 and Theorem 2. #

Comments:

(i) All this theory follows the assumption of a multivariate normal density function for the error process (c.f. Chapter 15). We iterate the orthogonality transform (as well as the sphericality one if appropriate) to convergence to gain a maximum likelihood interpretation but for consistent starts at least there is no penalty to asymptotic properties to stopping the iteration after S = 3, provided (16) is used.

(ii) The properties of estimators using inconsistent starts are rather obvious from theorems 4.5 and 4.6 given convergence and unique Y*. But the question of the properties of earlier iterates is still open. The important

point of Theorem 3 is that all convergent solutions of an iterative estimator must be checked for that corresponding to the global maximum of (2.35) if an arbitrary start is used. This patently has not been done in the past, although it is relatively common in direct maximisation of (2.35). If however a consistent start were used, Theorems 4.8 and 4.9 indicate we may converge on the global maximum solutions for sufficiently large samples and may claim the statistical properties mentioned there, and in Theorems 4.5 and 4.6 at convergence. The latter course of action will save considerable computation overall.

5.6 Conclusion

In many applications it may be of great value to use a least squares or iterative estimator, because it is very much easier to programme and is probably much cheaper to compute. Such an estimator does not require detailed knowledge or expertise in complicated iterative routines as is required by FIML for the efficient nonlinear optimisation of (2.35). Small machines could handle considerably larger models with repeated least squares than with a FIML routine. The maximum likelihood interpretation, however, of IIV and FIIV, or FP and FIFP asymptotically, depends on these estimators being iterated to convergence.

The real objective in convergence is the resulting maximum likelihood interpretation. As a result it is immediate that LIVE is one iterate of LIML[1], FIVE (or FIIV) one of FIML (as noted by Hausman (1974)); that 3SLS iterated only on the covariance matrix is not convergent to FIML, whereas if the instruments are also iterated it is.

However, the asymptotic equivalence of FP and IIV methods, and the equivalence of IIV and maximum likelihood methods, does not mean that we expect to see (approximately) identical estimates, even in the limit (were we able to observe such situations), because it is possible that each different algorithm

---

1. Although it requires a complete specification of the model which LIML does not.

would reach different solutions corresponding to various local maxima. Then, also, finite iteration only provides an approximation of maximum likelihood and in the FP and IIV comparison there are small sample differences, not to mention accumulations of numerical errors.

The FP and IIV estimators considered here are important due to their availability in the frequently encountered situation of $T \lesssim K$. Their efficiency has not been commented upon in previous work. Moreover the efficiency results provide formulae for the standard errors of any of the FP or IIV estimators. In the full information case, maximum likelihood estimation is usually achieved by means of a conjugate direction, or alternative algorithms which avoid the evaluation of derivatives. Hence standard errors, $E[\frac{\partial^2 L}{\partial\delta\partial\delta'}]$, cannot be estimated. This problem is immediately overcome in iterative estimation.

CHAPTER 6: GENERAL INTERDEPENDENT SYSTEMS - WHAT IS THE TRUE FAITH?

6.1 Introduction

The FP and IIV estimator families are said to share many properties of which consistency, asymptotic efficiency, and equivalences with conventional estimators were developed in Chapter 5, under the assumption that plim $T^{-1}Z'U = 0$. However inseparable from these estimators, since the proposal of FP by Wold (1965), is the notion of relaxing some of these zero asymptotic correlation assumptions. Evidently this is a general possibility in view of corollaries 4.2 and 4.3 but we shall be chiefly concerned with the particular cases which suit FP and IIV, popularly known as GEID (general interdependent) and symmetric GEID systems respectively. The original assumption set of section 2.2, sufficient for consistency, is called the classical specification.

In section 2GEID and symmetric GEID specifications are defined. The statistical properties of FP, IIV are investigated for these specifications in sections 3, 4, 5. Other GEID specifications are introduced in Sections 6 and 7 and autoregressive models are treated in Section 8.

6.2 Definitions

We note that (2.30) - the REID specification - includes a predictor specification when

$$E(u_{ti}|Z_t) = 0 \quad i = 1 \ldots M,\ t = 1 \ldots T \tag{1}$$

or equivalently

$$E(v_{ti}|Z_t) = 0 \quad i = 1 \ldots M,\ t = 1 \ldots T \tag{2}$$

by (2.8).

Evidently by corollary 4.2 this is only a sufficient condition for a predictor specification; necessary and sufficient on (2.30) would be the condition

$$E(v_{ti} | Y^*_{ti}, Z_{ti}) = 0 \quad i = 1 \ldots M, \; t = 1 \ldots T \tag{3}$$

The condition (3) allows consistent estimation by (4.27) if Y^* is known. The counterpart of (3) for consistent estimation by (4.28) is evidently by theorem 4.1

$$E(u_{ti} | Y^*_{ti}, Z_{ti}) = 0 \quad i = 1 \ldots M, \; t = 1 \ldots T \tag{4}$$

Again this is a necessary and sufficient condition by corollary 4.2. The condition (3) defines the GEID specification on (2.5) whereas (4) is the symmetric GEID specification.[1] If we are only concerned with asymptotic statistical properties then the expectations in (3) and (4) may be replaced by probability limits. Both may be summarised in

$$E(Y_{ti} | Z_t \Pi_i) = Y^*_{ti} \quad i = 1 \ldots M, \; t = 1 \ldots T \tag{4a}$$

The benefits claimed for the specification (4a) are:

(i) There is a "substantial" or "drastic" generalisation from REID; many less assumptions being necessary (Wold (1973a), Wold (1974), Lyttkens (1973), Lyttkens & Wold (1973)).

(ii) The smaller number of assumptions preserve the parity principle. If there is no parity principle the model is in an awkward situation as it is not generally possible to construct the model so that all the basic

---

1. A much less common usage but taken from Bergstrom (1974)

assumptions hold (Mosbaek & Wold (1970), Lyttkens (1973)).

(iii) GEID is as good as REID causally but has a more general predictor specification (Lyttkens (1973), Wold (1973)).

(iv) Classical specifications and REID are special cases of GEID (Lyttkens (1973)).

(v) GEID recognises that the statistical world is not perfect and that relaxing the assumed zero correlations is realistic because those that remain may have a greater chance of being upheld. "A GEID correlation may be thought of as caused by variables left out of a relation and hence summed in the residual" Agren (1972).

(vi) It is very easy to interpret and programme an estimator (Wold (1973b)).

Definition:

The parity principle holds whenever the number of zero correlation assumptions in the maintained hypothesis exactly equals the number of structural parameters to be estimated. Both (3) and (4) clearly display this property.

Lemma 1:

If Y* exists and is unique, then for convergent iterative estimators

(a) (4), (4a) and the condition of Corollary 4.2 are equivalent.

(b) (3), (4a) and the condition of corollary 4.3 are equivalent.

(c) a consistent $Y^{(0)}$ implies all subsequent IIV iterates are consistent if (4) holds; all subsequent RFFP iterates, but not FP iterates, are consistent if (3) holds.

Proof:

(a) Obvious.

(b) The error process in FP/IIV is given by

$$v_i^{(s)} = y_i - y_i^{(s)} = y_i - X_i^{(s-1)}\delta_i^{(s)} = u_i^{(s)} + V_i^{(s-1)}\beta_i^{(s)} \qquad (5)$$

from (4.6) since $y_i - X_i \delta_i^{(s)} = u_i^{(s)}$. The error process in RFF/RFIIV is therefore

$$v_i^{(s)} = y_i - y_i^{(s)} = y_i - X_i^{(s)}\delta_i^{(s)} = u_i^{(s)} + V_i^{(s)}\beta_i^{(s)} \qquad (6)$$

At Y* therefore $v_i = u_i + V_i\beta_i$ in either case.

(c) follows immediately by theorem 4.1 in IIV family estimators. For RFFP the corollary 4.4 still applies if $Y^{(0)}$ is consistent. However if (3) holds for Y* it will not generally hold for any other $Y^{(s)} \neq Y^*$. #

Comments:

(a) Specifications (3) and (4) are exactly the necessary and sufficient assumption sets for consistent estimates by FP and IIV respectively.

(b) (4a) provides no indication of which estimator is appropriate, although the wrong choice implies inconsistency. Thus the estimator choice becomes part of the maintained bypothesis, contravening the objective of this study.

(c) Naturally if (4a) is known to hold for P replacing Π then $Y^{(s)} = ZP$ will give consistent estimates in the appropriate estimator. (3) and (4) are therefore just two special cases of the set of possible generalisations.

(d) FP, IIV are the only efficient and optimal estimators in GEID systems at convergence (theorems 4.2 and 4.6). For previous iterates clearly theorem 12.3 below still holds.

(e) It is often claimed that FP, IIV are the only consistent estimators for GEID situations (e.g. Wold (1973a)). Fortunately this is false otherwise a consistent $Y^{(0)}$ would be impossible to construct.

6.3 Inconsistency of FP estimates:

In a recent comment, Agren (1975), it is argued that my proposition that one cannot in general expect FP under a GEID specification to be consistent (stated without proof in Hughes Hallett (1973)) is false. The

proof offered by Agren, while correct in logical detail, being of insufficient generality, is inapplicable for the consistency of FP estimators and is therefor misleading as well as inappropriate.

Let b^{ij} be a typical element from β^{-1}.

Agren's proof requires only the assumptions

$$\text{plim } T^{-1} Y_i^{*\prime} v_i = \text{plim } T^{-1} Z_i' v_i = 0 \quad i = 1 \ldots M \tag{7a-b}$$

If $C_{ij} = \text{plim } T^{-1} z_i' u_i$ then we may write (7) as

$$\sum_{j=1}^{M} \sum_{p=1}^{k} b^{pi} \Pi_{jk} C_{jp} = \sum_{j=1}^{M} b^{ji} C_{jk} = 0 \tag{8a-b}$$

if $\beta_{ki} \neq 0$, $\gamma_{ki} \neq 0$. Hence only $\sum_{i=1}^{M} (r_i + s_i)$ assumptions, on MK unknown limiting covariances, are claimed as necessary for consistent FP estimates within a GEID specification. This is the parity principle.

Agren nowhere restricts the covariance matrix E(U'U) to be diagonal or any other special form.[1] Indeed in his previous proof (Agren (1972)), he specifically takes it as a general symmetric matrix; as does Lyttkens (1973, p.362) to whose statement my proposition actually refers (Bergstrom (1974) and Bodin (1974) also do the same). Let

$$V = VR + f, \quad \text{that is } v_i = \sum_{j=1}^{M} r_{ij} v_j + f_i, \; i = 1 \ldots M \tag{9}$$

where f_i are stochastic errors, such that

$$\text{plim } T^{-1} Z_i' f_i = 0 = \text{plim } T^{-1} Y_i^{*\prime} f_i \tag{10}$$

follows (7a) naturally (see section 4 to justify this), and $r_{ii} = 0$.

---

1. Since FP is essentially OLS on (4), using an estimate of Y*, assumptions on E(V'V) would be more appropriate, though a priori unacceptable since we start from (2) and $V = (I-\beta)^{-1}$; the parameters being unknown.

Theorem 1:

FP estimates in a GEID specification are inconsistent in general unless $E(V'V) = \psi = \text{diag } \psi_{ii}$.

Proof:

(i) Recall in theorem 2.1, (a) is a necessary but (b) and (c) are sufficient conditions. Since FP involves repeated OLS regressions on (4), (7a-b) with $E(V) = 0$ are at least necessary conditions[1] for consistent estimates. For each $\beta_{pi} \neq 0$, $\gamma_{ki} \neq 0$ $i = 1 \ldots M$, $p = 1 \ldots r_i$, $k = 1 \ldots s_i$, (7a-b) imply $\text{plim } T^{-1} y_p^{*\prime} v_i = \text{plim } T^{-1} z_k' v_i = 0$; and using (9) and 910) we obtain

$$\sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} y_p^{*\prime} v_j = \sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} z_k' v_j = 0 \qquad \text{(11a-b)}$$

Thus necessary conditions for consistent estimates derived from[2] (6) imply not any general symmetric covariance matrix of reduced form errors but a highly specialised nondiagonal one. Since neither (11a-b) nor such a covariance matrix are assumed (neither Agren (1975) nor Lyttkens (1973) does so) FP estimates are inconsistent. Therefore in the general case a necessary condition for consistency is violated unless $E(V'V)$ is diagonal.

(ii)Although the theorem is proved by any counterexample, such as some failure in conditions (11a-b), we can furthershow it to be true even if (11a-b) had been assumed.

Consider the i^{th} equation. At least $\beta_{pi} \neq 0$, some $p \neq i$, and suppose $\gamma_{ki} \neq 0$. Then, if either $\beta_{pn} = 0$ or $\gamma_{kn} = 0$ for any $n = 1 \ldots M$, $n \neq i$, our assumption $E(v_n v_n') = \sigma_n^2 I_T$ gives, from the conditions (11a-b) corresponding to β_{pi}, γ_{ki},

---

1. Conditions at their most general with consistency treated in a large sample sense.
2. We assume GEID distinct from REID i.e. at least one overidentified equation, at least some $\text{plim } T^{-1} z_k' v_j \neq 0$ $j = 1 \ldots M$.

$$\frac{1}{\sigma_n^2} \sum_j r_{ij} \text{ plim } T^{-1} y_p^{*\prime} v_n v_n' v_j = \frac{1}{\sigma_n^2} \sum_j r_{ij} \text{ plim } T^{-1} z_k' v_n v_n' v_j = 0 \qquad (12)$$

and therefore,

$$\frac{1}{\sigma_n^2} \text{ plim } T^{-1} y_p^{*\prime} v_n \sum_j r_{ij} \text{ plim } T^{-1} v_n' v_j = \frac{1}{\sigma_n^2} \text{ plim } T^{-1} z_k' v_n \sum_j r_{ij} \text{ plim } T^{-1} v_n' v_j = 0 \qquad (13)$$

For the specification (6), the necessary conditions (11a-b) can, in either case, be transformed into necessary conditions on the reduced form error covariances

$$\sum_j r_{ij} \text{ plim } T^{-1} v_n' v_j = 0 \qquad (14)$$

Now (14) holds for every n = 1 .... M; $n \neq i$, since the necessary (order) condition for identification, in any submodel of equations i and n, requires at least one row of X_i to be excluded a priori from X_n. We can clearly obtain (14) for each i = 1 .... M. Noting that the limiting covariances are symmetric, we have M(M-1) equations in M/2(M+1) unknown covariances. Therefore when $M \geq 3$ the necessary conditions (14) cannot be satisfied[1] for the particular R matrix of a given general model. (14) becomes R plim $T^{-1} vv' = 0$ using the structure of R. The necessary conditions for consistent FP estimates are satisfied if and only if $R \equiv 0$ since plim $T^{-1} vv'$ contains plim $T^{-1} v_s v_s' \neq 0$ s = 1 .... M which is nonsingular (and certainly does not have n zero eigenvalues). The extra $\Sigma(r_i + s_i)$ conditions (11a-b) are independent of (7a-b); a total of $2\Sigma(r_i+s_i)$. Finally consider models having M = 2. In models for which $KM \leq 2\Sigma(r_i+s_i)$ REID ≡ GEID. One equation overidentified implies if M = 2, then $\Sigma r_i = 2$ and $\Sigma s_i \geq 3$ and hence REID ≡ GEID. #

---

1. There is one exact solution possible if M = 3; but it has a negligable probability of being observed and so cannot be treated as a general possibility.

Comment:

If $KM \leq 2\Sigma(r_i + s_i)$, any $C_{ij} \neq 0$ implies inconsistent FP estimates. Thus for example Summers, Grischick-Haavelmo, Bassman and Johnston's models, where FP has been applied as a GEID estimator, are inconsistently estimated.

Remark:

In conclusion, therefore the proposition above was worded "in general" and no reference was made to the particular assumptions in Agren (1975) nor to the lack of (11a-b). In order to satisfy the necessary conditions for consistent FP estimates we must assume either E(V'V) = a diagonal matrix or a REID specification (5). These are hardly more general conditions.

6.4 An Extension of Theorem 1:

It may be contended that the assumption (10) is an arbitrary assumption and is no part of the GEID specification and further that the proof of Theorem 1 depends entirely on (10). However consider the following arguments to refute these objections.

6.4.1

If $\text{plim } T^{-1} X_i^{*\prime} f_i \neq 0$ then replacing (11a-b) we have

$$\sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} x_p^{*\prime} v_j + \text{plim } T^{-1} x_p^{*\prime} f_i = 0 \quad p = 1 \ldots r_i + s_i \qquad (15)$$

Since (and neither Agren (1975) nor Lyttkens (1973) do so) no assumptions at all are made on the covariance structure of errors V, cannot generally hold in addition to (6). Yet when it does not, FP estimates are inconsistent. Consider Summer's model

$$y_1 = \beta_1 y_2^* + \gamma_1 z_1 + \gamma_2 z_2 + v_1$$

$$y_2 = \beta_2 y_1^* + \gamma_3 z_3 + \gamma_4 z_4 + z_2$$

If $E(y_2^{*\prime} v_2) \neq 0$ and $E(v_1' v_2) \neq 0$ then $E(y_2^{*\prime} v_1) = 0$ is a contradiction, unless the two nonzero covariances are mutally compensating. That is

asymptotically what (15) states. If (15) is not assumed in addition to (6) then the general model is inconsistently estimated.

6.4.2

We investigate further models with (6) and (11a-b) or (15). The definition of a GEID specification (6) cannot itself necessarily prevent a particular model satisfying (6), from also satisfying (10). The class of models satisfying (6) contains a nonempty subset which also satisfy (10) - namely those models for which also

$$\sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} x_p^{*\prime} v_j = 0 \quad p = 1 \ldots r_i + s_i, \; i = 1 \ldots M \tag{16}$$

Within (16) plim $T^{-1} x_p^{*\prime} v_j \neq 0$ if $\beta_{pj} = 0$ or $\gamma_{pj} = 0$ as required and (6) then holds iff (10) does.

6.4.3

Suppose again plim $T^{-1} X_i^{*\prime} f_i \neq 0$ $i = 1 \ldots M$. Retracing the argument, (13) becomes

$$\frac{1}{\sigma_n^2} \sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} v_s' v_j = \frac{\text{plim } T^{-1} x_p^{*\prime} f_i}{\text{plim } T^{-1} x_p^{*\prime} v_s} \tag{17}$$

for all $p = 1 \ldots r_i + s_i$, $i \neq s$ such that $\beta_{ps} = 0$ or $\gamma_{ps} = 0$, and $i = 1 \ldots M$.

Simplify notation with $\frac{1}{\sigma_s^2} \sum_{j=1}^{M} r_{ij} \text{ plim } T^{-1} v_s' v_j = a_{si}$,

$-\text{plim } T^{-1} x_p^{*\prime} f_i = b_{pi}$, plim $T^{-1} x_p^{*\prime} v_s = d_{ps}$. Then (17) may be rewritten

$$AW = 0 \tag{18}$$

where

$$A = \begin{bmatrix} 1 & 0 \cdots\cdots 0 & a_{11} & & 0 \\ 1 & & & \ddots & \\ \vdots & 0 & 0 & & a_{1M} \\ 1 & & & & \\ \hline 0 & 1 & a_{11} & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 1 & 0 & & a_{1M} \\ \hline & & & & \\ & & & & \\ & & & & \\ \hline & 1 & a_{M,1} & & 0 \\ 0 & \vdots & & \ddots & \\ & 1 & 0 & & a_{MM} \end{bmatrix}$$

and $w' = [b_{11} \cdots\cdots b_{M,r_M+s_M}\ d_{11} \cdots\cdots d_{s_1,r_1+s_1} \cdots\cdots d_{s_M,r_M+s_M}]$.

It is understood that a_{si} elements are only entered if (17), for the given i and p, indicates $\beta_{ps} = 0$ or $\gamma_{ps} = 0$, some s. Note however no row or column of A is composed solely of zeros (i.e. for columns, no equation has no variable, no variable is in every equation;[1] and rows by construction) and because of the structure of A no rows can be eliminated by elementary linear operations. Further the negation of (10) ensures all $a_{si} \neq 0$.

---

1. If a constant is in the model, say Z_o is the unit vector, then $C_{oi} = d_{os} = 0$ all i,s. Thus it is taken that K, and s_i make no allowance for any constant.

A has columns on the left corresponding to b_{pi}, i.e. $\sum_{i=1}^{M} (r_i+s_i)$ of them. It has columns on the right corresponding to d_{ps}, i.e. one for each zero in the matrix D for each variable in turn (less those zeros in positions β_{ii} $i = 1 \ldots M$ since (13) requires $s \neq 1$) i.e. $M(M + K) - \sum_{i=1}^{M} (r_i + s_i) - M$. Overall A has $M(M + K - 1)$ columns. A has rows collected in blocks - one for each variable in each equation. The number of rows in each block is the number of zeros in the row of D corresponding to the variable of the equation being treated. Define α_j as the number of nonzero elements in the j^{th} row of D. Then A has $\sum_{j=1}^{M+K} \alpha_j(M - \alpha_j)$ rows. It is obvious however from (17) itself that iff $\alpha_j = 1$, $j = 1 \ldots M + K$ (i.e. each variable appears only in one equation) the number of rows in A exactly equals the number of distinct d_{ps} elements.

One solution to (18) is $w = 0$ but by hypothesis we seek a second $w \neq 0$. This is only possible if A is singular, having more columns than rows. Thus GEID cannot exist if either

$$\sum_{j=1}^{M+K} \alpha_j(M-\alpha_j) \geq M(M+K-1) \text{ or } \sum_{j=1}^{M+K} (\alpha_j-1)(M-\alpha_j) \geq \sum_{i=1}^{M} (r_i+s_i) \qquad (19)$$

that is sufficient variables appear repeated in several equations. The reader will have no difficulty in checking (19) holds, with very few exceptions, for the model of his choice. Thus a substantial set of counter-examples are available for my proof even if (10) is not used.

For those few exceptions to (19) note the rows of A impose restrictions quite independent of any quantity in (7a-b). Even if $\alpha_j=1, j=1....M+K$, in total GEID requires the satisfaction of exactly the same number of restrictions as assuming plim T^{-1} $(Z'U) = 0$; and more if $\alpha_j > 1$ at least one j. Thus the maintained hypothesis is no way reduced, in contrast to the objective of GEID. Indeed Mosbaek and Wold (1970) write the parity

principle is "the key incentive" for GEID. Moreover whatever the probabilities of (6) and (18), or plim T^{-1} $(Z'U) = 0$, being true the latter avoids the logical difficulty of making assumptions involving the parameters to be estimated. For the general model GEID has more assumptions and therefore I conclude (18) cannot be the GEID system defined and without loss of generality I assume (10).

6.5 Comments

(a) The REID specification indicates no extra potential for FP over conventional estimators. Further, even in this case, since FP uses estimates $Y^{(s)}$, any of the multifarious possible misspecification biases in one equation will be propagated throughout system (4). Therefore my point remains: with reference to the matrices in (7a-b) "there will be nonzero elements in a GEID system and the same proofs extend for many REID system equations".

(b) A diagonal structure error covariance matrix does not imply consistent FP estimates, if (6) holds. Indeed assuming $E(V'V) = \psi$ implies

$$E(U'U) = \beta'\psi\beta, \quad \text{nondiagonal.} \tag{20}$$

This kind of assumption is unacceptable for several reasons. First it is rediculously specialised and has a negligable probability of being upheld precisely in practice in all elements. Secondly it involves a precise a priori knowledge of the parameters to be estimated and therefore it is logically impossible to make. Third, and most important, (20) differs in principle from an assumption on the structural error covariances for conventional estimators. The latter affects the efficiency, not consistency, of estimates and therefore is a testable and avoidable assumption - a decision on its acceptability may be taken after estimation. For FP estimation in GEID conditions, consistency is at issue so the

assumption becomes untestable and part of the maintained hypothesis. Fourthly, I am unaware of any econometric model where restrictions of this kind on either error covariance matrix have even been attempted. It seems an undesirable innovation. The only assumptions in use are, of course, E(UU) a general symmetric matrix or diagonal. A general case would imply the former.

It is argued that the parity principle is "the key incentive" for a GEID specification (Mosbaek & Wold (1970)) and most discussions refer to that property. Yet

Lemma 2:

Consistent FP estimates result only from specifications not satisfying the parity principle.

Proof:

REID requires the full set of KM zero correlation assumptions; and (6) with (20) implies, being sufficient and necessary conditions by Corollary 4.3 and Theorem 1 above, $\sum_{i=1}^{M} (r_i + s_i) + M/2\,(M - 1)$ of them. #

Evidently the originators of FP considered its use in REID conditions to be very uninteresting and hence it was largely ignored.

6.6 Comparisons with IIV estimates

With reference to the previous paragraph, one frequently reads that FP is the GEID estimator (e.g. Agren (1975)). But

$$E(u_{ti} \mid Y^{*}_{ti}, Z_{ti}) = 0 \tag{21}$$

is perhaps more natural than (6). Then

$$\sum_{j=1}^{K} \Pi_{jk} C_{ji} = 0 \text{ and } C_{ki} = 0 \tag{22a-b}$$

where $\beta_{ki} \neq 0$, $\gamma_{ki} \neq 0$, follow from (4), i.e.

$$\text{plim } T^{-1} [Y^{*}_{i} : Z_{i}]'u_{i} = 0 \quad i = 1 \ldots M. \qquad (22a\text{-}b)$$

are the conditions for IIV estimates corresponding to (8a-b). Thus;

Theorem 2

We add the condition that E(U'U) be diagonal to (22a-b) to obtain necessary (and sufficient) conditions for IIV estimates to be consistent in a GEID specification (16).

Proof:

As Theorem 1 with u_j for v_j throughout. #

Corollary 1:

There is no unique GEID estimator but rather corresponding separate GEID conditions of equal generality exist for each estimator giving consistent estimates in its own GEID specification.[1] #

Corollary 2:

Those conditions for FP and IIV are mutually exclusive - if one holds the other cannot.

Proof:

E(V'V) diagonal implies E(U'U) is not, and visa versa, by (20) - unless β is diagonal which by assumption it is not. #

Corollary 3:

It follows that REID conditions, which are merely sufficient for the consistency of FP and IIV estimators, are the highest (most general) common factor in the assumption sets under which both may be consistent - and thus have equivalent statistical properties. #

Comment:

The conditions (8a-b) and (22a) depend on a precise a priori knowledge of the parameters to be estimated. We can never know therefore, a priori,

---

1. In fact necessary and sufficient GEID type conditions for consistency may be obtained for any estimator, even OLS (see below).

which estimator to select.[1] Once again it is logically impossible to assume, a priori, that either holds, or even that either should hold precisely with more than a negligable probability. Moreover the assumption of (8a-b), (22a-b) or E(V'V) or E(U'U) being diagonal as a justification of FP or IIV methods is illegitimate since they are untestable and therefore augment the maintained hypothesis. Indeed if $MK \leq \Sigma(r_i+s_i)+\frac{M}{2}(M-1)$ (i.e. $M(M+1) \geq 2\Sigma q_i$) the REID maintained hypothesis is actually smaller. Although this may not happen frequently,[2] except for models with little overidentification, the reduction in the number of assumptions in a GEID or symmetric GEID maintained hypothesis is quite small - certainly relative to the parity principle (Table 1). The necessary assumptions can be observed with at best zero probability so the estimators have effectively no more general application than any other. The generalisation in GEID specifications is so slight as to be negligable.

Table 1: the number of zero asymptotic covariances in the maintained hypothesis.

	REID	PARITY PRINCIPLE	GEID
Summers	8	6	7
Klein I	21	12	15
Christ	52	16	20
Klein-Goldberger	253	35	90
Wharton	11988	146	812

(Constants are neglected since they are to be uncorrelated with errors — implies E(U) = 0. This Wold (1974) did for GEID, not REID, making GEID appear more favourable than here.)

---

1. Mosbaek and Wold write "the selection of a preferred method should not be separated from decision on the type of model specification" - but this seems to be taking the selection problem too far.

2. It holds for the Girschick-Haavelmo model.

6.7 GEID conditions for other estimators:

It is erroneously argued that "classical methods for estimating equation systems as well as OLS are inconsistent in GEID specifications" (e.g. Mosbaek and Wold (1970) p.20, p.317 or Wold (1973a)). Maximum likelihood estimators have GEID conditions also. Theorem 5.1 shows identical starts lead to identical estimates in IIV methods and LIML (or full information IIV and FIML). Thus the necessary and sufficient conditions for consistent LIML estimates are precisely those of IIV. However by Corollary 2 the asymptotic equivalence of FP and IIV no longer holds in any GEID specification.

Comment (c) of Section 4.2 provides a further example; it specifies the 2SLS GEID condition, which is evidently of no less generality than that of FP or IIV themselves, although equally unlikely to hold precisely. We may specify another case when 2SLS will be consistent. Sawa (1972) shows for certain models 2SLS is unbiased whenever the direct simultaneous bias (measured by $\beta_{ji}\, v_j$, $j = 1 \ldots r_i$) exactly cancels the indirect (measured by $r_{ji}\, v_j$, $i \neq j = 1 \ldots M$). One may conjecture this is a general result for all models including GEID specifications. In any event there will be one specification outside REID when the biases are mutually cancelling. This case has equal probability to that of GEID of being observed. Hence 2SLS has a more general GEID specification.

A GEID specification for OLS is also available. Each parameter is consistently estimated by OLS iff $\text{plim}\ T^{-1}(y_j'u_i) = \text{plim}\ T^{-1}Z_i'u_i = 0$ i.e.

$$\sum_j \Pi_{jk}\, C_{ji} + \sum_j \sigma_{ji}\, b^{kj} = \text{plim}\ T^{-1}\, Z_i'u_i = 0 \quad i = 1 \ldots M \qquad (23).$$

It can be seen this is very slightly less restrictive than (22a-b) due to the extra term allowing consistent estimates for some E(U'U) diagonal as well as nondiagonal. It has therefore all the attributes of GEID, except that REID is not a "special" case of it.

6.8 Other GEID conditions:

A $Y^{(s)}$ - GEID specification is defined as that specification of zero asymptotic correlations between Z and U such that (a) the IV or RLS estimator using $Y^{(s)}$ is consistent, (b) the probability of the specification being true is no less than that of GEID or symmetric GEID and (c) REID is a proper subset of it. We can now state

(i) Symmetric GEID is only one member of the set of all possible $Y^{(s)}$ - GEID specifications associated with IV estimators. The problem is that we have no means of defining an algorithm to estimate the relevant $Y^{(s)}$ without prior knowledge of the correlation structure of $X_i^{(s)}$ and u_i - inclusive of the special case of Y* - GEID. Hence symmetric GEID is an "ex post" justification of the usefulness of IIV, being merely the particular $Y^{(s)}$ - GEID which indicates the necessary and sufficient consistency conditions for that estimator.

(ii) A similar argumeent applies to GEID. $Y^{(s)}$ - GEID specifications associated with RLS estimators are those for which plim $T^{-1} X_i^{(s)\prime}(V_i^{(s)} \beta_i + u_i) = 0$. Again the problem is to specify the relevant algorithm without prior knowledge of the correlation structure of $X_i^{(s)}$ and v_i. However GEID is the special case for which FP is suitable.

(iii) Clearly every model has some $Y^{(s)}$ - GEID specifaction for RLS and IV estimators. But the set for which any one estimator is suitable is a set of measure zero. Hence any $Y^{(s)}$ - GEID has negligable probability of being precisely true.

(iv) Only one $Y^{(s)}$ - GEID specification for RLS and IV estimators is efficient and optimal; that is GEID and symmetric GEID where (4a) holds.

(v) Only one $Y^{(s)}$ - GEID has a maximum likelihood interpretation.

A more pertinent interpretation of (22a-b) with E(U'U) is therefore necessary and sufficient conditions for consistent LIML estimates. Thus contrary to Wold (1973a) classical and iterative estimators are competitive.

6.9 Autoregressive models and serial correlation:

Bodin (1974) demonstrates that if $y_{t-\ell,j} \in Z_i$ for any $\ell \geq 1$, $j = 1 \ldots M$, $i = 1 \ldots M$ then $E(v_{tk}\, v_{t'k}) \neq 0$ for any $t \neq t'$ and $k = 1 \ldots M$. The introduction of lagged endogenous variables at any point in a model with a GEID specification indues serial correlation in all the errors.

Theorem 3:

This result generalises to any $Y^{(s)}$ - GEID system for RLS or IV estimators.

Proof:

For RLS estimators[1], a $Y^{(s)}$ - GEID specifications implies that a linear association holds approximately between V_t and Z_t; say $V_t = Z_t A + \eta_t$ where $Z_t = [Y_{(\ell)t} : Z_t^*]$ of chapter 2 and η_t is a stochastic error, $E(\eta_t) = 0$, $E(\eta_{ti}\, \eta'_{ti}) = \sigma_i^2 I$, $i = 1 \ldots M$.

$$E(v_{tk}\, v_{t'k}) = a_k' E(Z_t' Z_{t'}) a_k + a_k' E(Z_t'\, \eta_{t'k}) + E(\eta_{tk}\, Z_{t'}) a_k + E(\eta_{tk}\, \eta_{t'k}) \qquad (24)$$

where $t \neq t'$ and a_k is the k^{th} column of A.

Even if we assume the last three terms to be zero, the first is not since $a_{ki} \neq 0$ some $i = 1 \ldots K$ and $E(Y'_{(\ell)t-1} Y_{(\ell)t'-1} = \sum_{j=1}^{t-t'} (\Pi_2^* \Pi_1^{*j})' E(Y'_{(\ell)t'-1} Y_{(\ell)t'-1})$

$\neq 0$.

For IV estimators write U_t for V_t throughout. #

Thus to estimate consistently we must either assume an equal and opposite serial correlation process in V_t or U_t as appropriate (which is so specialised as to have zero probability of being upheld) or design an autoregressive version of each FP and IIV estimator. As always there is

---

1. Suggested by Bodin (1974)

an estimator problem when autoregressive models with serially correlated errors are treated. A procedure based on (2.50) is given in Chapter 7. Theorem 3 is an important difference between classical (REID) systems and GEID systems which implies quite different estimation problems and quite different statistical properties for the estimators that have been discussed so far.

6.10 Conclusions:

The two important differences between the estimation theory of GEID and classical or REID systems are that the model must have either E(U'U), or E(V'V), diagonal and no lagged dependent variables and must not have serial correlation. The assumption that such a specification is correct is part of the mtaintained hypothesis. Even then only in two special cases are efficient and optimal estimates possible. The assumption of (3) or (4) is very restrictive in the relationship of the covariances to the point of being special cases which can never hold precisely - even if we knew which GEID condition might be assumed. Since suitable estimators for these cases are iterative and use restricted reduced forms, any small failure in the assumptions will cause inconsistencies to be propogated throughout the model. In fact it is much more likely that the model is misspecified, and has measurement errors, than it has exactly the special GEID specification.

It is unsatisfactory to include in the maintained hypothesis assumptions that we know cannot be upheld even in a correctly specified model in place of those which can, provided predetermined variables are genuinely so. Moreover the selection of the estimator is part of the maintained hypothesis (without perfect knowledge of the parameters to be estimated) and no estimator is more generally applicable than any other. Our intention was to relax the maintained hypothesis in order to compare estimators. Thus GEID specifications being mutually exclusive must be rejected in

favour of their highest common factor REID to fulful our objectives.

GEID systems are merely necessary and sufficient conditions for

consistency - a theoretical possibility but not a practical consideration.

CHAPTER 7: EXTENSIONS OF ITERATIVE ESTIMATION THEORY

7.1 Introduction

Equation (5.10) allows the extension of iterative estimation theory and, by the identity of IIV and maximum likelihood estimators, the extension also of the latter. A number of well known results in the context of maximum likelihood estimation are presented in Section 2 with proofs that are new and particularly simple. A new and more straightforward version of the FIML estimator, suitable for model (2.5) together with its relationship to the systems k-class estimator of (2.48) is presented in Section 3. Some discussion of the maximum likelihood interpretation of iterative estimators is given in Section 4. An isolated problem is the apparent conflict of the efficiency results of the previous chapter with those of Dhrymes and Pandit (1972). This is treated in Section 5. The remainder of the Chapter is devoted to designing iterative, and hence maximum likelihood estimators, for each information class.

In what follows we refer to a group of estimators as a class by the name of the information class for which they are asymptotically efficient. We distinguish between single equation and restricted estimators. The former is one whose normal equations depend only on the specification of the particular equation being estimated and a knowledge of the predetermined variables of the model as a whole. The latter has normal equations dependent on the specification of all equations. This distinction will facilitate the discussion of practical aspects of iterative estimation from the users point of view. Clearly no full-information estimators are single equation ones, but several limited information and autoregressive estimators are. LIML however occupies an ambiguous position. (2.43) suggests a single equation estimator but (5.10) a restricted one. Here the distinction turns on the use of restricted or unrestricted reduced form parameters.

7.2 Some results for (2.41), (5.10) and (4.27)

The results which follow are derived for explicit use in the maximum likelihood, IIV or asymptotic FP context of (2.41) or (5.10). They are well known results but the proofs here are offered as being a simplification on those used previously.

Theorem 1:

Identities may be ignored in the estimation phase of IIV and FP family estimators and consequently direct maximum likelihood estimation.

Proof:

(2.41) and (5.10). #

Theorem 2:

(i) (2.41) and (5.10) consequently LIML, FIML, also IIV and FP family members, are identical in full and limited information versions in the i^{th} equation if $\sigma_{ij} = 0 \quad i \neq j = 1 \ldots M$.

(ii) FP ≡ FIFP in all family members if that equation is just identified; it is also true for (2.41) and (5.10) asymptotically.

Proof:

(i) Suppose $\sigma_{ij} = 0$ all $i \neq j$, (2.41) becomes

$$\hat{\delta} = [\hat{X}^{*\prime}(\Sigma^{-1} \otimes I)\hat{X}]^{-1}[\hat{X}^{*\prime}(\Sigma^{-1} \otimes I)y]$$

$$= [\text{diag } \frac{1}{\sigma_{ii}} X_i^{*\prime}X_i]^{-1}(\text{diag } \frac{1}{\sigma_{ii}} X_i^{*\prime}y_i)$$

$$\text{diag } [(X_i^{*\prime}X_i)^{-1}X_i^{*\prime}y_i]$$

which is (5.10).

For FP replace $\hat{X}$ by $\hat{X}^*$ and X_i by X_i^* throughout. These results are invariant to particular FP and IIV members.

(ii) Any FP family member may be written as

$$\hat{\delta}_i = (Q_i Q'_i)^{-1} Q_i p_i$$

where $Q_i Q'_i = X_i^{*\prime} X_i^*$

which is symmetric and positive definite; Q_i is M x M triangular nonsingular (obtained by Nagar's method) and $p_i = Q'^{-1}_i X_i^{*\prime} y_i$.

Then $\hat{\delta} = (Q^*Q^*)^{-1}Q^*p = Q^*p$ where $Q^* = \text{diag } Q_i$; $p' = (p'_1 \dots p'_M)$.

Since $(\hat{X}^{*\prime}\phi^{-1}\hat{X}^*) = (S'(\Sigma^{-1} \otimes X^{*\prime}X^*)S)$ using the matrix S of Theorem 2.11, FIFP is $\hat{\delta} = (Q^*\phi^{-1}Q^*)^{-1}Q^*\phi p = Q^*p$ so FP ≡ FIFP for all family members.

This proof does not extend to (2.41) and (5.10) since the necessary condition for the existence of Q_i is that $X_i^{*\prime}X_i$ be symmetric, which holds only asymptotically. #

Comments:

(a) FIMD (Bergstrom 1974) i.e. FIML with Σ diagonal, should not be treated as a separate estimator; it is identical with LIML.

(b) Equations with structural errors mutually uncorrelated should be estimated in a separate block by LIML, IIV or FP family members. The separate block may include just identified equations if the estimator is an FP family member.

(c) Except in these cases full information methods dominate limited information ones.

Theorem 3:

Consistent and asymptotically efficient estimates result from (2.41), (5.10) or FP equivalents applied to individual subsystems of a block recursive model (as defined in comment (a) following Theorem 2.11).

Proof:

(2.41), (5.10) or (4.27). #

Remark:

Theorem 3 contains some results of Chow (1974) in simpler fashion.

Theorem 4:

Let $X_i = [X_{1i} \vdots X_{2i}]$ and $\delta_i' = [\delta'_{1i} \vdots \delta'_{2i}]$ and define X_{i*} be the residuals from the regression of X_{1i} on X_{2i}, X_{i**} be those from X_{1i*} on X_{2i*} and y_{i*} those from y_i on X_{2i} then the maximum likelihood, IIV and FP estimates of δ_{1i} are identical to those in $y_{i*} = X_{i**}d + u^*$, where the same normal equation system is used at all stages.

Proof:

(a) IIV and maximum likelihood:

$X_{i*} = X_{1i} - X_{2i}(X_{2i}^{*\prime} X_{2i})^{-1}X_{2i}^{*\prime}X_{1i} = Q_2X_{1i}$ so $y_{i*} = Q_2y_i$ and $X_{i**} = Q_1X_{1i}^*$

where $Q_1 = I - X_{2i}^*(X_{2i}^{*\prime}X_{2i})^{-1}X_{2i}^{*\prime}$. Thus $Q_1'Q_2 = Q_2$.

Then (5.13) is $$\begin{pmatrix} X_{1i}^{*\prime}X_{1i}\hat{\delta}_{1i} + X_{1i}^{*\prime}X_{2i}\hat{\delta}_{2i} \\ X_{2i}^{*\prime}X_{1i}\hat{\delta}_{1i} + X_{2i}^{*\prime}X_{2i}\hat{\delta}_{2i} \end{pmatrix} = \begin{pmatrix} X_{1i}^{*\prime}y_i \\ X_{2i}^{*\prime}y_i \end{pmatrix}$$

and eliminating $\hat{\delta}_{2i}$ between them gives $(X_{1i}^{*\prime}Q_2X_{1i})\hat{\delta}_{1i} = X_{1i}^{*\prime}Q_2y_i$.

But also $\hat{d} = (X_{i**}'X_{i*})^{-1}(X_{i**}'y_{i*}) = (X_{1i}^{*\prime}Q_2X_{1i})^{-1}X_{1i}^{*\prime}Q_2y_{i*}$.

(b) FP: OLS regressions imply both Q_1 and Q_2 contain $(X_{2i}^{*\prime}X_{2i}^*)^{-1}$ but still $Q_1'Q_2 = Q_2$.

(c) Extension to all other information class estimators by prior application of the sphericality transform matrix L of Lemma 4.1.

(d) This result is valid for any step S in an interative estimator. #

Comments:

Theorem 4 generalises a well known result of OLS estimation to other estimators. The applications are chiefly for deseasonalising or detrending data where X_{2i} are seasonal factors or a trend.

7.3 A k-class FIML estimator

A disadvantage of the FIML estimator is the substantial computation

involved either as (2.38) or (2.41). The problem is nonlinearities in the normal equations even for the linear model (2.5). It would be convenient to have a linear set of normal equations; this is presented in Theorem 5.

Theorem 5:

FIML estimates may be obtained from

$$\hat{\delta} = [\hat{X}'(\Sigma^{-1} \otimes I)(\text{diag } I - k_i Q)\hat{X}]^{-1}[\hat{X}'(\Sigma^{-1} \otimes I)(\text{diag } I - k_i Q)y] \qquad (1)$$

where $k_i = \lambda_i$ of (2.32) and $Q = I - Z(Z'Z)^{-1}Z'$.

Proof:

Consider Y^* as the convergence point of iteration (5.10) corresponding to the global maximum of (2.35). Under the transformation appropriate to LIML, the results of section 5.3 show estimates obtained by (5.10) at Y^* identical to LIML obtained from (2.47) with $k_i = \hat{\lambda}_i$, all i; $\hat{\lambda}_i$ from (2.32).

$$\therefore \quad \tilde{\delta} = (\hat{X}^{*\prime}\hat{X})^{-1}\hat{X}^{*\prime}y = [\hat{X}'(\text{diag}(I-k_iQ))\hat{X}]^{-1}[\hat{X}'(\text{diag}(I-k_iQ))y]$$

That this LIML estimate corresponds to the global maximum of (2.35) is established in Corollary 1 below. Thus the solutions to

$$(\hat{X}^{*\prime}\hat{X})\tilde{\delta} = \hat{X}^{*\prime}y \text{ and } [\hat{X}'(\text{diag}(I-k_iQ))\hat{X}]\tilde{\delta} = [\hat{X}'(\text{diag}(I-k_iQ))y] \qquad (2)$$

are identical. Define X_i^c to be the complement of X_i in X; and $\bar{X}_i$ to be X with every element of X_i^c set to zero; finally $\bar{X} = \text{diag } \bar{X}_i$. (2) becomes

$$(\bar{X}^{*\prime}\bar{X})\tilde{\delta}^* = \bar{X}^{*\prime}y \text{ and } [\bar{X}'(\text{diag}(I-k_iQ))\bar{X}]\tilde{\delta}^* = [\bar{X}'(\text{Diag}(I-k_iQ))y] \qquad (3)$$

where $\tilde{\delta}^* = R\tilde{\delta}$ and R is the selection matrix defined in theorem 2.11. Premultiplication of both parts of (3) by $R'(\Sigma^{-1} \otimes I)$ gives new estimates $\tilde{\delta}^* = R\tilde{\delta}$ identical in both systems. Implementing the selection matrices (after some conventional manipluation) we find $\hat{\delta}$ is obtained identically from (2.41), a FIML estimator, on the left and by (1) on the right. #

Remark:

To use (1) requires the solution of M eigenproblems as well as (1) itself. But estimates may no longer be obtained directly as the associated normalised eigenvectors as in LIML. Since I obtained theorem 5, another applied result was presented by Scharf (1976). In introducing the K-matrix class estimators, FIML is shown to be the solution $\hat{\delta}$ to

$$\sigma^{ih} X_i'(I-k_{ih}Q)X_{h(i,h=1....M)}\hat{\delta} = \sum_{h=1}^{M} \sigma^{ih} X_i'(I-k_{ih}Q)y_{h(i=1....M)} \qquad (3a)$$

where $k_{ih} = \frac{w^{ih}}{\sigma^{ih}}$ and w^{ih} is from $w^{-1} = [T^{-1}\beta'Y'QY\beta]^{-1}$ and σ^{ih} from Σ^{-1}. In practical terms, this formulation avoids computing M eigenvalues at a cost of a rather complicated nonlinear iteration to compute the matrix K simultaneously with the parameters (and Σ, which is required in (1) anyway). The convergence properties and computational advantages are considered in Chapter 13.

For the remaining information classes we replace $\Sigma^{-1} \otimes I$ in (1) by ϕ^{*-1} of section 2.4.

Comments:

(a) This result is obvious in that (2.41) is exactly the full information version of (5.10) and (1) that of (2.32 - 34).

(b) The full information k-class estimator (2.48) is defective in that k is required to be identical for all equations. This is not the case for (2.47). Thus Savin's "LVR estimator" is not the counterpart of LIML.

(c) We note the remarks of Section 3.3. apply to (1); plim $k_i = 1$ all i, and Theorem 2.16 holds.

(d) A suggested procedure for obtaining (1) is to obtain k_i by (2.32) and then use $\Sigma^{(1)}$ from LIML residuals. The resulting estimates only approximate (1) and for exact results we should iterate Σ to convergence. Convergence is however guaranteed (Theorem 6), and would be rather rapid.

(e) It is likely that (1) is computationally more efficient than FIML with direct optimisation but it is unclear if it is also more efficient than (2.41).

(f) (1) and (2.41) both give more insight to the FIML estimates than (2.38) can do. It is now obvious that both the orthogonality transformation and the sphericality transformation must be iterated to convergence (in (1) the orthogonality transform is completed in one step by obtaining M eigenvalues for k_i). The sphericality transformation may be ignored for LIML as (5.10) shows the estimates independent of the error variances but the orthogonality transformation must be iterated. Other k-class estimators (in particular 2SLS and 3SLS) take approximations, possibly consistent, to the orthogonality transformations together with estimates, possibly consistent, of the error variance - covariance matrix in all but limited information estimators. Various intermediate estimators exist in which some or all iterations of one, but not both, transformations are undertaken; but maximum likelihood estimation requires both to be fully iterated.

Theorem 6:

The iteration on δ in (1) is globally convergent for any fixed values k_i $i = 1 \ldots M$.

Proof:

the iteration of step S + 1 is written

$$\hat{X}'(\Sigma^{(s)-1} \otimes I)(\text{diag } (I-k_iQ))\hat{X}\hat{\delta}^{(s+1)} = \hat{X}'(\Sigma^{(s)-1} \otimes I)(\text{diag } (I-k_iQ))y$$

with $\Sigma^{(s)} = T^{-1}(XD^{(s)})'(XD^{(s)})$. It is globally convergent if for any arbitrary start $\Sigma^{(0)}$ and for all $s = 1, \ldots$ $\rho\left\{\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}^{(s)}}\right\} < 1$. Now $\left\{\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}^{(s)}}\right\}$ is square of order $\Sigma(r_i + s_i)$ with elements from

$$[\hat{X}'(\Sigma^{(s)-1} \otimes I)(I_M \otimes I_T - \hat{k} \otimes Q)\hat{X}]\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}_j^{(s)}} = \hat{X}'\left[\frac{\partial\Sigma^{(s)-1}}{\partial\hat{\delta}_j^{(s)}} \otimes I\right](I_M \otimes I_T - \hat{k} \otimes Q)\hat{u}^{(s+1)} \quad (*)$$

where $\hat{k} = \text{diag } (k_i)$ for $j = 1 \ldots \Sigma(r_i+s_i)$.

But $\frac{\partial \Sigma^{(s)^{-1}}}{\partial \hat{\delta}_j^{(s)}} = - \Sigma^{(s)^{-1}} \frac{\partial \Sigma^{(s)}}{\partial \hat{\delta}_j^{(s)}} \Sigma^{(s)^{-1}}$ and by

the definition of $\Sigma^{(s)}$, $\frac{\partial \Sigma^{(s)}}{\partial \hat{\delta}_j^{(s)}} = 0.$

We note the vector on the right of (*) is zero and that the matrix on the left of (*) is non singular, thus, for j=1 ..... $\Sigma(r_i+s_i)$,

giving the result $\left\{\frac{\partial \hat{\delta}^{(s+1)}}{\partial \hat{\delta}^{(s)}}\right\} = 0$ all $s > 0$. #

Remark:

Theorem 6 holds for all systems k-class estimators, including (1), where the k_i are fixed. The required number of iterations will be very small (theoretically zero).

Corollary 1:

(1) and (2.32 - 34) present unique estimates corresponding to a global maximum in the likelihood function (with no convergence problems) iff the smallest root in (2.32) is distinct.

Proof:

Obviously convergence is no problem noting comment (d) above. It is shown in the standard derivation of LIML that to maximise (2.35) is to maximise (e.g. Dhrymes (1970) equation 7.3.64) for the i^{th} equation,

$$L(\beta_i, Y, X) = C^* - \tfrac{T}{2}[\log \beta_i' W_{ii}^* \beta_i - \log \beta_i' W_{ii} \beta_i] \qquad (4)$$

where C* is constant. Since W_{ii}^* and W_{ii} are positive definite the turning points are given by

$$\frac{\partial L(\beta_i, Y, X)}{\partial \beta_i} = \frac{W_{ii}^* \beta_i}{\beta_i' W_{ii}^* \beta_i} - \frac{W_{ii} \beta_i}{\beta_i' W_{ii} \beta_i} = 0 \qquad (5)$$

or

$$(W_{ii}^* - \mu_i W_{ii}) \beta_i = 0; \quad \mu_i = \frac{\beta_i' W_{ii}^* \beta_i}{\beta_i' W_{ii} \beta_i} \quad \text{so} \quad \mu_i \geq 0. \qquad (6)$$

But (Chapter 2.3.4) $\mu_i \geq 1 \quad i = 1 \ldots M$.

Thus all turning points[1] of (4) are obtained by choosing $\hat{\beta}_i$ as the eigenvectors corresponding to the eigenvalues of $W_{ii}^{-1} W_{ii}^*$. Clearly from (6) by substituting μ_i into (4), the global maximum is achieved when $\hat{\beta}_i$ is so chosen with $\mu_i = \lambda_i$ the smallest of those eigenvalues. Uniqueness obviously follows λ_i being distinct. If it is not distinct LIML estimates will generally be $\hat{\beta}_i = \infty$ (Farebrother (1974)). Since λ_i is also used in (1) these results carry over to FIML. #

Remark:

Corollary 1 marks the distinct computation advantage of (1) and (2.32 - 4) over (2.41), (5.10) and direct maximisation algorithms. Neither checking over solutions or potential nonconvergence affect the computational burdens. However no relaxation of the assumptions of Section 2.2 is possible which is a noted disadvantage and leaves open the question of numerical stability, which according to Fareborther (1974) can be very important. Moreover (1) and (2.32 - 4) are only available for (2.5) with no further nonlinearities.

---

1. $\frac{\partial^2 L}{\partial \beta_i \partial \beta_i'}$ is generally indefinite; any combination of turning points may result.

Corollary 2:

FIML ≡ LIML, hence (2.41) ≡ (5.10), in the i^{th} equation when it is just identified.

Proof:

Using (2.32), if i^{th} equation is just identified then $\lambda_i = 1$ otherwise $\lambda_i > 1$ (Dhrymes (1970) proposition 7.1). Then FIML and (2.41) are identical to 3SLS in that equation by (2.48). But it is well known that 3SLS ≡ 2SLS in this case (Dhrymes (1970) Theorem 4.13). If $\lambda_i = 1$ LIML and (5.10) are also identical to 2SLS by (2.47). #

Remark:

We therefore include just identified equations in the separate block of comment (b) of the previous section for IIV and maximum likelihood estimators. Corollary 2 is invariant to the IIV family member chosen.

7.4 Discussion

7.4.1

Estimates (1) and (2.32 - 4) exist only if T > max (K,M) or T > K respectively. Despite Corollary 1, it is emphasised that (2.41) and (5.10) are particularly useful, even for estimating (2.5), if the data is limited because of their increased degrees of freedom. They may be the only maximum likelihood estimators available. For large models also they may be useful if computing capacity is limited.

We have seen that IIV estimators give estimates identical to maximum likelihood for identical starts. In general for any start all iterative estimators produce a sequence of estimates which converges stochastically on those of maximum likelihood if (a) the algorithm is convergent and (b) the likelihood function is defined. We have only then to find the global maximum. To achieve (a) we may attempt many different algorithms with and without accelerations. Although (2.41) and (5.10) give an IV interpretation of maximum likelihood estimates and also allow IIV (and asymptotically FP)

to have a maximum likelihood interpretation, these interpretations are strictly only applicable if the various estimators are defined (i.e. sufficient data for (1) and (2.32 - 4); and a convergent algorithm for the iterative estimators). Therefore these estimators are equivalent with respect to their alternative interpretations. Later we show that we can always find some convergent algorithm (Theorem 14.1) and hence the IV interpretation of maximum likelihood estimators always exists, but the reverse interpretation is "equivalent" unless T > K or T > max (K,M) as appropriate. These statements, as always for nonlinear iteration, presuppose the assumption that a fixed point Y* exists. By Corollary 1 this is equivalent to the assumption of the existence of a strict maximum to (2.35), i.e. (2.42) holds. This point is extended in Chapter 14, and repeated in another form in Theorems 12.25, 12.26.

7.4.2

Our analysis has obtained some results in iterative estimation for consistent starts and all iterates, and a few at convergence given inconsistent starts. Inconsistent starts are examined again in Chapter 11. Some examples commonly suggested are

(a) $Y^{(0)} = Y^{+}$ - the 2SLS start; available if T > K; consistent.

(b) $Y^{(0)} = Y$ - the OLS start (which is also the generalised inverse start, by Chapter 4.3, if $T \leq K$); satisfies the criterion of availability; inconsistent.

(c) $Y^{(0)}$ = constant -usually $Y^{(0)} = 0$ is taken; the zero start; all are consistent.

(d) $Y^{(0)}$ = a set of dummy variables; consistent.

(e) $Y^{(0)} = ZP(g)$; P(g) obtained by a preliminary regression on g principle components (or factors); consistent unless 2SPC is used.

To this list we have added various SOIV, (4.24) and (4.26) starts and have recommended (4.26) in particular. Thus all the estimators mentioned in this study may be subsumed into one scheme of estimation. Any estimator may

be obtained by specifying $\{Y^{(0)}$, the total number iterations required, the normal equation type, the information class, and the algorithm type$\}$. This quintuple specifies the estimator space.

7.4.3

It is useful to distinguish restricted from unrestricted estimators, and also estimators at convergence from earlier iterates. Unrestricted estimators (those that use Π^+ in the instrument or regressor formation) cannot be iterated to convergence and often are not available. Restricted estimators may be obtained from unrestricted ones, or independently, and satisfy the criterion of availability. The distinction is necessary as the restricted estimators will have different small sample properties; they may be influenced by their start variables in both small and large sample properties. If they are iterated to convergence they may be independent of their starts. The distinction is therefore only drawn for estimators not fully iterated to convergence to indicate their dependence on starts and possible small sample differences, but not to involve the problems of finding a global solution. Moreover since restricted estimators use the restrictions in β,Γ they might be more efficient in small samples (or generally have better[1] small sample properties), and also numerically more stable since they have no unrestricted reduced form parameters, which lack efficiency and are highly susceptible to multicollinearity. Against this a restricted estimator requires the specification of the whole model, and therefore any errors, or biases induced, are propogated throughout the entire system. Some equations are always less reliable than others and some misspecification is inevitable, so these are important practical points.

7.4.4

Finally, some iterative estimators may be considered as k-class estimators. Farebrother (1974) has provided an explanation of the frequently observed instability of LIML estimates in this context. He notes a LIML

---

1. In the sense their distribution approaches the asymptotic one faster since T - M > T - K. We cannot say however if the approach is monotonic.

estimate may be written as a sum of terms which are the ratio of reduced form parameters to the difference of the two smallest eigenvalues involved in the eigenproblem (2.32). When these eigenvalues approach each other the estimate explodes, and so would the iterated estimate. Since LIML and FIML are restricted estimators (like all iterated ones) they may be sensitive to misspecification. It is possible that the instability and lack of robustness in LIML or FIML, which has also been observed for some iterative estimates (e.g. Salmon and Eaton (1975), Bergstrom (1974)), is partly due to these two characteristics they have in common. In particular Salmon and Eaton (in testing a large number of estimators) find RFP estimates of a whole model to be especially sensitive and unstable. They conclude this is due to one or two badly specified and volatile equations in their model of 96 equations. Some further examples were reported in Chapter 3. 2SLS as both an unrestricted estimator, and a linear combination between LIML and OLS, Oi (1969), would necessarily appear more stable in poorly specified models. The same applies to 3SLS in relation to FIML and GLS as indicated by the following.

Lemma 2:

3SLS is the same linear combination of FIML and GLS as 2SLS is of LIML and OLS.

Proof:

Let $\Sigma = LL'$, L triangular and nonsingular, so that 3SLS estimates are given by $(\hat{X}^{+*\prime}\hat{X}^{+*})^{-1}X^{+*\prime}y^{*}$ where $y_i^{*} = L^{-1}y_i$ and $X_i^{+*} = L^{-1}X_i^{+}$; FIML by the same transformation in (1); and GLS by $(\hat{\tilde{X}}'\hat{\tilde{X}})^{-1}\hat{\tilde{X}}'\hat{\tilde{y}}^{*}$ where $\tilde{X}_i = L^{-1}X_i$. Now the proof follows that of Oi (1969) applied directing to these transformed estimators. The generalisation to FI - AR versions is obvious. #

7.5 Asymptotic inefficiency of limited information estimators when full information conditions apply.

It is often thought that an iterated 2SLS estimator (I2SLS) will be asymptotically more efficient since it utilises the information of the

specification restrictions. A simple proof using Theorem 5.1 (a) shows this is not so. It was suggested by Dhrymes and Pandit (1972) that I2SLS was of different efficiency (sometimes greater, sometimes less) to 2SLS. It is shown that their expression for the efficiency difference results from the implicit assumptions of limited information estimators, i.e. that Σ is diagonal and that every equation is just identified. Their expression measures the efficiency lost in using limited information estimators when those assumptions are not appropriate.

If $\hat{\beta}$ and $\hat{\Gamma}$ are the 2SLS estimates, and hence consistent, we have by Theorem 4.8(b) asymptotic efficiency and consistency for all subsequent iterates of I2SLS (regressors being formed through (5.18(a)) since they are RFFP with 2 2SLS start. What follows is, by Theorem 5.1, valid for all iterative estimators at each step that is consistent and efficient because only asymptotic properties are at issue.

It must be shown that I2SLS has exactly the same expression for asymptotic efficiency to confirm our results so far. From (4.10) I2SLS has first iterate

$$\tilde{\delta}_i - \delta_i = (\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'(\tilde{V}_i\tilde{\beta}_i + u_i); \quad \tilde{X}_i = [Z\hat{\Gamma}\hat{\beta}^{-1} \vdots Z]_i \tag{7}.$$

But $\tilde{V}_i = pU\tilde{\beta}_i^{-1}$, $p = I - X(X'X)^{-1}X'$, and $\tilde{\beta}_i^{-1} = \tilde{\beta}^{-1}S_{1i}$.

Hence

$$T^{-\frac{1}{2}}\tilde{X}_1'(\tilde{V}_1\tilde{\beta}_1^{-1} + u_1) = T^{-\frac{1}{2}}\tilde{X}_1'(pU\tilde{\beta}_1^{-1}\beta_1 + u_1) \tag{8}$$

taking the first equation as an example. By consistency plim $T^{-\frac{1}{2}}\tilde{X}_i'U$ = plim $X_i^{+}'U = \bar{X}_i'U$ where $\bar{X}_i = [Z\Pi \vdots Z]_i$.

$\therefore$ plim $T^{-\frac{1}{2}}X_1'(\tilde{V}_1\tilde{\beta}_1 + u_1)$

$$\text{and plim } T^{-\frac{1}{2}} \left\{ \bar{X}_1' u_1 + \sum_{j=2}^{M} \left(\bar{X}_1 - \left(\frac{\bar{X}_1'\bar{X}_j}{T} \right) \left(\frac{\bar{X}_j'\bar{X}_j}{T} \right)^{-1} \bar{X}_j' \right) u_j T_{j1} \right\} \tag{9}$$

are distributed identically, using the zero restrictions in β_1 on $\tilde{V}_1$ and writing $T_{j1} = \tilde{\beta}_1^{-1} \beta_{j1}$.

Using the same steps as Dhrymes and Pandit we may conclude that the second term on the right of (9) is distributed asymptotically $N(0,\psi)$ where $\psi = \sigma_{11} \bar{M}_{11} + \psi^*$ and

$$\psi^* = 2 \sum_{j=2}^{M} T_{j1} \sigma_{1j} (\bar{M}_{11} - \bar{M}_{1j}\bar{M}_{jj}^{-1}\bar{M}_{j1}) \tag{10}$$

$$+ \sum_{j=2}^{M} \sum_{i=2}^{M} T_{i1}T_{j1}T_{ij} (\bar{M}_{11} - \bar{M}_{1j}\bar{M}_{jj}^{-1} \bar{M}_{j1} - \bar{M}_{1i} \bar{M}_{1i} \bar{M}_{ii}^{-1} \bar{M}_{i1}$$

$$+ \bar{M}_{1j} \bar{M}_{jj}^{-1} \bar{M}_{ji} \bar{M}_{ii}^{-1} \bar{M}_{i1})$$

where $\bar{M}_{ij} = \lim_{T\to\infty} T^{-1}\bar{X}_i'\bar{X}_j$; by the Lindeberg-Feller theorem. Consequently 2SLS asymptotic efficiency is

$$T^{\frac{1}{2}}(\hat{\delta}_{.1} - \delta_{.1}) \sim N(0, \sigma_{11} \bar{M}_{11}^{-1} + \bar{M}_{11}^{-1} \psi^* \bar{M}_{11}^{-1}) \tag{11}$$

which is exactly that obtained by Dhrymes and Pandit for I2SLS under the same assumptions. There can therefore be no difference in their asymptotic efficiencies.

The question arises why is I2SLS not more efficient as suggested? Clearly ψ^* is a consequence of the special assumptions implicit in limited information estimators. Normally it is assumed $\Sigma = I$ and it is well known that limited information methods in working equation by equation treat every other equation as just identified. But notice in (10) if $\Sigma = I$ then all $\sigma_{ij} = 0 \quad j \neq i$ so

$$\psi^* = \sum_{j=2}^{M} T^2_{j1}\, \sigma_{jj}\, (\bar{M}_{11} - \bar{M}_{ij}\bar{M}_{jj}^{-1}\bar{M}_{j1}) \qquad (12)$$

and if all equations are just identified then

$$\bar{X}_i(\bar{X}_i'\bar{X}_i)^{-1}\bar{X}_i'\bar{X}_1 = \bar{X}_1 \qquad (13)$$

so both m atrices in ψ^* are null, and $\psi^* = 0$ in (12). Since 2SLS implicitly requires $\Sigma = I$ and works equation by equation using an unrestricted reduced form it imposes these assumptions and has the conventional efficiency expression. If these assumptions are both unjustified ψ^* shows the efficiency foregone by this imposition. (12) gives the efficiency loss due to working equation by equation in either method (identification restrictions were used in (9); and ψ^* less this quantity, the loss due to a nondiagonal covariance matrix (used in Lindeberg-Feller theorem). $\bar{M}_{11}^{-1}\,\psi^*\bar{M}_{11}^{-1}$ is positive semidefinite so the loss in efficiency from working equation by equation is always positive in overidentified equations. The loss due to a nondiagonal covariance matrix may be of either sign and any size depending on the true structural parameters β and Σ. A system estimator for the limited information case will always improve efficiency; whereas if it fails to allow for a nondiagonal Σ it can be more or less efficient depending on the sign of the second efficiency loss and if it outweighs the first.

7.6 Autoregressive Iterative Estimators

7.6.1

IIV and FP family estimators have been specified for both limited and full information classes. Here autoregressive versions for those information classes are given to complete the theory. This only involves specifying an algorithm to include estimates of the error covariance matrix so any acceleration may be used to calculate the next stage regressors or instruments. The same family members exist for these information classes as exist for the limited information case. We take up the ideas of Section 2.3.4.

Initially assume Z contains no lagged dependent variables. If $\phi = \text{diag } \Sigma_i$, limited information autoregressive estimators may be calculated as

$$\tilde{\delta}^{(s+1)} = (\hat{N}^{(s)'}\hat{X})^{-1}N^{(s)'}y \tag{14}$$

for IIV family members, or

$$\hat{\delta}^{(s+1)} = (\hat{N}^{(s)'}\hat{X}^{(s)})^{-1}\hat{N}^{(s)'}y \tag{15}$$

for FP family members.

In either case

$$\hat{N}^{(s)'} = \hat{X}^{(s)'}\,\hat{\phi}^{(s)-1} \tag{16}$$

where

$$\hat{\phi}^{(s)-1} = \text{diag } \Sigma_i^{(s)-1},\ \hat{X}^{(s)} = \text{diag } X_i^{(s)} \tag{17}$$

where $X^{(s)}$ is calculated by (5.18a) or (5.18b) possibly with accelerations as mentioned in Section 4.7.3. As always the estimation of $\Sigma_i^{(s)}$ may follow some assumption on the serial correlation proces.[1] The general theory of the transformation matrix L applies. The transformation matrix for this case could be used:

$$L^{(s)'-1}L^{(s)-1} = \phi^{(s)-1} \quad \text{where} \quad L^{(s)} = \text{diag } L_i^{(s)}; \tag{18}$$

In fact due to the simplicity of $L_i^{(s)-1}$ it will be computationally efficient to use it with the usual FP or IIV programmes, thus avoiding extra computation. The general formulation is to find

$$\Sigma_i^{(s)} = T^{-1}(y_i - X_i\,\delta_i^{(s)})(y - X_i\,\delta_i^{(s)})' \tag{19}$$

---

1. For example a first order scheme in u_i would imply at step S, $\Sigma_i^{(s)-1}$ given by the well known matrix (8.32) Johnston (1972) where $\rho_i^{(s)}$ replaces ρ_i being the OLS regression coefficient of $u_{t,i}^{(s)}$ on $u_{t-1,i}^{(s)}$. The sphericality transform matrix $L_i^{(s)}$ is given in (8.35) Johnston (1972) with $\rho_i^{(s)}$ replacing ρ_i. Iterations then proceed on $\delta_i^{(s)}$ and $\rho_i^{(s)}$ $i = 1 \ldots M$.

and calculate $L_i^{(s)}$ by Nagar's method similarly to corollary 1 of Theorem 2.21.

By Theorem 5.1(a), (14) and (17) are an autoregressive verions of LIML. Similar to Theorem 5 we could also use

$$\hat{\delta} = (\hat{X}'\hat{\phi}^{-1}(\text{diag } I - k_i Q)\hat{X})^{-1}(\hat{X}'\hat{\phi}^{-1}(\text{diag } I - k_i Q)y) \qquad (20)$$

where $\hat{\phi}^{-1} = \lim_{S\to\infty} \phi^{(s)-1}$ is best obtained from $\lim_{S\to\infty} L^{(s)}$.

Under the assumptions of this section (15) and (17); (14) and (17) and hence (20) obviously have exactly the properties, in their information class, claimed for the corresponding limited information estimators. Such theorems may be proved from the sphericality transformation matrix $L^{(s)}$ and then the appropriate theorem of the limited information case applied to the transformed data. Note, in particular, at convergence we are merely using GLS or GIV estimation of REID equations since, there, consistency of $\delta^{(s)}$ automatically implies the consistency of $\Sigma_i^{(s)}$. For consistent starts we can claim these properties for earlier iterates.

We can construct full information autoregressive estimators. At convergence we require $\hat{N}^* = \hat{X}^* \phi^{*-1}$ in either (14) or (15); ϕ^* from Section 2.4. The most obvious estimator sets $\phi^{*(s)} = T^{-1}\{y - \hat{X}\delta^{(s)}\}\{y - \hat{X}\delta^{(s)}\}'$ and finds $L^{(s)}$ by Nagar's method. $L^{(s)}$ being triangular is comparatively easy to invert and to transform $\hat{X}^{(s)}$, y (and $\hat{X}$ for (14)). However following Parks (1967), who analysed the nonsimultaneous case, we can simplify the computation somewhat.

$$\phi^* = E(\text{vec } U)(\text{vec } U)' = Q^*\phi Q^{*\prime} \qquad (21)$$

which has typical T x T submatrices $\sigma_{ij}Q_iQ_j'$ where $Q^* = \text{diag } Q_i$ and Q^* is defined as the sphericality transformation associated with the serial correlation component of the error process. That is $U = Q^*\varepsilon$ such that $E(\varepsilon'\varepsilon) = \Sigma$

but $E(\varepsilon_i\varepsilon_i') = \sigma_{\varepsilon_i}{}^2 I$. Then writing $\Sigma = R^*R^{*\prime}$, R^* the sphericality transformation matrix associated with the contemporaneous component of the error process, we have $L = Q^*(R^* \otimes I)$ for $LL' = \phi^*$ the overall sphericality transform. Thus $\hat{N}^*$ is efficiently computed with

$$\phi^{*-1} = Q^{*\prime-1}(R^{*\prime-1} \otimes I)(R^{*-1} \otimes I)Q^{*-1} \tag{22}$$

since inversion involves a number of operations given by the cube of the order of the matrix. M inversions of order T and one of order M is preferable to one of order TM (any T,M > 1); especially if R and Q_i are all triangular (calculated by Nagar's method). This suggests in (14) or (15) we use $\phi^{*(s)-1}$ obtained from (21), where $R^{*(s)}$ is given by $\Sigma^{(s)} = R^{*(s)}\ R^{*(s)\prime}$ with $\Sigma^{(s)} = T^{-1}(Y\beta^{(s)} + X\Gamma^{(s)})'(Y\beta^{(s)} + X\Gamma^{(s)})$ and $Q_i{}^{(s)}$ given by $L_i{}^{(s)}$ i.e. $\Sigma_i{}^{(s)} = Q_i{}^{(s)}Q_i{}^{(s)\prime}$ in the general formulation given above (19). The last remark is justified since ϕ^* in (21) can be written $\phi^* = \Sigma \otimes BB'$ where $B' = (Q'_1 \ldots Q'_M)$.

Once again we can also write autoregressive FIML as (20) with $\hat{\phi}^{-1} = \lim_{S\to\infty} \phi^{*(s)-1}$. Similarly, also, the full information autoregressive estimators proposed have exactly the properties in the information class claimed for the corresponding limited information estimators of Section 4.7 and Chapter 5 for precisely the same reasons. These properties are explicitly proved by Parks (1967) for nonsimultaneous equations. Those proofs therefore may be applied exactly to the REID specification (2.30). It is clear autoregressive 2SLS and 3SLS are obtained by using X^+ in place of $X^{(s)}$ in the respective estimators above. Evidently the maximum likelihood equivalances remain. $\frac{\partial(\text{vec } U)}{\partial y}$ has typical element $\frac{\partial u_{si}}{\partial y_{tj}} = 1$ if $i = j$, $-\beta_{ji}$ if that is nonzero, 0 otherwise and $s \neq t$.

Hence $\frac{\partial(\text{vec } U)}{\partial y} = \beta \otimes I_T$ and (2.35) may be written

$$L(\beta,\Gamma,\phi^*) = C + {}^{T}/_{2} \log\det \phi^{*-1} + T \log|\det \beta| - {}^{T}/_{2}\ \text{tr}[T^{-1}\phi^{*-1}(y-\hat{X}\delta)'(y-\hat{X}\delta)]$$

so by (22)

$$L(\beta,\Gamma,\phi^*) = C + {}^{T}/_{2}\{\log\det\Sigma^{-1} + \log\det(Q^*Q^{*\prime})^{-1}\} + T\log|\det\beta|$$

$$- {}^{T}/_{2}\,\mathrm{tr}[T^{-1}(\Sigma^{-1}\otimes I)(Q^*Q^{*\prime})^{-1}(y-\hat{X}\delta)'(y-\hat{X}\delta)] \qquad (23)$$

which adds an extra necessary condition to (2.39), amends the remainder, and yields the iterative estimator (14) with (22) in place of (2.41).

It is obvious that there are many other ways of designing the iterative schemes for these autoregressive estimators. In particular, whichever of (5.18a) or (5.18b) is used, we can either accept $\phi^{*(s)}$ from (22) to obtain new parameter estimates and thus $Y^{(s+1)}$ directly or we can iterate $\phi^{*(s)}$ to convergence, with $Y^{(s)}$ fixed, before using the latest estimates to calculate $Y^{(s+1)}$. Clearly there is no difference to the asymptotic properties of the estimates.

In Chapter 12 it is shown that the local convergence condition is that the spectral radius of a certain matrix be less than unity. For all iterative estimators that matrix is $0,\beta$ or some rational function of only β and the acceleration parameters. Therefore the convergence properties are identical under the various schemes - at least locally. Naturally computation per iteration is increased if $\phi^{*(s)}$ is iterated with $Y^{(s)}$ fixed, and thus the latter scheme is always, at least locally, computationally inefficient relative to the former; also globally to the extent we can obtain global convergence conditions in Chapter 11.

The reader is reminded that we estbalished a hierarchy of information classes in Section 2.4 and thus the estimators in this section dominate those from lower information classes.

7.6.2.

We now relax the restriction that Z has no lagged dependent variables. One estimator is that discussed at (2.50 - 1) in the framework of the

iterative estimators above, in the information class appropriate to the form of the covariance matrix (2.51). In that notation

$$y = y_{-1}r + (\hat{X} - \hat{X}_{-1}r)\delta + vec(\varepsilon) \tag{24}$$

where r = vec (R). (2.50) indicates that the counterpart to (5.18a) is

$$y_i^{(s+1)} - y_i^{(s+1)}\,\rho_i^{(s)} = (Y_i^{(s)} - Y_{i,-1}^{(s)}\,\rho_i^{(s)})\beta_i^{(s)} + (Z_i - Z_{i,-1}\,\rho_i^{(s)})\gamma_i^{(s)} \tag{25}$$

and to (5.18b) is (25) but with $Y_i^{(s+1)}$ and $Y_{i,-1}^{(s+1)}$ on the right. Having obtained $Y^{(s+1)}$ we construct $\hat{X}^{(s+1)}$ and estimate (24) as in (14) or (15) to obtain $\delta^{(s+1)}$ and $r^{(s+1)}$. Noting $E(\varepsilon_i\varepsilon_i') = \sigma_\varepsilon^2 I \quad i = 1 \ldots M$ we may use

$$\Sigma^{(s)} = T^{-1}[(Y - Y_{-1}R^{(s)})\beta^{(s)} + (Z - Z_{-1}R^{(s)})\Gamma^{(s)}]'[(Y - Y_{-1}R^{(s)})\beta^{(s)} + (Z - Z_{-1}R^{(s)})\Gamma^{(s)}] \tag{26}$$

if (2.51) is not diagonal. Note substituting X^+ for $X^{(s)}$ in (24) gives 2SLS and 3SLS variants. However naturally the iterative versions economise on degrees of freedom. (2.50) requires T > 2K + M whereas (24) requires only $T > \max_i(r_i + s_i + 1)$ a reduction which for most models (certainly for 8 of the examples considered in this study) will be essential if the method is to be feasible. We note also that we do not replace y_{-1} by $y_{-1}^{(s)}$ nor $\hat{X}_{-1}$ by $\hat{X}_{-1}^{(s)}$ in (24) for the same reasons that we do not replace $\hat{Y}_{-1}$ in $\hat{X}$ nor Y_{-2} in $\hat{X}_{-1}$ - it would merely confound our attempts to remove serial correlation.

Once again the estimators above are not unique. It will be noted that the 2SLS/3SLS versions above are similar to those of Dhrymes et al (1974) and are by their results, asymptotically, maximum likelihood. However they also propose an IIV variant in which, rather than attempt to remove the serial correlation itself, if a consistent start may be found then Y and Y_{-1} may be replaced in the instrument set by $Y^{(s)}$ and $Y_{-1}^{(s)}$ from the final form

$$Y^{(s)} = Z^*[\Pi_2^{(s)} : \Pi_3^{(s)}] [I - \Pi_1^{(s)}\theta]^{-1} \qquad (27)$$

where θ is the lag operator. Iteration then may continue to convergence at Y^*. The notation is that of (2.50) i.e. $\Pi_1^{(s)} = \beta^{(s)}{}_R{}^{(s)}\ \beta^{(s)-1}$; $-\Pi_2^{(s)} = \Gamma^{(s)}\ \beta^{(s)-1}$; $\Pi_3^{(s)} = \Gamma^{(s)}\ R^{(s)}\ \beta^{(s)-1}$. Clearly this method shares the properties of the above method by a reformulation of the REID specification (2.30) eliminating any correlation between RHS variables and stochastic errors. To be efficient, using (27), we must employ the autoregressive estimators of the previous section. The method obviously extends to FP family members. However this method has three major disadvantages. First it cannot be generalised to a model with more nonlinearity than (2.5) since (27) is only possible when the reduced form is defined. But (25), with $Y_i^{(s)}$ or $Y_i^{(s+1)}$ on the right, can be simulated by any iterative solution of nonlinear equations for a nonlinear model (see (11.29) and (11.30)). Secondly it precludes IIV and FP family members which use (5.18a). Thirdly, and most important, it clearly involves an increased computational burden at each step - we can set the inversion of $(I - \Pi_1^{(s)}\theta)$ against the extra reestimation $r^{(s)}$ and calculation of $\delta^{(s)}$. Thus the Dhrymes et al. method is not recommended.

CHAPTER 8: AN ANALYSIS OF THE PROBLEM OF COMPLETE MODEL SOLUTION I - LINEAR, FIRST ORDER STATIONARY METHODS.

8.1 The use of algorithms

The notation of Part III does not follow that of Chapter 2 but we return to that notation in Parts IV and V.

Consider a general nonlinear simultaneous system

$$F_i (Y_{ti} ; Y_{t-1,i} ; Z_{tk}) = U_{ti} \qquad (1)$$

$i = 1 \ldots n$ equations, $j = 1 \ldots \ell$ lags, $k = 1 \ldots g$ exogenous variables. We have future values of the exogenous variables.

The only concise statement on the problem of complete model solution readily available to econometricians is that contained in Fromm and Klein (1967). There it is pointed out that iterative solution methods are equivalent to taking a linear approximation to (1) and solving the linear system iteratively - the Newton method. That is, setting exogenous, predetermined and stochastic parts of (1) at known values we obtain

$$f_i (Y_{t1} \ldots Y_{tn}) = 0 \quad i = 1 \ldots n \qquad (2)$$

Taking a first order Taylor's series expansion about some $Y_j^{(0)}$,

$$f_i (Y_1^{(0)} \ldots Y_n^{(0)}) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial Y_j} \Bigg|_{Y = Y^{(0)}} (Y_j - Y_j^{(0)} = 0 \qquad (3)$$

we obtain n linear equations in Y_j $j = 1 \ldots n$ unknowns. We may solve linearly for Y_j, call it $Y_j^{(1)}$, and repeat until $\left| \frac{Y_j^{(s)} - Y_j^{(s-1)}}{Y_j^{(s+1)}} \right| < \tau$

each j, where τ is some tolerance level. Thus while all but one Y_j are taken as fixed $Y_j^{(0)}$ we are evaluating the partial derivatives of f_i i=1....n with respect to Y_j at $Y_j^{(0)}$ for each j = 1 .... n in turn. Upon normalisation (2) becomes $Y_{ti} = g_i\ (Y_{t1} \ldots Y_{tn})$ and we could make iteration on (3) quicker by reducing the Newton procedure to solving at the s^{th} iteration only,

$$Y_i^{(S)} = g_i\ (Y_1^{(s-1)} \ldots Y_n^{(s-1)}) \text{ each } i \qquad (4)$$

(4) is much quicker computationally since it does not involve extensive evaluation of derivatives and matrix inversion at each iteration, though it is somewhat less accurate at each iteration. The problem with (4) is there is now no guarantee of convergence even locally, and, even if it does converge that it will to the desired solution. However, we can invoke a contraction mapping proof to find convergence criteria to the correct solution if $Y^{(0)}$ is in a neighbourhood of the desired solution Y*. (4) is also very sensitive to normalisation, there being in general n! different possible normalisations to the system.

8.2 The current knowledge

We are only concerned with the endogenous part of the system. If that were linear, $BY_t + F(Y_{t-1} \ldots Y_{t-\ell} Z_t) = 0$ has solution $Y_{t+1} = -B^{-1}F(Y_t$ .. .. $Y_{t-\ell+1}\ Z_{t+1})$ so each period might be handled linearly and a sequence of solutions obtained. If there were block recursiveness in B, and there was some method of solving within the blocks then we could solve the system using the usual recursive procedures. If the system were linear in blocks of equations, some algorithm to solve between blocks would be the only requirement. The method (4) is merely treating the number of blocks as equal to the number of equations. Therefore we need only consider equations within

blocks of any block recursive scheme that may exist, i.e. we assume B an irreducible matrix. Steward (1962) has suggested a number of algorithms which may be used to identify any block recursive scheme. This should be done first but is unlikely to lead to much practical improvement. The recursive blocks which may be separated off have usually been found to be very small (e.g. Brookings, Wharton or London Business School models, Salmon & Eaton (1974)). Even if economic behaviour can, in the last analysis, be considered recursive it is most unlikely with data at quarterly, or worse at annual intervals that it can be represented in such terms.

We now write the linear system given by (2) and (3)

$$AY_t = b_t \qquad (5)$$

where A has unit diagonal and elements - $g_{ij}(\cdot)$, and b_t is the fixed pre-determined vector, one element for each equation evaluated at t. The separability of $g_i(\cdot)$ is ensured by linear approximation (3). A variant on (4) is the Gauss-Seidel method, where iterating in (4) we replace each RHS $Y_j^{(s-1)}$ by $Y_j^{(s)}$ as it is calculated and use it in the determination of the remaining elements of $Y^{(s)}$.

$$\therefore \quad Y_i = g_i(Y_1^{(s)} \ldots Y_{i-1}^{(s)} Y_{i+1}^{(s-1)} \ldots Y_n^{(s-1)}) \quad i = 1 \ldots n \qquad (6)$$

The advantages are a much quicker iteration process, half as much storage requirement in the computer so larger systems may be satisfactorily handled, and no added complexity in programming. Disadvantages are that the convergence now depends on the ordering of the equations as well as the normalisation and convergence conditions are harder to apply. A further modification which may be used on either (4) or (6) is the addition of a damping factor α. We use instead

$$Y_i^{(s)} = \alpha g_i(Y_1^{(s-1)} \ldots Y_n^{(s-1)}) + (1 - \alpha)\, Y_i^{(s-1)} \qquad (7)$$

or

$$Y_i^{(s)} = \alpha g_i(Y_1^{(s)} \ldots Y_{i-1}^{(s)}\, Y_{i+1}^{(s-1)} \ldots Y_n^{(s-1)}) + (1 - \alpha) Y_i^{(s-1)} \qquad (8)$$

It is found that for some α this frequently improves the convergence properties of (4) and (6).

There is some confusion in the literature on the names of these various methods. Here (4) will be referred to as the Jacobi method, (6) Gauss-Seidel, (7) Jacobi over-relaxation (JOR), (8) successive over-relaxation methods (SOR).

8.2.2

An iterative solution method is said to be linear and stationary of first order if it can be written

$$Y^{(s)} = GY^{(s-1)} + k \qquad (9)$$

where Y is an n-vector of unknowns, s the number of iteration, G is an n x n matrix of constants and k an n x 1 vector of constants. We may apply such methods to (5).

Theorem 1:

The method (9) is convergent from an arbitrary $Y^{(0)}$ if and only if the spectral radius $\rho(G) < 1$.

Proof:

If $\varepsilon^{(s)} = Y^{(s)} - Y^*$ then $\varepsilon^{(s)} = G^s\, \varepsilon^{(0)}$ and $G^s \to 0$ as $S \to \infty$ if and only if $\rho(G) < 1$. Then $\lim_{s\to\infty} Y^{(s)} \to Y^*$ which defines convergence. (By Theorem 1.4 in Varga (1962)). #

We also require A^{-1} to exist for a solution to be defined and then a completely consistent and unique solution Y^* follows for (5) (the terminology

is that of Young (1971)). The general nonlinear case is treated by Theorem 12.1 below.

We now define $A = D + L + U$ where D is a diagonal matrix, D^{-1} exists, L is the strictly lower triangular and U the strictly upper triangular matrix of A. In (5) we had defined A such that $D = I$. The Jacobi method has $G = D^{-1}B$ where $L + U = B$, JOR has $\alpha B + (1 - \alpha) I = B_\alpha$, Gauss-Seidel has $(I - L)^{-1} U = R$ and SOR $(I - \alpha L)^{-1} (\alpha U + (1 - \alpha)I) = R_\alpha$.

Convergence theorems obtained for these methods have been sufficient conditions for convergence. They may be summarised as follows:

Any method here converges if A is symmetric and positive definite or has nonpositive off diagonal elements with positive diagonal ones (which implies a nonnegative inverse). (Corollary 1 p.85 Varga (1962).)

These sufficiency conditions are not very useful. We have to check that we have a special kind of matrix, but it is most unlikely that a general econometric model can be relied on to be exactly an L, M or symmetric matrix (see Theorem 9.1) - or we have to check that certain matrices are positive definite which may be tedious. And this is to ignore the stochastic properties of any estimated endogenous coefficients. The only a priori restrictions we have on A are that it is n x n real matrix (which may have unit diagonals under some normalisation) and nonsingular and that its elements are endogenous coefficients B of the system, or separable linear approximations of fixed functions as in (3) so that we may consider a linear stationary first order method. In the latter case all the results which follow are locally valid in the sense of Theorem 12.1.

The rate of convergence of an algorithm is defined as, either an average rate of convergence = $\frac{1}{S} \log ||G^S||$ for some norm, or an asymptotic rate of convergence = $- \log \rho(G)$. We can see the number of iterations required for convergence $\simeq \log \tau / \log \rho(G)$, since $\rho(G) = \lim_{S \to \infty} (||G^S||)^{1/S}$ and the asymptotic rate is the limit of the average rate. Thus in optimisation we

shall aim to minimise $\rho(G)$ although a lower asymptotic rate may not imply a lower average rate at every stage. We do not specify the norms to be used in average convergence rates - many results are available in Young (1971) - as the spectral norm is the only generally applicable one for convergence analysis.

8.3 Stability conditions for polynomials

In Gantmacher (1959) details, and proofs, are given of the Routh test for the number of roots in the right hand half plane of the complex domain for any given real polynomial. The Routh test is the basis of the best known test of stability of a polynomial. The difference equation (9) is stable iff the characteristic equation of $G \Rightarrow \rho(G) < 1$. The Routh test applies to the stability of differential equation systems i.e. that G has no eigenvalues with positive real parts. It turns out (Theorem 5 below) that this test, with G + I in place of G, is what we require for a difference equation system (9).

Definition:

A polynomial is stable if the real parts of all its roots are negative.

The Routh-Hurwicz criterion: the necessary and sufficient condition for stability of the real polynomial

$$f(Z) = a_0 Z_n + a_1 Z^{n-1} + \ldots + a_n = 0; \quad a_0 \neq 0; \tag{10}$$

is $\Delta_1 = a_1 > 0$, $\Delta_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix} > 0$, $\ldots$, $\Delta_n = \begin{vmatrix} a_1 & 2_3 & a_5 & \ldots & 0 \\ a_0 & a_2 & a_4 & \ldots & 0 \\ 0 & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ 0 & & & & a_n \end{vmatrix} > 0$

if $a_0 > 0$. Write $\phi(\lambda) = |G - \lambda I|$. To test $\lambda_{max} < 1$ in real part we need only transform $\phi(\lambda)$ to the form (10) by $Z = \lambda + 1$, i.e. we test G + I not G itself. The coefficients a_i in $\phi(\lambda)$ are the sum of all principle minor determinants of G, order i. The remaining difficulties, Re $(\lambda_{max}) \lesseqgtr -1$, or if $|\text{Re}(\lambda_{max})| < 1$ but $\rho(G) \gtrsim 1$ due to the imaginary parts, have not been considered. This it turns out is the function of α and enables us to apply the Routh-Hurwicz criterion directly to G + I.

An improvement on the Routh-Hurwicz criterion for stability of (10) is the derived Lienard-Chipart criterion (also proved by Gantmacher). Using (10) under the same transformation, necessary and sufficient conditions for stability are <u>one</u> of the following four must be satisfied for $a_0 > 0$:

(a) $a_n > 0, a_{n-2} > 0, \ldots$ <u>and</u> $\Delta_1 > 0, \Delta_3 > 0, \ldots$

(b) $a_n > 0, a_{n-2} > 0, \ldots$ <u>and</u> $\Delta_2 > 0, \Delta_4 > 0, \ldots$

(c) $a_n > 0, a_{n-1} > 0, a_{n-3} > 0, \ldots$ <u>and</u> $\Delta_1 > 0, \Delta_3 > 0, \ldots$

(d) $a_n > 0, a_{n-1} > 0, a_{n-3} > 0, \ldots$ <u>and</u> $\Delta_2 > 0, \Delta_4 > 0, \ldots$

Samuelson (1941) has given necessary and sufficient conditions for (10) to have $|Z_i| < 1 \quad i = 1 \ldots n$. Farebrother (1974) has given a method of conducting the test. However, from Theorems 5 and 10 below, this is not the test we want for the general case.

Even Lienard-Chipart tests are arduous for small n, and involve more computation than direct computation of $\rho(G)$ for large n. Meyer-Spasche (1973; a number of the necessary proofs are below or in Meyer-Spasche (1972)) has provided a computationally efficient test of the stability of the characteristic polynomial of any matrix. With it we may test the stability of the polynomial as required for the convergence of (9) (in the sense of Theorems 5 and 10). Neither the solution of the characteristic polynomial nor the evaluation of any determinants are required. Instead one matrix

multiplication and reducing two matrices to Hessenburg forms are involved, which may be done very rapidly by Householder transforms (see Businger and Golub (1965)). For any size model the Meyer-Spasche test will be computationally efficient especially in contrast to calculating $\rho(G)$ directly.

The Meyer-Spasche test rests on Lyapunov's stability theorem. Let N be the set of all negative definite matrices and π the set of all positive definite ones. Let M be an arbitrary (n x n) real matrix.[1] Then

Theorem 2:

M is stable (i.e. has stable characteristic polynomial) iff $\exists$ $H \in \pi$ such that $MH + HM' \in N$ (proof: Theorem 2 Taussky (1964)). #

Let M have q eigenvalues with negative real parts, r with positive ones, v that are purely imaginary; $n = q + v + r$. Inertia $In(M) = (q, v, r)$. If we can find an H as in Lyaopunov's theorem we can find one that is symmetric; then $In(M) = In(H)$ (proof: Theorems 1,2 Meyer-Spasche (1973)). Therefore calculate H, and make it Hessenburg form (i.e. tridiagonal, H being symmetric);

$$H = \begin{pmatrix} a_1 & b_1 & & & & \\ b_1 & \ddots & \ddots & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & b_{n-1} \\ & 0 & & \ddots & \ddots & \\ & & & b_{n-1} & a_n & \end{pmatrix}$$

Define the sequences f_i, p_i, v_i $\quad i = 1 \ldots n$.

$$f_o = 1; \quad f_1 = a_1; \quad f_i = a_i f_{i-1} - b_{i-1}^2 f_{i-2}$$

$$p_o = 0; \quad p_i = \begin{cases} p_{i-1} + 1 \text{ if } f_i f_{i-1} \neq 0 \\ p_{i-1} + 2 \text{ if } f_{i-1} = 0 \text{ and } b_{i-1} = 0 \\ p_{i-1} \text{ otherwise} \end{cases}$$

$$v_o = 0; \quad v_i = \begin{cases} v_{i-1} + 1 \text{ if } f_i f_{i-1} < 0 \text{ or } f_i f_{i-1} = 0, f_i f_{i-2} < 0 \\ v_{i-1} \text{ otherwise} \end{cases}$$

---

1. The results of this section apply to complex matrices also but we do not need them.

then $q = v_n$ and $r = p_n - v_n$ (provided f_i is restored to 1 after p_i and v_i were calculated whenever $b_i = 0$) and $v = n - q - r$. The test is $r = 0$.

H is obtained much faster if M is in Hessenburg form. The Meyer-Spasche procedure may be improved. We start with an arbitrary vector as the first row of H, h_1, where h_{11} is real. Then the first column of H, $h_{.1} = h_{1.}$. Now the first row of HM may be obtained. If M is upper Hessenburg.

$$h_{21} M_{11} + h_{22} M_{21} = (HM)_{21} = -(HM)_{12} \qquad (11)$$

may be solved for h_{22}. Thus in general when the $(j-1)^{th}$ row of H is known, so is that of HM. We can obtain the j^{th} column (hence row) from

$$h_{\ell j} = -\frac{1}{M_{j,j-1}} \sum_{k=1}^{j-1} h_{\ell k} M_{kj-1} + (HM)_{j-1,\ell} \qquad \ell = j \ldots . n. \qquad (11a)$$

(11a) depends on

Theorem 3:

If M is stable, $MH + HM' = N_1$ can be solved with some $H \varepsilon \pi$ for every $N_1 \varepsilon N$. (Proof: Taussky (1964) Theorem 3). #

In particular it can be solved for $N_1 = -I$ say. This theorem simplifies the test and improves the proofs. Therefore by Theorems 1 and 2 Meyer-Spasche (1973)), choosing a suitable N_1 we can have H symmetric with an arbitrary first row, provided M is stable. [This replaces part of Theorem 1, all of Theorem 2 of Meyer-Spasche (1973) and has been incorporated in (11) and (11a).]

In the case of equation (9) we should set $M = G + I$ in the Meyer-Spasche test.

8.4 The Jacobi and JOR methods

So far the basic convergence test would be to calculate $\rho(G)$ and check it was less than unity. The following theorems show it is not the right test and would not help in any reorganisation, or calculation of α, that may be required for a particular system.

We shall need:

Theorem 4:

The eigenvalues of a matrix are continuous, first order differentiable functions of the elements of that matrix, i.e. if X is a square n x n matrix with eigenvalue γ then $|\gamma' - \gamma| < (n + 2)(n^2 x)^{1/n}$ where $x = \max_{i,j} |X_{ij}|$.

Proof:

Ostrowski (1960) appendix K. #

The Jacobi method has $G = B = I - A$ if we normalise the system such that $D = I$. Then JOR has $B_\alpha = \alpha B + (1 - \alpha)I$. By Theorem 1, convergence tests imply the direct calculation of $\rho(G)$. The following results summarise the use of JOR methods.

Lemma 1:

$\alpha \neq 0$ in JOR becuase if it were $\rho(B_\alpha) = I$ which is not strongly convergent. #

Theorem 5:

Under the iteration method JOR, the necessary and sufficient conditions for convergence with an arbitrary starting vector are that the real part of each characteristic root of B shall either be less than unity, or greater than unity.

Proof:

The JOR method has an iteration process

$$Y^{(s)} = B_\alpha Y^{(s-1)} \tag{12}$$

for system (5). Let $\mu = a + ib$ be the dominant eigenvalue of B.

$\therefore$ $\rho(B_\alpha) = \hat{\mu} = \alpha\mu + (1 - \alpha)$. The necessary and sufficient condition of convergence of B_α is $\rho(B_\alpha) < 1$. Assume $a > 1$. But $\hat{\mu} = \alpha a + \alpha ib + (1 - \alpha)$ $= \alpha(a - 1) + 1 + i\,\alpha\, b$. Hence real $(\hat{\mu})$ and $|\hat{\mu}|$ both > 1 for all $\alpha > 0$. Supposing $a < 1$, for convergence we require the complex circle $r^2 = [\alpha(a - 1) + 1]^2 + \alpha^2 b^2 < 1$ for $\rho(B_\alpha) < 1$. Under these conditions whatever b there is an α small enough such that $r^2 < 1$. Therefore if $a < 1$ there always exists a permissable α such that $\rho(B_\alpha) < 1$ and JOR converges. Conversely if $a > 1$ then if $\alpha < 0$ implies $\rho(B_\alpha) < 1$ is possible. #

Corollary 1:

If $\rho(B) = \mu = a + ib$, then a Meyer-Spasche test on $\phi(Z)$, Z the roots of B + I, is a test for the necessary and sufficient conditions of convergence of JOR method for a given real B matrix, and μ_i of B no longer need be calculated. If Theorem 5 is satisfied B is said to be convergent. #

Comment:

By Theorem 5, a Meyer-Spasche test on B implies, if satisfactory, there exists some α for which convergence is assured and the problems mentioned of $a \lesssim -1$ and $|b| \gtrsim 1$ are removed. In geometrical terms as α decreases the centre of the circule of stability moves along the negative real axis, and its radius increases as α^{-1}. For small α convergence is assured but will be very slow and an alternative method would be attractive. Hereafter we consider $\alpha > 0$, since if $a > 1$ we merely take $-\alpha$ from the case of $a < 1$.

Theorem 6:

If B has eigenvalues $\mu_j = a_j + ib_j \quad j = 1 \ldots n$ then eigenvalues of B_α, λ_j, are minimised with $\alpha_j^* = \dfrac{-(a_j - 1)}{b_j^2 + (a_j - 1)^2}$ and each $|\lambda_j| < 1$ iff $0 < \alpha_j < 2\alpha_j^*$

Proof:

By Theorem 1,

$$\alpha_j < \frac{-2\,(a_j - 1)}{b_j^2 + (a_j - 1)^2} \qquad (13)$$

is required each $j = 1 \ldots n$ for convergence since $\alpha_j > 0$ and $\alpha_j \neq 0$ by Lemma 1. Similarly

$$r_j = +(\alpha_j{}^2[(a_j - 1)^2 + b_j{}^2] + \alpha_j\, 2(a_j - 1) + 1)^{\frac{1}{2}} \tag{14}$$

$$\therefore \quad \frac{dr}{d\alpha} = \tfrac{1}{2}(\alpha_j{}^2[(a_j - 1)^2 + b_j{}^2] + 2\alpha_j\,(a_j - 1) + 1)^{-\frac{1}{2}}.$$

$$(2\alpha_j[(a_j - 1)^2 + b_j{}^2] + 2(a_j - 1)) \;=\; 0$$

$$\therefore \quad \alpha_j{}^* = \frac{-(a_j - 1)}{b_j{}^2 + (a_j - 1)^2} \;=\; \tfrac{1}{2}\beta_j \tag{15}$$

where β_j is the expression in (13), the upper bound for convergence. Second order conditions may be checked for a minimum but it can be seen the coefficient of $\alpha_j{}^2$ in (14), a quadratic function of α_j, is positive. #

Theorem 7:

Let β = the maximum value α may take to allow convergence in (5). JOR is convergent all $0 < \alpha \leq \beta = 2 \min\ (\alpha_j{}^*)$.

Proof:

By Theorem 6 obvious from $\min_j\ (\beta_j)$ since $r_j{}^2 < 1$ is required all $j = 1 \ldots n$ for $\rho(B_\alpha) < 1$. #

Theorem 8:

The optimal relaxation factor for JOR methods, i.e. $\min_\alpha\ \rho(B_\alpha)$, is in the notation of Theorem 3 given by

$$\alpha^* = \min\left(\min_j \left| \frac{\tfrac{1}{2}\beta_j\, r_j + 1}{r_j + r_k} \right| ;\quad \beta_{k/2} \right) \tag{16}$$

where $r_k = \min_k \left(+ [(a_k - 1)^2 + b_k{}^2]^{\frac{1}{2}} \right)$, $k = 1 \ldots n$ and

$$r_j = + [(a_j - 1)^2 + b_j{}^2]^{\frac{1}{2}} \text{ for } j = 1 \ldots k - 1,\ k + 1 \ldots n.$$

Proof:

Substituting α_j^* of (15) into the relation $\lambda_j = \alpha\mu_j + 1 - \alpha$, each $j = 1 \ldots n$, we get

$$\lambda_j^* = \frac{-(a_j - 1)}{b_j^2 + (a_j - 1)^2}(a_j + ib_j) + 1 + \frac{(a_j - 1)}{b_j^2 + (a_j - 1)^2}$$

$$= \frac{b_j}{[b_j^2 + (a_j - 1)^2]^{\frac{1}{2}}} = b_j/r_j \qquad (17)$$

Further since $\lambda_j = \alpha(\mu_j - 1) + 1$, when $\alpha = 0$, though outside the permitted range, $\lambda_j = 1$ all $j = 1 \ldots n$ and we can write

$$|\lambda_j| = |-\alpha\hat{r}_j + 1| \qquad (18)$$

as we define $\hat{r}_j$ as the positive root and consider only $\alpha > 0$; $\therefore$ also $|\lambda_j^*| = |-\alpha_j^*\hat{r}_j + 1|$. Evidently $|\lambda_j|$ is, for increasing α, first a linearly decreasing function with gradient $-\hat{r}_j$ and after the minimum point it is a linearly increasing function with gradient $\hat{r}_j$, i.e. the function traces out an isosceles triangle $(0,1)$, $(\alpha_j^*, b_j/\hat{r}_j)$ the minimum, and $(\beta_j, 1)$ in the permitted interval $0 < \alpha \leq 2\alpha_j^*$. (For each $j = 1 \ldots n$ this is a continuous function and since all values are absolute there is a 1 - 1 mapping between $|\lambda_j|$ and r_j^2, which is differentiable as in (14).)

Restricting attention to situations where $a_j < 1$ all $j = 1 \ldots n$, $\hat{r}_j \neq 0$ any j. Now $\rho(B_\alpha)$ as a function of α is the frontier of the set $|\lambda_j| = |-\alpha_j\hat{r}_j + 1|$ bounded elsewhere by $\alpha = 0$, $|\lambda_j| = 0$ and $\alpha = \beta$. For $0 < \alpha \leq \beta/2$ the upper bound is $|\lambda_k|$ a decreasing function of α where $\hat{r}_k \leq \hat{r}_j$ all $j \neq k = 1 \ldots n$. If $\beta = \beta_k$ then $\alpha^* = \frac{1}{2}\beta_k$ since for all $\alpha > \frac{1}{2}\beta_k$, $|\lambda_k| > |\lambda_k^*|$ as α_k^* strictly minimises $|\lambda_k|$. If on the other hand $|\lambda_j| = |\lambda_k|$ for some $\alpha_0 < \alpha_k^*$ some $j \neq k$, then still $|\lambda_k| > |\lambda_j|$ for $\alpha < \alpha_o$ but also $|\lambda_j| > |\lambda_k|$ for $\alpha > \alpha_0$ as well as $\alpha > \alpha_k^*$. If $|\lambda_j| = |\lambda_k|$ only for $\alpha > \alpha_k^*$ all $j \neq k$

then we do better to use α_k^* where $|\lambda_k^*| < |\lambda_j| = |\lambda_k|$. Finally if $|\lambda_k| = |\lambda_j|$, $|\lambda_k| = |\lambda_l| \ldots. |\lambda_k| = |\lambda_m|$ for a subset j, l .... m of eigenvalues at $\alpha_{0j}, \alpha_{0l}, \ldots. \alpha_{0m} < \alpha_k^*$ j, l, ... m $\neq$ k then $\alpha^* = \min_{p,j=1,\ldots m} \{\alpha_{0p}\}$ since any $\alpha > \alpha^*$ thus defined implies $|\lambda_p| > |\lambda_p|$, the value of λ_p at this intersection, and any $\alpha < \alpha^*$ implies $|\lambda_k| > |\lambda_{0p}|$. But since $|\lambda_k| = |\lambda_p|$ p $\neq$ k only for $\alpha > \beta/2$ we have from (18) $|\lambda_k| = -\alpha\hat{r}_k + 1$, and λ_p is now the mirror image of $-\alpha\hat{r}_p + 1$ in the line $\lambda_p = -\alpha_p^* \hat{r}_p + 1$ i.e. $|\lambda_p| = \alpha\hat{r}_p + (-\alpha_p^* \hat{r}_p + 1) - 1$. Hence $|\lambda_k| = |\lambda_p|$ implies $\alpha_{0p} = \dfrac{-\alpha_p^* \hat{r}_p - 1}{-(\hat{r}_k + \hat{r}_p)}$.

Putting these three cases together and using the formula for α_{0p} we get (16) since $\alpha_p^* = \frac{1}{2}\beta_p$. #

Comment:

Theorem 8 may be used to give $\rho(B_\alpha^*) = \rho^*$ by substituting α^* into (18) for λ_k since that is one root at the intersection. We get

$$\rho^* = -\left(\frac{\alpha_j^* \hat{r}_j + 1}{k + r_j}\right)\hat{r}_k + 1 = \frac{(a_j - 1)\,\hat{r}_k\,\hat{r}_j + \hat{r}_j}{\hat{r}_k + \hat{r}_j} = \frac{[(a_j - 1)\hat{r}_k + 1]}{\hat{r}_k/\hat{r}_j + 1} \qquad (19)$$

where j is the index indicating the minimum in (16).

Corollary 2:

Therefore ρ^* depends on μ_j and μ_k only. Hence any renormalisation or reordering that may be undertaken to reduce $\rho(B_\alpha)$ must be done before α^* is determined to ensure overall minimal asymptotic rate of convergence. This is a lexographic ordering of priorities. #

Corollary 3:

In the case that all μ_j j = 1 .... n are real or that μ_k and μ_j selected by formula (16) are real, $\alpha^* = \dfrac{2}{r_k + r_j - 2}$. For a real μ_j, $\hat{r}_j = r_j - 1 = a_j - 1$ yet by (15) $\alpha_j^* = \dfrac{-1}{a_j - 1}$ which gives the result. Note (17) implies if μ_j is real $\lambda_j^* = 0$. If all μ_j are real, r_k is from $|\mu_k - 1|$ = minimum and r_j is from $|\mu_j - 1|$ = max for k, j = 1 .... n, because all the increasing functions of α are mirrored in $\lambda^* = 0$ so μ_j thus defined has

minimum α_j^*, maximum r_j and hence minimum α_{0j}. #

Corollary 4:

An underestimate of α^* is preferable to an overestimate since by Theorem 4 $r_k \leq r_j$ all $j \neq k$ and all the λ_j, λ_k are linear functions of α. The more $\hat{r}_j, \hat{r}_k$ differ the more this is preferable. The increase in iterations required for convergence is at least as little for an underestimate as for an overestimate and is strictly less if α^* is given by min α_{0p}. #

Corollary 5:

If JOR converges for some α_0 then it converges for $0 < \alpha < \alpha_0$ since by Theorem 4 $|\lambda_k|$ is the frontier of the convergent set and is monotonically decreasing between 0 and α^*. #

The above statements are all obvious from geometric considerations. We may caricature the situation:-

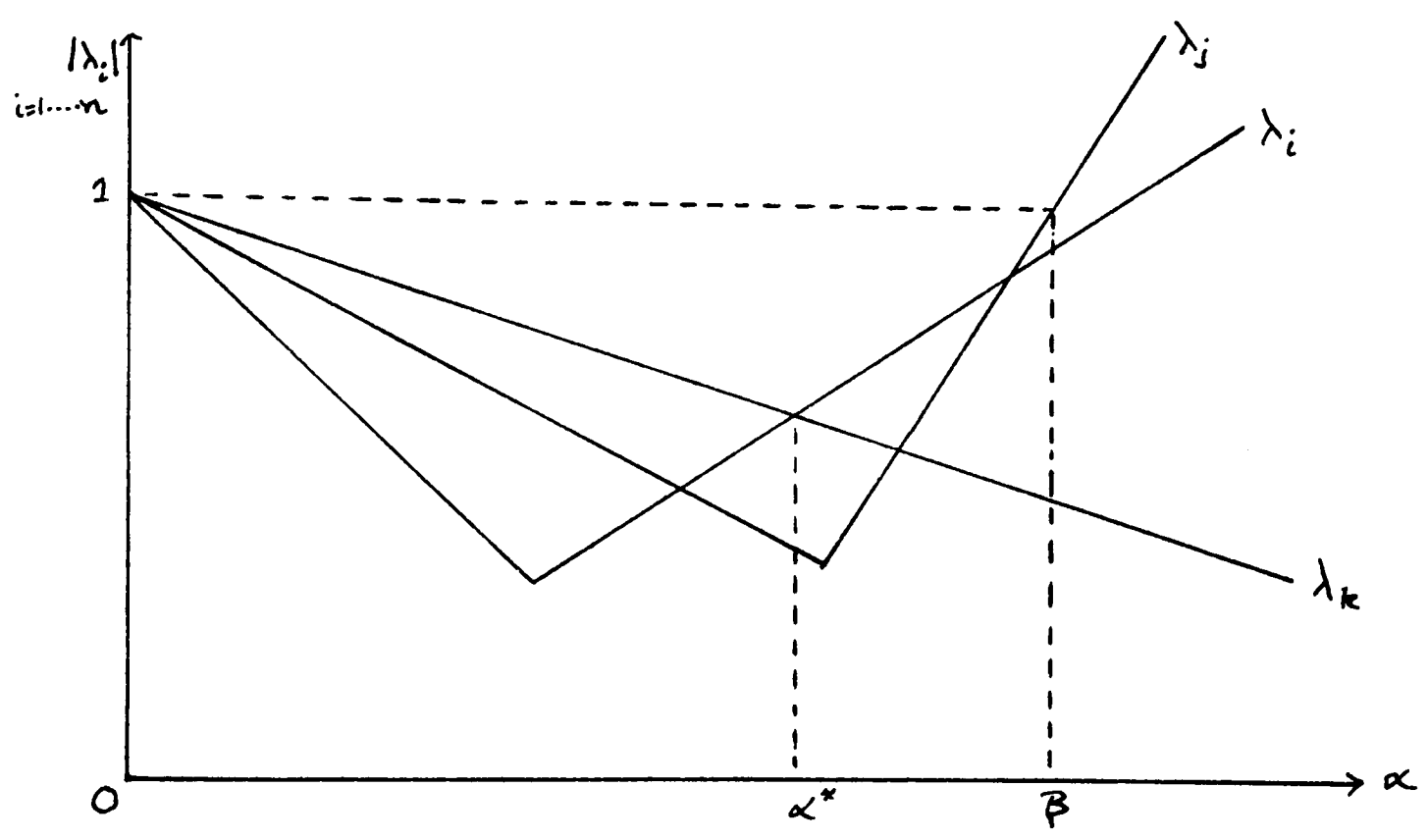


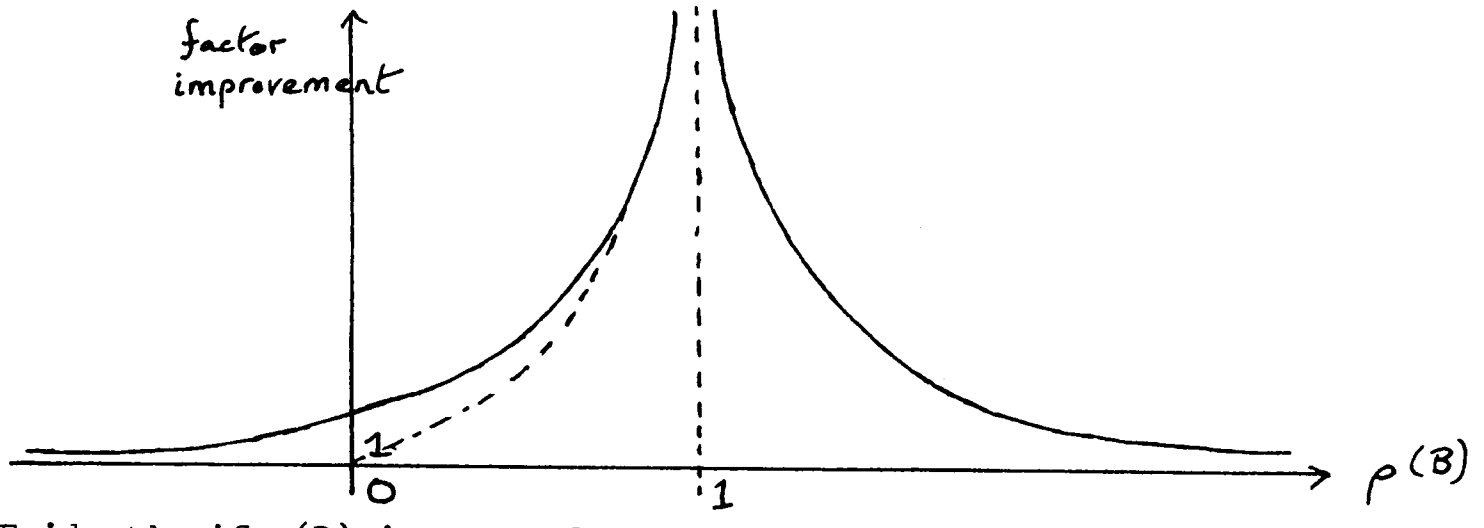
Figure 1 : The roots of B_α as functions of α.

It is possible to represent the formula (16) in different forms but this is most efficient for attaching to a general eigenvalue computer programme.

Corollary 6:

For each unrepeated, nonzero or one half of a conjugate pair root, $\hat{r}_j$, β_j and α_{0j} need only be calculated. These numbers are repeated for the other member of a conjugate pair; and zero roots have $\beta_j = 2$, $\hat{r}_j = 1$ and may be used directly in (16). #

We can analyse the improvement in convergence from using α^* using the definition in Section 2 for the number of iterations required for convergence. For fixed τ, the ratio of the number of iterations using α^* to that for Jacobi is (the factor improvement) $\log_e \rho(B)/\log_e \rho(B_{\alpha^*})$. The number of iterations required as a function of α is proportional to $-1/\log_e \rho(B_\alpha)$ i.e. figure 2[1] : convergence rate improvements due to α.



Evidently if $\rho(B)$ is very close to 1 we can expect very great improvements for any $\alpha \neq 1$ including those close to 1. However if $\rho(B)$ is substantially less than one, the scope for improvement is much more limited. The cases when calculation of α^* may not be expected to improve convergence much are:

(a) If α^* is very close to one - but it is arbitrary to compare against the case $\alpha = 1$ anyway and there is no a priori test for this except where $\rho(B)$ is substantially within the unit circle.

(b) $\hat{r}_k$ is very small i.e. either β_k very large, or α_k^* very small but b_k large (so a_k close to one).

(c) $\hat{r}_j$ used in (16) large with α_j^* small, i.e. large negative a_j and/or large b_j.

(d) $\hat{r}_j$ used in (16) quite small, i.e. large positive a_j and large b_j.

(e) If α^* close to β - no a priori test.

Except, possibly for (a) (b), these cases will be rather unlikely but generally large improvements may be expected. Indeed this has been found in actual examples (see table A.1 Chapter 13).

---

[1] the solid curve is the factor improvement given which we know to be an approximation. The left half will actually behave more as the broken curve. It is not clear how the approximation fares on the right.

8.5 Gauss-Seidel and SOR methods

Under some normalisation, the SOR method has $G = (I - \alpha L)^{-1}(\alpha U + (1-\alpha)I)$ where the Gaus-Seidel method is the same with $\alpha = 1$. The immediate question is, do Theorems 6 - 8 and their corollaries also hold for SOR methods?

Lemma 2:

(8) is identical to a recursive acceleration on JOR, i.e. the iteration (6) applied to (7).

Proof:

Obvious. #

Lemma 3:

Equivalence noted in JOR case for $-\alpha$ in place of α also holds here, because if R_α has all $a_j > 1$ then $R_{-\alpha}$ has all $a_j < 1$. #

We note α^* of (16) no longer applies however.

Theorem 9:

(Kahan) For $R_\alpha = (I - \alpha L)^{-1}(\alpha U + (1 - \alpha)I)$ with real or complex A, $\rho(R_\alpha) \geq |\alpha - 1|$ with equality if and only if all eigenvalues of R_α are equal. Hence SOR converges only if $0 < \alpha < 2$.

Proof:

Varga (1962). #

Theorem 10:

Under the iteration method SOR the necessary and sufficient conditions for convergence for an arbitrary starting vector are that the real part of each characteristic root of B shall be less than unity.

Proof:

For (6) $0 = |R - \lambda I| = |(I - L)^{-1} U - \lambda I|$ and

$$\therefore \quad |U + \lambda L - \lambda I| = 0 \tag{20}$$

Therefore a test for $\rho(R) < 1$ is equivalent to a test that the polynomial

$$\begin{vmatrix} \lambda & b_{12} & \cdots & \cdots & b_{12} \\ \lambda b_{21} & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \lambda b_{n,n-1} & \lambda \\ \lambda b_{n1} & \cdots & \cdots & & \end{vmatrix} = 0$$ has all roots within the unit circle.

This is simple since it only involves B so the polynomial is easily calculated.[1]

Now let $\rho(R_\alpha) = \lambda^*$ and $\rho(B) = \mu$.

$$\therefore \quad 0 = \phi(\lambda^*) = \left|(I - \alpha L)^{-1}(\alpha U + (1 - \alpha)I) - \lambda^* I\right|$$

$$\therefore \quad \phi(\lambda^*) = \left|\alpha U + \lambda^* \alpha L - (\lambda^* + \alpha - 1)I\right| \tag{21}$$

Let $\rho(U + \lambda^* L) = \bar{\mu}$, then

$$\therefore \quad \lambda^* = \alpha(\bar{\mu} - 1) + 1 \tag{22}$$

$$\therefore \quad \left|(I - \alpha L)^{-1}(U + (1 - \alpha)L) - \bar{\mu} I\right| = 0 \tag{23}$$

Since by Theorem 4 any eigenvalue is a continuous function of the elements of its matrix, so is $|\bar{\mu}|$ (and so is a where $\bar{\mu} = a + ib$) a continuous function of α for fixed B. Specifically if $\alpha = 1$, $|\bar{\mu}| = |\lambda|$; and if $\alpha = 0$ $|\bar{\mu}| = |\mu|$ and $\bar{\mu} = \mu$. By continuity in (23) as $\alpha \to 0$, $|\bar{\mu}| \to |\mu|$ and therefore $a \to \mathrm{Re}(\mu)$. $\therefore$ For $0 < \alpha < 2$, $\exists\ \alpha > 0$ sufficiently small such that $|\bar{\mu}| = |\mu| + |\delta| \leq |\mu + \delta|$ for some $\delta \neq 0$ arbitrarily small. Since from (22) and Theorem 5 we are only concerned with small α and $a < 1$, the necessary and sufficient condition that there should exist a permitted α sufficiently small for $\rho(R_\alpha) < 1$ is μ should have real part less than unity, which follows if $\bar{\mu} \to \mu-$. By Theorem 9.3, to follow, we see $\bar{\mu}$ is always bounded above by μ if α sufficiently small; thus $\mathrm{Re}(\bar{\mu})$ by $\mathrm{Re}(\mu)$. #

1. The polynomial with the calculation of the dominant root would supply the necessary and sufficient test of convergence of the Gauss-Seidel method. Note: SOR with any specific α may be done this way replacing λ by $\lambda\alpha$. Selection of optimal α by direct calculation would be best done this way.

Corollary 7:

A Meyer-Spasche test on B + I is a test of necessary and sufficient conditions for convergence of an SOR method. #

Thus if JOR converges, SOR does and vice versa , i.e.

Corollary 8:

For given B if $\exists$ some α_1, $0 < \alpha_1 < \beta$, such that JOR converges then $\exists$ α_2, $0 < \alpha_2 < 2$ such that SOR converges. If there is no such α_1 then there is no such α_2. This is a generalisation of the Stein-Rosenberg theorem, but does not cover their relative speeds of convergence, nor any relation between α_1 and α_2 (but see Theorem 13 for an improvement). #

Corollary 9:

SOR, not Gauss-Seidel, is independent of the order of the equations for convergence in any model where B + I satisfies the Meyer-Spasche test. #

Using $\bar{\mu} = \psi_1(\alpha_2,B)$ from (23), with (22) and a lot of manipulation it would be possible to find $\lambda^* = \psi_2(\alpha_2,\mu)$. Similarly we have (17) in the JOR case for α_1. There will therefore be some relationship between max α_1 and max α_2 for convergence. Immediately it is of interest to compare JOR and SOR speeds of convergence. The complexity of $\psi_2(\alpha_2\mu)$ implies suitable values of α_1 and α_2 will generally be quite different, as will their optimal values α_1^* and α_2^*. The interesting analytic comparison is of rates of convergence when α_1^* and α_2^* are in use. Some comparison and comments will be made later.

An alternative proof of Theorem 10 would seek the form of $\bar{\mu} = \psi_1(\alpha,B)$ for use in (22). Then we would choose α such that the function

$$\frac{\lambda^* + \alpha - 1}{\alpha} - \psi_1(\alpha,B) = 0 \qquad (23a)$$

had minimal roots, λ^*. To do this we need the form of $\psi_1(\alpha,B)$ in terms of α, and λ^*. $\psi_1(\alpha,B)$ turns out to be very awkward to handle. We may approximate

its shape in a number of ways. Below we approximate it with a smoothed function and find the minimum of that.

Let $\psi_1(\alpha,B) = \psi_{11}(\alpha,B) + i\,\psi_{12}(\alpha,B)$ be the complex valued function (since ψ_1 is shown below to be a polynomial). By (22), the necessary and sufficient condition for $|\lambda^*| < 1$ is $\psi_{11} < 1$ is valid so long as ψ_{11} is not some polynomial (or approximately so), in negative powers, α^j, $j < -1$. We may write (23) as

$$\phi(\bar{\mu}) = \bar{\mu}^n + (n\bar{\mu})\bar{\mu}^{n-1} + [(n-1)\bar{\mu}^2 + C_{1,1}\bar{\mu} + C_{1,2}]\bar{\mu}^{n-2}$$

$$+ [(n-2)\bar{\mu}^3 + C_{2,1}\bar{\mu}^2 + C_{2,2}\bar{\mu} + C_{2,3}]\bar{\mu}^{n-3} + \ldots.$$

$$+ [\bar{\mu}^{n-1} + C_{n-2,1}\,\bar{\mu}^{n-2} + \ldots. + C_{n-2,n-1}]\bar{\mu} = 0 \qquad (24)$$

because the i^{th} coefficient of the characteristic polynomial is the sum of i^{th} order principle determinants (here of $U + \lambda^*L$). In this case we have coefficients, $n\bar{\mu}$, $\lambda^* \sum_{j=1}^{n-1} b_{j+1j}\, b_{jj+1}$,

$$\sum_{j=1}^{n-2} \{\lambda^{*2} b_{j+1j}\, b_{j+1j+2}\, b_{jj+2} + \lambda^*\bar{\mu}(b_{jj+2}\, b_{j+2j} + b_{jj+1}\, b_{j+1j} + b_{j+2j+1}\, b_{j+1j+2}\}$$

$$+ \lambda^* \sum_{j=1}^{n-2} b_{j+2j}\, b_{j+1j+2}\, b_{jj+1}, \ldots\ldots\ldots \text{ and so on. Alternatively by (23)}$$

we have

$$C_{1,1} = \alpha \sum_{j=1}^{n-1} b_{j+1,j}\, b_{j,j+1}\,; \quad C_{1,2} + (1-\alpha) \sum_{1}^{n-1} b_{j+1,j}\, b_{j,j+1}$$

$$C_{2,1} = \alpha^2 \sum_{1}^{n-2} b_{j+1,j}\, b_{j+1,j+2}\, b_{j,j+2}$$

$$+ \alpha \sum_{1}^{n-2} (b_{j,j+2}\, b_{j+2,j} + b_{j,j+1}\, b_{j+1,j} + b_{j+2,j+1}\, b_{j+1,j+2})$$

$$C_{2,2} = 2\alpha(1-\alpha) \sum_{1}^{n-2} b_{j+1,j}\, b_{j+1,j+2}\, b_{j,j+2}$$

$$+ (1-\alpha) \sum_{1}^{n-2} (b_{j,j+2}\, b_{j+2,j} + b_{j,j+1}\, b_{j+1,j} + b_{j+2,j+1}\, b_{j+1,j+2})$$

$$+ \alpha \sum_{1}^{n-2} b_{j+2,j}\, b_{j+1,j+2}\, b_{j,j+1}$$

$$C_{2,3} = \alpha(\alpha-2) \sum_{1}^{n-2} b_{j+1,j}\, b_{j+1,j+2}\, b_{j,j+2}$$

$$+ (1-\alpha) \sum_{1}^{n-2} b_{j+2,j}\, b_{j+1,j+2}\, b_{j,j+1}$$

and so on, becoming rather complicated to evaluate. Hence (23a) may be written either in terms of λ^* or α.

$$\therefore \quad \phi(\bar{\mu}) = [1 + n + (n-1) + \ldots + 1]\bar{\mu}^{n-1} + [C_{1,2}\bar{\mu} + C_{1,1}]\bar{\mu}^{n-3} + \ldots$$

$$\ldots + [C_{n-2,1}\bar{\mu}^{n-2} + \ldots + C_{n-2,n-1}] = 0 \text{ (and } \bar{\mu} = 0). \qquad (25)$$

$$\therefore \quad \phi(\bar{\mu}) = N\bar{\mu}^{n-1} + K_1\,\bar{\mu}^{n-2} + K_2\bar{\mu}^{n-3} + \ldots + K_{n-1} = 0 \qquad (26)$$

where $N = \frac{1}{2}(n+1)n + 1 \geq 0$; $\quad K_1 = \sum_{j=1}^{n-2} C_{j,1}$; $\qquad K_2 = \sum_{j=1}^{n-2} C_{j,2}$;

$$\ldots\ldots\ldots\ldots ; \quad K_{n-1} = \sum_{j=1}^{n-2} C_{j,n-1} = C_{n-2,n-1}$$

But the typical element of K_i is $C_{j,i} = \sum_{k=1}^{j} \alpha^k\, d_{i,k} = a_j$ where a_j is the sum of j^{th} order pirinciple minor determinants of $A_\alpha = U + \lambda^*L + I$, $j=1 \ldots n-2$ a new index, and $d_{i,k}$ are the coefficients of the polynomial in α, made up of the elements of B times positive and negative multiplying constants.

$$\therefore \quad K_1 = \sum_{j=1}^{n-2} a_j = \sum_{j=1}^{n-2} \sum_{k=1}^{j} \alpha^k\, d_{1,k} \quad \text{or generally } K_i = \sum_{j=1}^{n-1-i} \sum_{k=1}^{j} \alpha^k\, d_{i,k} \qquad (27)$$

because each a_j (j - 1 .... n - 2) contains only one term in λ^{*j}, each of which by expanding (22) has only one term[1] in α^j. Similarly for each j = 1 .... n - 2 th expansion of λ^{*j} yields terms in α^k, (as well as a_j having other terms $\bar{\mu}\lambda^{*j-1}$, $\bar{\mu}^2\,\lambda^{*j-2}$, ... , $\bar{\mu}^{n-1}\,\lambda^*$ giving terms in α^k; k = 1 .... j = 1.) by the form of $\lambda^* = \alpha(\bar{\mu} - 1) + 1$. Hence K_1 .... K_{n-2} are all polynomials of order n - 2 in α. Finally $\bar{\mu} = -\phi^{-1}(N\bar{\mu}^{n-1} + \ldots + K_{n-1})$; that is the $(n-1)^{th}$ root of a polynomial of order n - 2 in α. Therefore for large α, $\bar{\mu}$ is like $\alpha^{\frac{n-2}{n-1}}$ and for small α, $\bar{\mu}$ is like $\alpha^{\frac{-1}{n-1}}$ and Theorem 10 is confirmed. However ϕ^{-1} is a very difficult function to handle analytically and α^* may only be approximated.

Faced with an intractable expression ϕ^{-1} which requires optimising, we try to approximate it with "smoothed" functions whose minima are comparatively more easy to find numerically (analytic results are still unavailable). Three possible methods are presented here. They may be used directly in the hope that the approximations are, in any particular case, not too wild. In any practical example it will not be feasible to assess the quality of the approximation involved for ϕ^{-1}, let alone its turning points. The methods given below are therefore well suited to provide starting points for the search procedures of Chapter 9.6.2.

Using the theory of the roots of polynomials we have from (26)

$$|\bar{\mu}| \geq \frac{(K_1/N)^{\frac{1}{2}}}{\binom{n-1}{2}} \; ; \quad |\bar{\mu}| \geq \frac{(K_2/N)^{1/3}}{\binom{n-1}{3}} \; ; \ldots\ldots ; \quad |\bar{\mu}| \geq \left(\frac{K_{n-1}}{N}\right)^{\frac{1}{n-1}}$$

and so for each i = 1 , ...n j - 2

$$|\lambda^*| \geq \alpha \left[\frac{\left(\frac{K_i}{N}\right)^{\frac{1}{i+1}}}{\binom{n-1}{i+1}} - 1 \right] + 1 \qquad (28)$$

---

1. from the definition of a determinant being the "sum" of all the permutations of products of j elements taken never more than one from any row and column.

Equality only holds in (28) if all $\bar{\mu}_i$ are equal which cannot be generally expected unless all $\bar{\mu}_i = 0$ i.e. $\alpha \to \infty$ which is not allowed. The degree of inequality in (28) is of powers of α higher than $\frac{1}{i+1}$ only. (28) is a better approximation the smaller $\alpha^{\frac{1}{i+1}}$. Thus an α^*, to minimise (28), is preferred from the case of i small if $\alpha^* < 1$, and from i large if $\alpha^* > 1$.

$$\text{Now } \frac{\partial |\lambda^*|}{\partial \alpha} = \left[\frac{\left(\frac{K_i}{N}\right)^{\frac{1}{i+1}}}{\binom{n-1}{i+1}} - 1 \right] + \alpha \left[\frac{\left(\frac{1}{N}\right)^{\frac{1}{i+1}}}{\binom{n-1}{i+1}} \cdot \left(\frac{1}{i+1}\right) K_i^{\frac{1}{i+1}-1} \cdot \frac{\partial K_i}{\partial \alpha} \right]$$

$$\text{and } \frac{\partial K_i}{\partial \alpha} = \sum_{j=1}^{n-1-i} \sum_{k=1}^{i} k \, \alpha^{k-1} d_{i,k}$$

$$\therefore \frac{\partial |\lambda^*|}{\partial \alpha} = \frac{\left(\frac{K_i}{N}\right)^{\frac{1}{i+1}-1}}{\binom{n-1}{i+1}} \left\{ K_i + \frac{\alpha}{N(i+1)} \frac{\partial K_i}{\partial \alpha} \right\} - 1$$

$$\text{but } \alpha \frac{\partial K_i}{\partial \alpha} = \sum_{j=1}^{n-1-i} \sum_{k=1}^{j} k \, \alpha^k d_{i,k}$$

$$\therefore K_i + \frac{\alpha}{N(i+1)} \frac{\partial K_i}{\partial \alpha} = \frac{1}{N(i+1)} \sum_{j=1}^{n-1-i} \sum_{k=1}^{j} (k+1) \, \alpha^k d_{i,k}$$

Therefore our minimisation problem is to find α^* to solve

$$\frac{\left(\frac{K_i}{N}\right)^{\frac{1}{i+1}} - 1}{\binom{n-1}{i+1}} \left\{ \sum_{j=1}^{n-1-i} \sum_{k=1}^{j} \frac{k+1}{N(i+1)} \alpha^k d_{i,k} \right\} - 1 = 0 \qquad (29)$$

all i = 1, ... , n - 2 using also (25) - (27). (29) appears inconvenient but will be quite easy to solve, for each i, by numerical means searching the interval [0,2].[1] That, acceptable,[2] value of α^* which gives the largest value in (38) should then be used. Of course within that interval plural α^*

---

1. We need only search $[0,\beta]$ in fact by Theorem 13(ii) below for diagonally dominant matrices, and only $\left|0, \frac{p}{p-1}\right|$ for p-cyclic matrices - Varga (1962).
2. The solutions of (29) need not lie in the search interval for all i.

solutions, for some i, may result from (29), in which case reference to (28) will indicate the desired value - this method is effectively checking second order conditions. For large n one may assume that at least for some i (28) and (29) are reasonably accurate approximations.

In view of its simplicity and the remarks following (28) when $\alpha^* > 1$, (29) with $i = n - 2$ is (there is little simplification for $i = 1$)

$$\left(\frac{K_{n-2}}{N}\right)^{\frac{2-n}{n-1}} \cdot \frac{2}{N(n-1)}\,\alpha\, d_{n-2} - 1 = 0$$

giving

$$\alpha^* = \frac{N}{d_{n-2}} \left(\frac{n-1}{2}\right)^{n-1} \tag{30}$$

which could be used as a start in solving (29), $i \neq n - 2$.

The following simpler methods may be used to locate a more precise area of search for α^*. Frobenius and Wittmeyer have computed upper bounds of the maximum characteristic root in terms of the elements of a matrix (Bodewig (1956)). By (20) and Frobenius $|\bar{\mu}| \leq \max_i \{\lambda^* \sum_{j<i} |b_{ij}| + \sum_{j>i} |b_{ij}|\}$ where b_{ij} are elements of $B = U + L$ and since $\bar{\mu}$ from B is the same as from B', i refers to both row and column sums. Therefore we have 2n upper bounds. We use lowest maximum of the row, or column sets

$$|\bar{\mu}| \leq d_1 + d_2 \lambda^* = e_1 + e_2 |\bar{\mu}| \qquad \text{by (22)}$$

$$\therefore \quad |\bar{\mu}| \leq \frac{e_1}{1-e_2}\,; \quad e_1 = d_1 - d_2\alpha + d_2,\ e_2 = d_2\alpha$$

$$\therefore \quad |\lambda^*| \quad \frac{\alpha d_1 - d_2\,\alpha^2 + d_2\alpha}{1 - d_2\alpha} - \alpha + 1$$

$$\therefore \quad \frac{d|\lambda^*|}{d\alpha} \simeq \frac{d_1 + 2d_2\alpha + d_2}{1 - d_2\alpha} + \frac{d_2\,(\alpha d_1 - d_2\,\alpha^2 + d_2\alpha)}{(1 - d_2\alpha)^2} - 1 = 0$$

$$\therefore\ \alpha^2(2d_2) + \alpha(5d_2 - 2d_2^2) + (d_1 - 1) \simeq 0 \qquad (31)$$

$$\therefore\ \alpha^* \simeq \frac{d_2\,[2d_2 - 5] \pm (d_2^2(2d_2 - 5)^2 + 8d_2\,(d_1 - 1))^{\frac{1}{2}}}{-4d_2}$$

One of these will represent an approximate minimum; the selection may be carried out by investigating (28) in the event that both $\varepsilon\ [0,2]$. Naturally d_1 and d_2 will vary among the 2^n constraints according to which λ^* is in use. This may be estimated by Wittmeyer's theorem and (29) and (30).

$$|\bar{\mu}| \leq [\lambda^{*2} \sum_{j<i}\sum b_{ij}^2 + \sum_{j<i}\sum b_{ij}^2]^{\frac{1}{2}}\ [\lambda^{*2} f_i + f_2]^{\frac{1}{2}}$$

$$\therefore\ \bar{\mu}^2 \leq \alpha^2\bar{\mu}^{2-2} f_1 + \bar{\mu}\,(1 - \alpha)2\alpha f_i + [\alpha(\alpha - 2) + 1]\, f_1 + f_2;\quad f_1, f_2 > 0$$

$$\therefore\ |\bar{\mu}| \leq \frac{(1 - \alpha)2\alpha f_1 \pm [(1 - \alpha)^2\, 4\alpha^2 f_1^2 + 4(1 - \alpha^2 f_1)\,[(\alpha(\alpha - 2) + 1)f_i + f_2]]^{\frac{1}{2}}}{2(1 - \alpha^2 f_1)} \qquad (32)$$

From (22) $\frac{d|\lambda^*|}{d\alpha} = (|\bar{\mu}| - 1) + \alpha(\frac{d|\bar{\mu}|}{d\alpha}) = 0$, we can obtain (two possible) $\frac{d(\bar{\mu})}{d\alpha}$ from (32) to give a quadratic in α to be solved. Once again the α^* which gives a minimum has to be selected by reference to (28) - though it should be close to that of (31). By Wittmeyer $|\lambda^*| \leq \alpha[(\lambda^{*2} f_1 + f_2)^{\frac{1}{2}} - 1] + 1$

$$\therefore\ \frac{d|\lambda^*|}{d\,\alpha} \simeq [(\lambda^{*2} f_1 + f_2)^2 - 1] + \alpha[\tfrac{1}{2}(\lambda^{*2} f_1 + f_2)^{-\frac{1}{2}}\, 2\lambda^* \frac{d|\lambda^*|}{d}\ ]$$

and setting $\frac{d|\lambda^*|}{d\,\alpha} = 0$

$$\lambda^{*2} f_1 + f_2 \simeq 1 \text{ so at optimum } |\lambda^*| \leq \left(\frac{1 - f_2}{f_1}\right)^{\frac{1}{2}} \qquad (33)$$

$|\lambda^*|$ from (33) may be used to determine d_1 and d_2. We have now several estimates of α^*, (29), (30), (31) and from (32), any of which may be used; or each may be used in a search procedure as described in the next chapter.

Corollary 10:

The upper bound of permitted α is no longer a simple function of α^* as was the case for JOR. #

For example from above

$$|\lambda^*| < 1 \Rightarrow \frac{\left(\frac{K_i}{N}\right)^{\frac{1}{i+1}}}{\binom{n-1}{i+1}} < 1$$

and solving for the upper bound is to solve $(n-2)^{th}$ order polynomial for α. We need only to choose $\alpha^* < 2$. We have an estimate of ρ^* from (33). Others may be obtained similarly from the other approximations. However since they are only approximations and since they are in terms of the elements of B instead of μ_j, as ρ^* for JOR was, it will be very hard to obtain any analytical comparison with the JOR case. As stated earlier, this was the purpose of obtaining expressions for r^* and so these problems will not be pursued.

8.6 Alternative methods for calculating α^*

In defence of the estimates of α^* proposed in the previous section, it is obvious we have traded exactness for comparative computational simplicity. We used the relation $\frac{d|\lambda^*|}{d\alpha} = (|\bar{\mu}| - 1) + \alpha\frac{d|\bar{\mu}|}{d\alpha} = 0$ with various approximations to $\frac{d|\bar{\mu}|}{d\alpha}$ to obtain α^*. If the approximations are inaccurate in some cases, α^* will be badly estimated. But $\frac{d|\bar{\mu}|}{d\alpha} = \frac{d\psi_1}{d\alpha}$ will be a polynomial of order $\frac{n-2}{n-1} - 1$, since $\bar{\mu} = \psi_1(\alpha, B)$ is of order $\frac{n-2}{n-1}$. Thus $\frac{d|\lambda^*|}{d\alpha}$ is a polynomial of order $\frac{n-2}{n-1}$ in α. There will be a large number of turning points of $|\lambda^*|$ with respect to α, for large n. Many of these may be outside the range [0,2], some will be maxima, complex, etc. but in general there can still be a considerable number of local minima in the permitted range depending on B.

To obtain an exact α^* from (27) not only involves solving a difficult polynomial and sorting out which solutions were local minima but also checking each solution for the global minimum. By taking bounds on ψ_1 we approximated the general shape of $\frac{d|\lambda^*|}{d\alpha}$ and hence immediately probably have an approximate global minimum. This would seem worthwhile trade especially if the global minimum is distinct from local ones. The loss of exactness is unlikely to be important.

In other cases a precise value of α^* may be obtained provided A has special properties. For these cases Young and Varga have derived α^*. A is said to be weakly cyclic of order p if and only if it can have its elements permuted such that A becomes block cyclic with p non-null submatrices. Young, p.159-60, has a computer programme to test if a matrix is weakly 2-cyclic. It is then proved that if λ^* is eigenvalue of R_α and μ satisfies

$$(\lambda^* + \alpha - 1)^p = \lambda^{*p-1}\ \alpha^p\ \mu^p \qquad (34)$$

then μ is an eigenvalue of B; and visa versa. Secondly in the case that $|\mu|, |\lambda| < 1$, $|\lambda| = |\mu|^p < 1$ where $\rho(R) = \lambda$ so Jacobi converges p times slower than Gauss-Seidel[1], $p > 1$.

Next

$$(\mu\alpha^*)^p = [p^p(p-1)^{(1-p)}]\ (\alpha^* - 1) \qquad (35)$$

from which unrestricted α^* may be obtained, as unique positive real root less than $\frac{p}{p-1}$. Finally it is shown that $|\lambda^*| = (\alpha^* - 1)(p - 1)$ and $\rho(R_\alpha) < 1$ for $0 < \alpha < \frac{p}{p-1}$ from which we can deduce the rate of convergence of optimal SOR (assymptotically) $R(R_{\alpha^*}) \simeq (\frac{2p^2}{p-1})^{\frac{1}{2}}\ [R(B)]^{\frac{1}{2}}$ which is a great improvement over the Jacobi method. (Theorems 4.3, 4.4 and their corollaries: Varga.) In many econometric models B is sparse and so it is possible that A is weakly p-cyclic or can be reordered and renomalised to be so. In view of the simplicity of these results the correct strategy would be to run a Meyer-Spasche

---

1. which confirms theorem 10 for this A.

test on B for the feasibility of SOR, then use Young's programme to test if A is weakly 2-cyclic. If it is (35) may be used to calculate α^* and SOR converges very much faster than any method mentioned. Young proves further that these results hold approximately if A is a positive definite L-matrix. This could also be tested. Tables are given in Young for the calculation of α^* if A weakly 2-cyclic for μ real and when μ is complex (pp.195-200). If μ is complex Young's calculations involved require knowledge of all the eigenvalues of B and are iterative and need approximately as many operations as optimal SOR itself and so are of no practical help. No help is given if B is not 2-cyclic.

8.7 Ordering and normalisation problems

It is commonly asserted that for any normalisation of A for which $\rho(B) \geq 1$ there is another normalisation such that $\rho(B^*) < 1$ (e.g. Fromm and Klein (1967)). However this proposition is not generally true. A counter example is $A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$. Then clearly if $|a| > \frac{1}{2}$ no renormalisation will reduce the spectral radius below unity since $B = I - A$ implies $\rho(B) = \max\{a, -a, 2a\}$. However there is the general possibility that if $\rho(A) \geq 1$ then another normalisation may have A^* such that $\rho(A^*) < 1$. If so, which particular normalisation will have $\rho(B) = \mu$ such that the spectral radius of any method is at a minimum? Using the same example, Lemma 4: if an ordering implies a divergent Gauss-Seidel process, then there may not be another ordering for which it is convergent.

Proof:

Renormalisation is to permute the columns of A, and set the diagonal elements to 1. Reordering is to permute the rows of A. By (20) $\rho(\lambda L + U) = \lambda$, for some $|\lambda| > 1$, and $\rho[\lambda L + U] = \rho[\lambda L' + U'] = \lambda$ the optimal reordering problem has the same eigenvalues as the optimal renormalisation problem of

the transposed matrix which we know may not be satisfactory. #

Let P,Q be permutation matrices (i.e. every row and column have exactly one nonzero element which is unity) or rank n. Now the similarity transform of A is $S^{-1} A S = \Lambda$ where S is the matrix of eigenvectors and Λ is the Jordan cononical form matrix.

$$\therefore \quad S^{-1} P'P A Q Q'S = \Lambda$$

$$\therefore \quad PAQ = (PS) \Lambda (S^{-1}Q) \qquad (36)$$

The object is to choose P,Q such that the largest diagonal element of Λ in absolute value was minimised. If P = Q we merely permute the diagonal elements. Hence we are only interested in $P \neq Q$. For fixed α (36) also holds for B_α. For Gauss-Seidel however we cannot get an equivalent representation to (36) since $R = (I - L)^{-1} U$ involves the inverse of a difference and we cannot generalise the elements of R* (permuted) in terms of R, where as B*=PIQ+PLQ+PUQ. If α is fixed this also holds for R_α. Hence, if A has eigenvalue $\gamma, \gamma = \mu + 1$, since $|P| = |Q| = 1$, $|A - \gamma I| = 0$ implies $|PAQ - \gamma PQ| = |A^* - \gamma P^*| = 0$. Therefore A* had different eigenvalues if and only if $P^* \neq I$ i.e. $P \neq Q \neq Q' \neq P'$. P* is of course just another normalisation so for JOR there is no need to look for a renormalisation and reordering. The best renormalisation theorem is at least as good.

Theorem 11:

Given A, there may not be $A^* = PAQ$ such that $\rho(B^*) < 1$, $B^* = I - A^*$ when $\rho(B) \geq 1$. If there is such an A* then we consider only all P such that $|\mu|$ from $|B - \mu P| = 0$ gives ρ^* a minimum for JOR methods. Only in this sense do JOR methods not depend on the order of equations. #

Corollary 11:

There are n! possible renormalisations (in general). #

SOR methods are more difficult. To discover how many orderings affect convergence we have to isolate the system from the different normalisations

affecting convergence properties; that is to consider the difference between the Gauss-Seidel equivalent of (3) and (3) itself, for some α, say $\alpha = 1$.

Theorem 12:

the number of orderings which affect convergence properties of SOR is $\frac{n!}{2!}$.

Proof:

Let $D^{(S)} = Y^{(S)G} - Y^{(S)J}$, the difference between the Gauss-Seidel and Jacobi solutions at the S^{th} iteration. The Gauss-Seidel version of (3) is

$$Y^{(S)G} = EY^{(2)} + E'Y^{(1)} - \sum_{k=1}^{S}\left[E\frac{\partial Y}{\partial G}\Big|_{Y^{(k+1)}} + E'\frac{\partial Y}{\partial G}\Big|_{Y^{(k)}}\right] \cdot \left[E\ G(Y^{(k+1)}) + E'\ G(Y^{(k)})\right]$$

where $G(\cdot) = (g_1(\cdot) \ldots , g_n(\cdot))$ and E is n x n strictly lower triangular matrix with unit elements. Now

$$D^{(S)} = E\ D^{(S)} + E'\ D^{(S-1)} + E\frac{\partial G}{\partial Y}\Big|_{D^{(S)}} G(D^{(S)})$$

$$\therefore\quad D^{(S)} = [(I-E)^{-1}E']^{S-1}D^{(1)} + \sum_{k=1}^{S}[(I-E)^{-1}E']^{k}\ (I-E^{-1}E\frac{\partial Y}{\partial G}\Big|_{D^{(k)}}\ G(D^{(k)})$$

The necessary and sufficient condition $\lim_{k\to\infty} E\ \frac{\partial Y}{\partial G}\Big|_{D^{(k)}} G(D^{(k)}) \to 0$ will ensure eventual convergence such that $D^{(S)} \to 0$ as $S \to \infty$. Since E is strictly lower triangular any permutations of rows of $G(\cdot)$ will only not affect this limit if they are confined to first and last rows. Hence the number of orderings which affect convergence is the number of ways of permuting the remaining rows, which is $\frac{n!}{2!}$. #

Corollary 12:

In a block recursive model with k block each with n_j equations, j = 1....k there are $k/2!\ \sum_{j=1}^{k}(n_j!)$ orderings which make a difference. #

By Theorem 10 SOR converges for some $0 < \alpha < 2$ provided Real $(\mu) < 1$ whatever the ordering. However this may be of little practical use since the

α^* may be very small and hence convergence very slow. For small α, $\frac{d|\lambda^*|}{d\alpha} \simeq \alpha^{\frac{1}{n-1}-1}$ from Theorem 10 so $\bar{\mu} \to \mu$ as $\alpha^{\frac{1-n}{n-1}}$ which approaches infinity (more so the larger the model). In difficult cases the necessary α will be very small indeed. It would be useful to have a test for the existence of a reordering which improves convergence.

We need the following theorems of Perron and Stein-Rosenberg

(a) Let A,B be two n x n matrices $0 \leqslant |B| \leqslant A$, then $\rho(B) \leqslant \rho(A)$

(Proof: Theorem 2.8 Varga). #

(b) Let $B \geqslant 0$ then one and only one of the following is valid:-

(i) $\rho(B) = \rho(R) = 0$ (ii) $\rho(B) = \rho(R) = 1$

(iii) $0 < \rho(R) < \rho(B) < 1$ (iv) $1 < \rho(B) < \rho(R)$.

(Proof: Theorem 3.3 Varga). #

We can further generalise the Stein/Rosenberg theorem beyond Corollary 8.

Theorem 13:

If A is irreducibly diagonally dominant[1] then

(i) both JOR and SOR are convergent for some $0 < \alpha < 2$

(ii) both are convergent for the same α, so β is maximum value for α in SOR also.

Proof:

(i) is clear from Theorem 10 as a special case.

(ii) normalisation implies $b_{ij} = a_{ij}/a_{ii}$ $i, j = 1 \ldots n$.

$\therefore \sum_{j=1}^{n} |b_{ij}| \leqslant 1$ $i = 1 \ldots n$ with inequality in at least one i by diagonal dominance. By Frobenius' upper bound on the largest eigenvalue (Section 5)[2] $\rho(|B|) < 1$. By (a) above $\rho(B) \leqslant \rho(|B|) < 1$ so Jacobi converges. But $R=(I-L)^{-1}$

---

1. i.e. A irreducible and $|a_{ii}| \geqslant \sum_{j=1}^{n} |a_{ij}|$ $j \neq i$ and strict inequality at least one i.

2. Either $\rho(A) = \sum_j a_{ij}$ all i or $\min_i \sum_j a_{ij} < \rho(A) < \max_i \sum_j a_{ij}$ for $A \geqslant 0$

(Proof: Lemma 2.5(Varga) is a direct corollary of Gerschgorin's theorems bounding $\rho(A)$

so by (a) above $\rho(R) = \rho[(I - |L|)^{-1} |U|]$. Also $\rho(|B|) < 1$ and $|B| \geq 0$ using (b) above implies $\rho[(I - |L|)^{-1} |U|] \leq \rho(|B|) < 1$ so Gauss-Seidel converges. Since $|B_\alpha| = \alpha|B| + |1 - \alpha|I$ and $|R_\alpha| = (I-\alpha|L|)^{-1}(\alpha|U|+|1-\alpha|I)$ because $\alpha > 0$ we have $\rho(B_\alpha) \leq \rho(|B_\alpha|)$ and $\rho(R_\alpha) \leq \rho(|R_\alpha|)$ and provided $0 < \alpha \leq \beta$, $\alpha \neq 0$, diagonal dominance is preserved because $\alpha\Sigma|b_{ij}|+|1-\alpha| \leq 1$ so $\alpha[\Sigma|b_{ij}|-1] \leq 0$ still holds all $i = 1 \ldots n$, and (ii) is proved. #

It is tempting to try to prove SOR faster than JOR in this case but it cannot be done as we do not know by how much $\rho(B_\alpha) \leq \rho(|B_\alpha|)$ relative to the extent $\rho(R_\alpha) \leq \rho(|R_\alpha|)$ except for $\alpha = 1$.

Corollary 13:

The greater the degree of dominance by the diagonal in the row or column sum with the greatest dominance, the more $\rho(B)$ and $\rho(R)$ are less than unity. This comes from considering the largest inequality $\sum_j |b_{ij}| < 1$ $i=1,\ldots,n$ and the Frobenius upper bound of the spectral radius since they are not equal, all i. #

Corollary 14:

SOR is always superior to JOR if $B \geq 0$ for any α.

Proof:

$$R_\alpha = (I - \alpha L)^{-1}(\alpha U + (1 - \alpha)I).$$

Write $(I-\alpha L)^{-1} = I+\alpha L+\alpha^2L^2 + \ldots\ldots + \alpha^{n-1}L^{n-1}$ since L is strictly lower triangular. This is nonnegative, so therefore is R_α.

$\therefore \quad \exists\ X \geq 0$ such that $R_\alpha X = \lambda X$; $\quad \lambda = \rho(R_\alpha) \geq 0$. Therefore using (20) $(\alpha L + \alpha/\lambda\ U)X = X$, containing two nonnegative irreducible matrices for $\alpha > 0$ and may be rewritten, $\alpha L = \hat{L}$, $\alpha U = \hat{U}$, hence $(\lambda\hat{L} + \hat{U})X = \lambda X$. This is in the notation of the Stein/Rosenberg theorem and the proof is completed identically.#

We have a quick and easy test if there exists a normalisation or an ordering (given α fixed) which enable convergence if $\rho(B) \geq 1$ and a method for finding normalisations or orderings which improve convergence. We need only find P,Q such that irreducible diagonal dominance in A results (then $\rho(B) < 1$, $\rho(R) < 1$, etc.). Among all those possibilities, select those for

which the degree of dominance in the most doninant is largest. Such P and Q may be found by search procedures. However such a test does not handle cases when irreducible diagonal dominance is not possible, nor does it sort out the best P, Q from those "good" ones so obtained. It is not clear what progress may be made by permutations which affect the degree of dominance of other row and column sums.

A further good ordering and normalisation may be obtained using the following theorem[1]:

Theorem 14:

if B is block triangular with diagonal blocks B_{ii} $i = 1 \ldots k$ then $\rho(B_\alpha) = \max_i \rho(B_{\alpha i})$ and $\rho(R_\alpha) = \max_i \rho(R_{\alpha i})$ where $B_{\alpha i}$ and $R_{\alpha i}$ are diagonal blocks defined by B_{ii} $i = 1 \ldots k$.

Proof:

$|B_\alpha - \mu I| = 0$. If B is block triangular so is B_α such that $B_{\alpha i} = \alpha B_{ii} + (1-\alpha)I$. $\therefore \prod_{i=1}^{k} |B_{\alpha i} - \mu I_i| = 0$ and $\rho(B_\alpha) = \max \mu_j$, $j = 1 \ldots n$ so here $\rho(B_\alpha) = \max_i \rho(B_{\alpha i})$. But by (20) $|U + \lambda L - \lambda I| = 0$ and by (21) $|\alpha(U + \lambda L) - (\lambda + \alpha - 1)I| = 0$. So if B is block triangular we have similarly $\prod_{i=1}^{k} |\alpha(U_i + \lambda L_i) - (\lambda + \alpha - 1)I_i| = 0$. From (21) this gives eigenvalues of the diagonal blocks $R_{\alpha i}$. For some fixed α $\rho(R_\alpha) = \max \lambda_j$ $j = 1 \ldots n$ so $\rho(R_\alpha) = \max_i \rho(R_{\alpha i})$ in the block triangular case. #

A rigorous test may now be introduced for the existence of a renormalisation or reordering which will enable convergence when $\rho(B) = \mu$ has real part greater than one, or $\rho(B_\alpha) \gtrsim 1$ for fixed α. We need Markov's determinant theorem (Gantmacher (1959)). If for $S_0 \ldots S_{2m-1}$ we have either

(i) $S_0 = a_0, \ldots, S_{2m-1} = a_{2m-1}$ or

(ii) $S_0 = b_0, \ldots, S_{2m-1} = b_{2m-1}$ where a_i and b_i are the coefficients

of a real polynomial (10) then if also

---

1. Suggested by Agren (1972) for JOR case only.

$$\Delta_1 = S_0,\ \Delta_2 = \begin{vmatrix} S_0 & S_1 \\ S_1 & S_2 \end{vmatrix},\ \ldots\ldots,\ \Delta_m = \begin{vmatrix} S_0 & \ldots\ldots & S_{m-1} \\ \vdots & & \vdots \\ S_{m-1} & \ldots\ldots & S_{2m-1} \end{vmatrix}$$

$$\Delta_1^* = S_1,\ \Delta_2^* = \begin{vmatrix} S_1 & S_2 \\ S_2 & S_3 \end{vmatrix},\ \ldots\ldots,\ \Delta_m^* = \begin{vmatrix} S_1 & \ldots\ldots & S_m \\ \vdots & & \vdots \\ S_m & \ldots\ldots & S_{2m-1} \end{vmatrix}$$

are all positive with $a_0 \geq b_0,\ b_1 \geq a_1,\ \ldots\ ,\ a_{2m-2} \geq b_{2m-2},\ b_{2m-1} \geq a_{2m-1}$ then $\Delta_1 \ldots\ \Delta_m,\ \Delta_1^* \ldots\ \Delta_m^*$ remain positive for all possible values of S_i such that $a_0 \geq s_0 \geq b_0,\ \ldots\ ,\ b_{2m-1} \geq S_{2m-1} \geq a_{2m-1}$ and $A_k \geq \Delta_k \geq B_k$, $A_k^* \geq \Delta_k^* \geq B_k^*$ in the obvious notation, $k = 1 \ldots\ m$.

Evidently if the conditions for this theorem to hold are satisfied for some characteristic equation coefficients then a matrix yielding that equation is stable by Routh-Hurwicz (as in Gantmacher). Let the basic matrix A be stable and have characteristic equation coefficients b_i. This theorem indicates if we can permute rows and columns of A such that we can find A* with characteristic equation coefficients a_i satisfying the conditions of this theorem with respect to b_i, A* is also stable. If however A was not stable, A* may or may not be stable; but it is "closer" to stability by the last inequalities. In particular we can find those permutations which make A_k and A_k^* as big as possible relative to B_k B_k^* each k: note S_i of any permutation is the sum of the principle minor determinants of order i in that permutation. Hence we want $S_0 > 0$ as large as possible (it will be unity for characteristic polynomial) $S_1 > 0$ but as small as possible i.e. select from the elements so that the product of the diagonal elements is the smallest positive number available. Let the numbers actually chosen so far be a_0 and a_1. Next test if an arrangement can be found for first upper and lower diagonal elements such that $S_2 > \frac{a_1^2}{a_0}$. If so is there an arrangement of the next diagonals such that $S_3 > \frac{S_2^2}{S_1} = \frac{a_1^3}{a_0}$? A

sequential test using Δ_k then Δ_k^*, k = 1 .... m may be performed. Further if at each stage of all the S_i which pass the i^{th} test we choose the permutation which alternatively minimises or maximises the inequality we will finish with the permutation which minimises the spectral radius of the set of matrices with the given elements a_{ij}. This is because S_i, Δ_i and the spectral radius are all conintuous functions of the elements of A, A*, etc. we will get the series A_k, A_k^* as large as possible. Therefore this procedure will improve convergence properties most for given α by continuity and the final inequality of the theorem.

For the large system this rigorous test involves a very large amount of computation and trials.

Theorem 15:

A test of necessary and sufficient conditions for an improvement in convergence properties of any linear stationary method to be possible by reordering equations and renormalising, is to reorganise the elements on successive diagonals to convert characteristic equation coefficients b_i to a_i by at least the amounts and in directions indicated above such that the Markov determinant theorem still holds. If this is done optimally it is equivalent to finding P,Q to select the matrix with smallest spectral radius of the set of matrices $\{a_{ij}\}$. #

Corollary 15:

A lexographic strategy may also be used for SOR as was proposed for JOR in Section 5, by (27). However it cannot be said that necessarily optimal SOR follows from B* optimal for JOR. It may be the case but returning to (26) for fixed α the reorganising of b_{ij} in the coefficients of $\bar{\mu}$ to obtain K_i i = 1 .... n - 2 prevents any general statement being made that B* also is optimal for SOR even though $|\lambda^*| = \alpha(|\bar{\mu}| - 1) + 1$. #

Hence the correct strategy for SOR is to go through the procedures for optimising the ordering and normalisation. This procedure should be performed on (26) for some fixed α (for convience α = 1). Since

$$|\lambda^*| = \alpha(|\bar{\mu}| - 1) + 1 \qquad \alpha > 0$$

if we get a minimum $|\bar{\mu}|$ by the same rules as the previous paragraph, we will have $|\lambda^*|$ minimised.

8.8 Conclusion

This chapter has been confined to linear stationary first order methods for solving (5). The results are (i) a very simple test for convergence valid for all methods considered, (ii) optimisation in minimising the number of iterations to convergence via a damping factor and a demonstration of the role of this factor, (iii) optimisation in the same sense by reordering and renormalisation. All these theorems are operational.

Very little has been obtained in comparison theorems. However there is a strong supposition that a good SOR is very likely to be at least as good as optimal JOR - though simplicity, accuracy and minimisation of over-all computations must not be overlooked. In any case (33) by direct calculation may be compared with (17) in specific cases and the method with smaller spectral radius (approximately) chosen unless they are close, when SOR might be preferred.

CHAPTER 9: EXTENSIONS OF SOR THEORY

9.1 Introduction

A number of results useful in applications of the SOR method have been developed in the previous chapter. We have necessary and sufficient conditions, and hence an exact a priori test for SOR convergence in the general case. We can calculate an approximate α^* and the optimal reordering or renormalisation. All these theorems are operational. Nevertheless they are computationally involved, and very likely it is not worthwhile to undertake them completely as then the total work done in solving (8.5) may exceed that for nonoptimal solutions. At this stage it is conjectured the minimum total work, optimising and iterating, will come from the simpler JOR method. Hence a plan for good orderings and normalisation was given. There remain the problems (i) how should (8.5) be solved if $a > 1$? (ii) If SOR is convergent but only for very small α, convergence will be very slow since $R_\alpha \to I$ with $\alpha \to 0$. Can this be improved? It is now assumed that a "good" if not optimal normalisation and ordering has been chosen. (iii) Can we test the conjecture that some JOR are superior to a particular sensible SOR? (iv) Is there a good α to use for α^* in SOR which is easy to calculate and perhaps more exact than that of Chapter 8.5? (v) Can we test if B is p-cyclic? Problems (iii) - (v) are handled now; a partial answer is provided for (iii) but it is taken up again the the next chapter, where also (i) is answered.

First tests of the superiority of SOR methods over JOR are proposed; then a rule for quickly calculating a good α, and also an improved ordering. These results taken with the previous SOR ones constitute simple working rules for the SOR method in the general case. It is emphasised that the following are not optimal results, indeed most of the tests are

only sufficient conditions, so the scheme may fall some way short of optimal SOR performance. It is possible the simplifications involved might show an apparent superiority of JOR, when in fact this was not necessarily the case. The object is to provide the first comparison theory for SOR and, secondarily, translate the theory into practical tests which are computationally sufficiently simple.

9.2 Comaprison theorems:

Very little has been obtained to date for SOR comparison theorems. Known theorems state SOR is superior to JOR if

(a) B is p-cyclic[1] (Varga (1962) Theorem 4.3), for some α values;

or (b) $B \geq 0$ and irreducible[2] (Theorem 8.9 corollary 2 above), all α;

or (c) A is a Stieljes or M-matrix (Varga (1962) Theorem 3.15 corollary 2);

or (d) $A^{-1} > 0$ (Varga (1962));

or (e) A positive definite L-matrix (Young (1971)).

We note a M-matrix, A, is defined to be real with $a_{ij} \leq 0 \quad i \neq j$ for A nonsingular and $A^{-1} \geq 0$. The normalisation may be taken as $a_{ii}=1$ $i=1....n$. A Stieljes matrix A is defined to be real with $a_{ij} \leq 0$ $i \neq j$ for A symmetric and positive definite. Therefore we conclude:-

Theorem 1

(i) $A^{-1} > 0 \Rightarrow a_{ii} > 0$, $a_{ij} \leq 0$ $j \neq i = 1 .... n$ iff A is M-matrix.

Proof:

Identical to theorem 3.10 Varga but using this normalisation on A. #

(ii) If A is a Stieljes matrix then it is an M-matrix but not conversely.

Proof:

Theorem 3.11 corollary 3 Varga. #

---

1. for p = 2 and nonnegative eigenvalues - the generalisation for all p and negative eigenvalues is in Section 5.1 below.
2. Recall we assume irreducibility throughout.

(iii) A a positive definite L-matrix ⇒ A is an M-matrix.

Proof:

L-matrices X,Y with eigenvalues λ,μ, are defined to have the property that the pencil $cX + dY$ (c,d scalars) has eigenvalues $c\lambda + d\mu$. M-matrices in the polynomial $p(X,Y)$ have eigenvalues monotonically related $p(\lambda,\mu)$ (see Taussky (1964)). Positive definiteness ensured these linear relations to be monotonic and increasing. #

(iv) Under this normalisation, A and M-matrix ⇒ $B \geq 0$.

Proof:

$B = I - A$ so the result is immediate from the definition. #

Corollary 1

SOR is superior to JOR if, either (a) holds, for some α or if (b) holds for any α. This is a special comparison theorem.

Proof:

If (b) holds by Theorem 9 corollary 2 for any α. If (a) holds by Varga Theorem 4.5 corollary 1, $R(R_{\alpha^*}) \simeq \left(\frac{2p^2}{p-1}\right)^{\frac{1}{2}} \left(R(B)\right)^{\frac{1}{2}}$ where $R(\cdot)$ = asymptotic convergence rate. #

This corollary does not say any SOR is superior to any JOR, but that optimal SOR is superior to optimal JOR, and that SOR is superior JOR with the same α if $B \geq 0$; and if B is p-cyclic at least some SOR is superior to some JOR. Using the continuity of eigenvalues in the elements of the matrices we can extend these areas of superiority to some extent.

We can test if B satisfies either (a) or (b) immediately and use SOR if either are satisfied. It is not possible to proceed further with exact tests, without calculating the two spectral radii in a given case for some α, since there is no available relation between $\rho(R_\alpha)$ and $\rho(B_\alpha)$ unless (a) is strictly true. We can deduce more general comparison theorems:-

Theorem 2A:

Let V be an n x n matrix and $\eta_1 > 0$ such that $R_\alpha = C + V$ where $C = (I - \alpha L)^{-1}(\alpha U + (1 - \alpha)I)$, $B_\alpha = \alpha B^+ + \alpha D$, $B^+ = L + U$, $D = L_D + U_D$. Then SOR is superior to JOR for some $0 < \alpha < 2$ when $B^+ \gtrsim 0$, if $V \in S$, some neigh-bourhood about C with radius η_1; i.e. $S \in R^n$ with $||S - C||_r = \eta_1, r \in [1, \infty]$. In particular optimal SOR is superior to optimal JOR if $V \in S$ for α^* of JOR.

Proof:

Let $\rho(B_\alpha) = \mu'$, $\rho(B^+_\alpha) = \mu$, $\rho(C) = \lambda$, $\rho(R_\alpha) = \lambda'$. For any $0<\alpha<2$, $\lambda<\mu$. Let Y^* be solution to (8. 5) and $e^{(m)} = Y^{(m)} - Y^*$

$\therefore\ e^{(m)} = \prod_{i=0}^{M} (C + V)\, e^{(0)} = R_\alpha^m\, e^{(0)}$. If $V \in S$, $||V||_r \lesssim \eta_1$. By Theorem 1, Chapter 16, Ostrowski (1960),

$$||R_\alpha^m||_r \lesssim \sigma(\lambda + \varepsilon)^m \qquad (1)$$

if we choose ε such that $\lambda + \varepsilon < \mu'$ and define (as Ostrowski) $\sigma = ||X||_r\ ||X^{-1}||_r$ where X is the eigenvector matrix of C, and $\eta_1 = \varepsilon/2\sigma$.

$\therefore\ \sigma > 0$ and finite for some normalisation on the eigenvectors.

$\therefore\ ||e^{(m)}||_r \lesssim ||R_\alpha^m||_r\ ||e^{(0)}||_r \lesssim \sigma(\lambda + \varepsilon)^m\ ||_e(0)||_r$

From (1) $||R_\alpha^m||_r < \sigma\mu'^m$ and since $\rho(R_\alpha) = \lim_{m\to\infty} ||R_\alpha^m||_r^{1/m}$, SOR is superior in asymptotic rates of convergence for any norm, because $\frac{1}{m}\log||R_\alpha^m||_r < \frac{\log\sigma}{m} + \log\mu'$. Thus the asymptotic rate

$$\log\rho(R_\alpha) = \lim_{m\to\infty}\left(\frac{1}{m}\log||R_\alpha^m||_r\right) < \lim_{m\to\infty}\left(\frac{\log\sigma}{m} + \log\mu'\right) = \log\mu'. \qquad (2)$$

$\therefore\ \rho(R_\alpha) < \rho(B_\alpha)$ by monotonicity of logarithms. If $||V||_r \lesssim \eta_1$ with α^*, this inequality also holds for α^* of optimal JOR. Optimal SOR has spectral radius at least as small as $\rho(R_{\alpha^*})$ for this α^* and probably smaller.

Naturally theorem can only be valid if $\varepsilon > 0$

i.e. $\lambda < \mu'$ (3)

But from theorem 1 we know $\lambda < \mu$. By the continuity of eigenvalues (Ostrowski (1960)) if $\max_{i,j} |d_{ij}| = \gamma$ then

$$|\mu' - \mu| < (n + 2)\,(n^2\gamma)^{1/n} \tag{4}$$

so for γ sufficiently small (3) will always be satisfied and $V \in S$. #
The existence of $\varepsilon > 0$ S is certain for γ sufficiently small, but the allowable bound, $||V||_r \leq \eta_1$, becomes rather difficult to use in a test. It is really in two parts; i.e. $||V||_r$ is such that $|\lambda' - \lambda| + |\mu' - \mu| < |\lambda - \mu|$, and then η_1 is implicit.

Theorem 2B:

With the notation of theorem 2A, when B^+ is p-cyclic and α^* given for SOR on B^+ by (8.35), $\rho(R_{\alpha}{}^*) < \rho(B_\alpha)$ for a particular α if $V \in S$ some neighbourhood about C with some radius η_2. Optimal SOR is then superior to JOR.

Proof:

Identical to that of Theorem 2A but with $\varepsilon > 0$ defined as $(\alpha^* - 1)\,(p - 1) + \varepsilon = \mu'$. But $(\alpha^* - 1)\,(p - 1) < \mu$ by corollary 1, for α sufficiently close to 1. α^* is obtained from μ in (8.35); and $\eta_2 = \varepsilon/2\sigma$, σ from the eigenvectors of this C matrix. The spectral radius of optimal SOR is at least as small as $\rho(R_{\alpha^*})$; α^* is optimal only for B^+. #
We can restate theorem 2B in various versions for more general α values:

Corollary 2:

R_{α^*} for SOR is superior to optimal JOR if $\rho^* > (\alpha^* - 1)(p - 1)$ for ρ^* of (8.19) and $V \in S$ with that α^*. Therefore optimal SOR is preferable to optimal JOR in this case.

Corollary 3:

If for some α_o Theorem 2B is satisfied then SOR with α_o within the interval $[\alpha_1 \leq \alpha_o \leq \alpha_2]$ is superior to some particular JOR. α_1, α_2, are determined as those values for which $||V|| = \eta_2$; $\eta_2 = \varepsilon_1/2\sigma$; $\lambda + \varepsilon_1 = \mu'$.

If λ' is a monotonically increasing function of $\alpha(\mu'$ fixed) as $\alpha \to 0$, or $\alpha \to 2$, then α_1, α_2 are unique and α^{**} (the optimal parameter for B), α^* are also in the interval. This property holds for λ given any α, and so by (4) for V with small enough elements. It seems highly likely that if B is close enough to B^+ as in Theorem 2B then λ', given μ', will display the same property. But it cannot be proved for all $||V|| \leq \eta_2$, since we are only bounding $|\lambda' - \lambda|$ and have no relation between λ' and α. If the bound widens, since λ' is a polynomial of order $\frac{n-2}{n-1} + 1$ with respect to α by (8.26), this could imply nonunique α_1 α_2 despite the satisfaction of Theorem 2B. #

Corollary 4:

If Theorem 2B is satisfied for a particular SOR against JOR with α_o, then there is an interval about α_o for which all JOR methods are inferior. We could attempt to determine that intervals as above. #

Corollary 5:

In summary, if V ε S of Theorem 2B $\exists$ at least α_o for which any SOR using $\alpha \varepsilon [\hat{\alpha}_1 \leq \alpha_o \leq \hat{\alpha}_2]$ where $\hat{\alpha}_1 \geq \alpha_1$ and $\hat{\alpha}_2 \leq \alpha_2$ is superior to any JOR with $\alpha \varepsilon[\alpha_3 \leq 1 \leq \alpha_4]$. #

The proof of these versions is immediate from that of Theorem 2B involving merely different forms of the bound η_2. However the determination of the interval boundaries will be very burdensome in any particular case. In view of the extreme simplicity of Theorem 8.8, the sensible strategy is to apply corollary 2. Then, if desired, a better α may be sought for SOR if it proves superior.

Any version of Theorems 2A or B is inconvenient for practical testing. Their use is theoretical in proving that SOR will be superior in asymptotic rates of convergence if B is nearly p-cyclic or non-negative, and in indicating a form of test. Note: (i) the average rate may not be superior even if the asymptotic one is. Since $\mu' \leq ||B||_r$ from (2).

$$\frac{1}{m} \log ||R_\alpha^m||_r < \frac{\log \sigma}{m} + \frac{1}{m} \log ||B_\alpha^m||_r$$

so the average rate is superior only for all $m > m_o$; m_o given by $(||R_\alpha^{m_o}||_r / ||B_\alpha^{m_o}||_r)^{1/m_o} = \sigma$. (ii) Even if $V \notin S$, SOR may be superior on asymptotic rates as $\rho(R_\alpha) \lesssim ||R_\alpha||_r$ so $||V|| > \eta$ may hold with $(\rho(R_\alpha))^m \lesssim \sigma(\lambda + \varepsilon)^m \lesssim ||R_\alpha^m||_r$. The test indicated is a <u>sufficient</u> condition; the test will be achieved by bounding eigenvalues.

9.3 Tests for the superiority of SOR

The test indicated by Theorems 2A, 2B is in terms of $\eta = \varepsilon/2\sigma$. ε is known from the two dominant eigenvalues μ' and λ and then $||V||_r < \eta$ may be tested for some norm; $r \in [1,\infty]$. Of the norms commonly used $||V||_1 = \max_j \sum_i |V_{ij}|$, $||V||_\infty = \max_i \sum_j |V_{ij}|$ and $||V||_2 = [tr(V'V)]^{\frac{1}{2}}$ are simplest to calculate. Hereafter $||V||_r$ will mean the minimum of those three.

If X is to be calculated for σ, involving considerable computation, we will obtain λ'_i, given α, $i = 1 \ldots n$ anyway; and if we have tested $\rho(B)$ for convergence we could compare λ' with μ' directly. To save computation we sum the tightest bounds for $|\lambda' - \lambda|$ and $|\mu' - \mu|$ available in terms of V, D elements and test if the sum exceeds $\mu - \lambda$. This test requires a knowledge of one of λ,λ' and one of μ,μ'. In the $B^+ \gtrsim 0$ case, we know if SOR is superior for one α, it is for all values where $V \in S$. We calculate only the two dominant eigenvalues and if SOR is superior find a good α; if not find α^* for JOR. (At least SOR with α^* for JOR is faster than optimal JOR if the test is satisfied.) In the B^+ p-cyclic case, we know (8.34) gives λ from μ given α. Therefore obtain α^* from (8.35) using μ of B^+. Then set $\lambda = (\alpha^* - 1)(p - 1)$. μ' is calculated for some α e.g. if the alternative is optimal JOR, ρ^* from (8.19). In either case the bounds will be in the form of ovals about one spectral radius, say λ. It is easier to perform the test by establishing if a certain point, say μ' or ρ^*, lies within the ovals, rather than determine ε and test if it exceeds the sum of the two bounds.

In either case, computation may be further reduced if μ' is not calculated; the Meyer-Spasche test or a knowledge of the eigenvalues of B only is required. Then if $\alpha < \beta$ in Theorem 8.6 convergence is assured. One problem is, for SOR unless Theorem 8.13(ii) applies, we cannot determine a priori the upper limit of permitted α values. Then we must test for a point of intersection of the ovals of λ and μ, directly.

9.4 Bounds for testing:

D may be expected to have very few nonzero elements, perhaps none. Typically B itself has been sparse in large econometric models. It is therefore very likely B is itself p-cyclic but if not that D has very few nonzero elements. For example the Klein-Goldberger, Dutta and Su, Michigan and Wharton econometric models had only 8.7, 5, 5.7, 2.6% of B positions filled respectively. Since

$$R = [I + L + L_D + (L + L_D)^2 + ... + (L + L_D)^{n-1}] [U + U_D] \qquad (5)$$

$$\therefore \quad V = [L_D + ... + \sum_{j=1}^{n-1} \binom{n-1}{j} L^{n-1-j} L_D^j] U + [I + (L+L_D) + ... + (L+L_D)^{n-1}] U_D \qquad (6)$$

There is no analytical or computational advantage to be gained from the extreme sparseness of D. But since V is made up of the polynomial $p(L,L_D)$ every term is at least either L_D or U_D times a matrix factor. V tends to be full but if all the elements of D were of the same order $V \to 0$ as D gets emptier since more terms become annihilated within the summations. Now redefine V such that $R + C + \delta V$, $\delta > 0$ chosen such that $|V_{ij}| \leq 1$ $i,j=1....n$; C,X etc. as before. Let $\rho(R) = \lambda'$. Then by (4) $|\lambda' - \lambda| < (n+2)(n^2\delta)^{1/n}$ which provides a secondary proof of the existence of S.

Let $X(C+\delta V)X=(\text{diag } \lambda_i) + \begin{vmatrix} \beta_{11}/S_1 & \cdots & \beta_{1n}/S_1 \\ \vdots & & \\ \beta_{n1}/S_n & \cdots & \beta_{nn}/S_n \end{vmatrix}$ (7)

where $S_i = x^i x_i$; x^i is i^{th} row of X^{-1} and x_i is i^{th} column of X both normalised so that $||x_i||_2 = 1$ $\therefore$ $\beta_{ij} = x^i V x_j$.

$$\therefore \quad |S_i| = |x^i x_i| \leq ||x^i||_2 \, ||x_i||_2 = 1. \tag{8}$$

$$\therefore \quad |\beta_{ij}| = |x^i(Vx_j)| \leq ||x^i||_2 \, ||Vx_j||_2 \leq ||V||_2 \, ||x^i||_2 \, ||x_j||_2 = ||V||_2. \tag{9}$$

Adapting an idea of Wilkinson (1965) we may bound $\rho(V)$ by Gerschgorin's theorems for use as η for (3). λ'_i lies in a disc, centre $(\lambda_i + \delta\beta_{ii}/S_i)$ and radius $\frac{\delta}{|S_i|} \sum_{j\neq i} |\beta_{ij}|$. For $\rho(R)$ consider i = 1, and if λ_1 is simple λ'_1 remains simple since a small enough δ keeps the first disc isolated.

However mutliplying first column by a constant z and first row by 1/z where $z = k/\delta$ leaves the centre unaltered. k is a constant chosen so the first disc is as small as possible but no other disc overlaps the first to such an extent that the spectral radius increases,

i.e. $|k \, \beta_{i1}/S_i| \leq |\lambda_1 - \lambda_i|$ all i = 2 .... n

$$\text{Now } X^{-1}(C+\delta V)X=(\text{diag } \lambda_i)+\begin{vmatrix} \delta\,\beta_{11}/S_1 & \delta^2\beta_{12}/kS_1 & \cdots & \delta^2\beta_{1n}/kS_1 \\ k\beta_{21}/S_2 & & & \\ \vdots & & & \\ k\,\beta_{n1}/S_n & \delta\beta_{n2}/S_n & \cdots & \delta\beta_{nn}/S_n \end{vmatrix} \tag{10}$$

The first disc has radius = $\sum_{j=2}^{n} |\delta^2\beta_{ij}/k\, S_1| \leq \dfrac{\delta^2||V||_2(n-1)}{k\,|S_1|}$ which is an

order of magnitude improvement for small δ. And $k = \min_i |(\lambda_1 - \lambda_i)S_i/\beta_{i1}|$ so $k^{-1} = \max_i |\frac{\beta_{i1}}{(\lambda_1-\lambda_i)S_i}| \leq \max_i |\frac{||V||_2}{(\lambda_1-\lambda_i)S_i}|$. We could bound $|\lambda'_1 - \lambda_1|$ by this method. We may tighten the bound further by using Brauer's (1964) Cassini ovals theorem in place of Gerschgorin's circles. Repeat the operations with z and to obtain (10) on the second rows and columns.

$$X^{-1}(C+\delta V)X+(\text{diag } \lambda_i)+ \begin{vmatrix} \delta\beta_{11}/S_1 & \delta\beta_{12}/S_1 & \delta^2\beta_{13}/kS_1 & \cdots\cdots\cdots\cdots \\ \delta\beta_{21}/S_2 & \delta\beta_{22}/S_2 & \delta^2\beta_{23}/kS_2 & \cdots\cdots\cdots\cdots \\ k\beta_{31}/S_3 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \\ & & \ddots & \delta\beta_{nn}/S_n \end{vmatrix} \qquad (11)$$

then $|\lambda_1'-\delta\beta_{11}/S_1-\lambda_1||\lambda'-\delta\beta_{22}/S_2-\lambda_2| \leq \left|\left(\frac{\delta^2}{kS_1}\sum_{j=3}^{n}\beta_{1j}+\frac{\delta}{S_1}\beta_{12}\right)\left(\frac{\delta^2}{kS_2}\sum_{j=3}^{}\beta_{2j}+\frac{\delta}{S_2}\beta_{21}\right)\right| = g$

(12)

$$\therefore \quad g = \left|\frac{\delta^4}{k^2S_1S_2}(\Sigma\beta_{1j}\,\Sigma\beta_{2j}) + \frac{\delta^3}{kS_1S_2}(\beta_{12}\,\Sigma\beta_{2j} + \beta_{12}\,\Sigma\beta_{1j}) + \frac{\delta^2}{S_1S_2}\beta_{12}\,\beta_{21}\right|$$

$$\leq \frac{||V||_2^{\,2}\delta^2}{|S_1S_2|}\left[\frac{\delta(n-2)}{k} + 1\right]^2 \simeq \frac{||V||_2^{\,2}\,\delta^2}{|S_1\,S_2|} \quad \text{for small } \delta/k. \qquad (13)$$

Now to test if a particular point, (in this case[1] μ^* or μ'), a + ib lies in the ovals (12) let $2d = |\lambda_1 + \delta\beta_{11}/S_1 - \lambda_2 - \delta\beta_{22}/S_2| \simeq |\lambda_1 - \lambda_2|$, SOR is preferable if

$$|b| > \{-a^2 - d^2 + (g^2 + 4d^2\,a^2)^{\frac{1}{2}}\}^{\frac{1}{2}} \qquad (14)$$

---

1. $\mu^* = \alpha^*\,\mu' + (1-\alpha^*)=(\alpha^*-1)a + 1 + i\alpha^*b$ for α^* from (8.23)

Using Theorem 35 of Brauer (1952) we can tighten the bound yet further.

Replace g of (12) by

$$g = \delta|\beta_{12}/S_1|\left(\frac{\delta^2}{kS_1}\sum_{j=3}^{n}\beta_{ij} + \frac{\delta\beta_{12}}{S_1}\right) + \delta|\beta_{12}/S_2|\left[\frac{\delta^2}{kS_2}\sum_{j=3}^{n}\beta_{2j} + \frac{\delta\beta_{21}}{S_2} - \delta|\beta_{12}/S_1|\right] + \delta^2\left|\frac{\beta_{12}\beta_{21}}{S_1 S_2}\right|$$

$$+ \sum_{\substack{t<q \\ t,q>2}}\left|\frac{\delta^4}{k^2}\left(\frac{\beta_{1t}}{S_1}\frac{\beta_{2q}}{S_2} + \frac{\beta_{1q}}{S_1}\frac{\beta_{2t}}{S_2}\right)\right| + \sum_{\substack{t<q \\ t,q\leq 2}}\left|\frac{\delta^2}{S_1 S_2}\left(\beta_{1t}\beta_{2q} + \beta_{1q}\beta_{2t}\right)\right| \qquad (15)$$

Note for (10), (13) or (15), $||V||_2$ enters the bound after considering either $\frac{\Sigma\beta_{1j}}{|S_1|}$ or $\frac{\Sigma\beta_{2j}}{|S_2|}$ so by (9) we may write $||V||_2$ wherever $|\beta_{ij}|$ occurs.

Although (15) looks involved it is very easy to compute using all $|\beta_{ij}| \leq ||V||_2$. Brauer proves the bound (12) is contained within (10) and g of (15) less than that of (13). (15) is still $O(\delta^2)$ but it should be used in (14). One can further improve the bound if the matrices of (11) are real - but this we cannot guarantee even when R is real. The disadvantage of the Gerschgorin procedure is that it requires the calculation of all S_i, and that $(\lambda_1 - \lambda_i)$ is likely to be very small and so k^{-1} very large. And $|S_i|$ could be arbitrarily small. Not only have we improved the bounds in (12), (14) but the bound has a difference in powers of δ and k of +2 in place of +1 in all terms. The requirement to calculate beyond S_1 and S_2, λ_1 and λ_2 is avoided. The bounds are a factor of (n-1) smaller. We can therefore except a substantial improvement in (14) over simpler bounds when used in (11). If the approximation (13) is used for g the computation of k is avoided.

Since $|V_{ij}| \leq 1$, $||V||_2 \leq n$, $r \in [1,\infty]$. But

Lemma 1:

$||V||_2 \leq \sqrt{n}/\delta$ if $\lambda < 1$

Proof:

Theorem 65 of Westlake (1968) states $||V||_2 \leq \sqrt{n}\,\rho(V)$.
$R^m = (C + \delta V)^m = \sum_{j=0}^{m} \binom{m}{j} C^{m-j}\ \delta^j V^j$, so if $\lambda < 1$ then $\lim_{m\to\infty} R^m = 0$ implying
$\lim_{m\to\infty} \delta^m V^m = 0$: i.e. $\delta\rho(V) < 1$. #

Naturally Lemma 1 is helpful when $\sqrt{n} < 1/\delta$ which follows cases where B is nearly p-cyclic or nonnegative. Certainly if $\lambda \gtrsim 1$ Lemma 1 is not true but since R is not convergent in this case no bounds on $|\lambda' - \lambda|$ will indicate convergence and so the α being used is unsatisfactory. This possibility will never be considered in comparison tests. For very quick estimates of the foregoing bounds $\sqrt{n}/\delta$ may be substituted (or n if $\sqrt{n} \gtrsim 1/\delta$) for $||V||_r$. For that reason I have quoted $||V||_2$, since (9), being likely to have the smallest bound.

If λ_1 is not simple, or C has nonlinear elementary divisors for λ_1, the bounds are not as tight as quoted here. Wilkinson provides Gerschgorin bounds for such cases, and similar amendments to those above are easily constructed using a parallel development. They are therefore not considered separately.

Bounds for $|\mu' - \mu|$ may be obtained in exactly the same fashion. Writing $B = B^+ + \gamma D$ where γ is chosen such that $|d_{ij}| < 1$, we may take advantage of the extreme sparseness of D - it contains only the "mistakes" of an already sparse matrix B - though γ may well not be so small as δ. $|S_i|$ now refers to the eigenvectors of B^+ and $|\beta_{ij}| \lesssim ||D||_r$ and k, if used, implies $k^{-1} \lesssim \max_i \left|\frac{||D||_r}{(\mu_i - \mu_i)S_i}\right|$. But now $\min\{||D||_1\ ||D||_\infty\} \lesssim 1$, or the lesser of the largest number of nonzero elements in any row or column if there are more than one in any row or column. Any of the norms are very simple to compute and compare. We substitute γ for δ, $||D||_r$ for $||V||_2$, μ_i for λ_i etc. in (10), (13), (15). This will not be necessary if μ' is known; if not it is simplest to test for the intersection of the two bounds $|\lambda' - \lambda|$ and $|\mu' - \mu|$.

Finally we may consider explicitly representing $\alpha \neq 1$ in the bounds. Inserting α in (6).

$$V_\alpha = [\alpha L + ... + \alpha^{n-1} \sum_{j=1}^{n-1} L^{n-1-j} L_D^j] \alpha U + [I + \alpha L + L_D + ... + \alpha(L+L_D)^{n-1}]\alpha U_D.$$

Hence for small α, $V_\alpha \simeq \alpha U_D$. We conclude for α sufficiently small.

Theorem 3:

$\exists\ \alpha_o$ such that for $0 < \alpha < \alpha_o$ SOR is superior to JOR, any B.[1]

Proof:

Regard any B as "nearly" non-negative, for sufficiently small α. $V_\alpha \simeq \alpha U_D \Rightarrow$ we can replace δV with $\delta \alpha U_D$. $\alpha \leqslant \alpha_o$ can be chosen where $\delta\alpha_o$ satisfies Theorem 2A. Additionally if $U_D = 0$ α_o increases by an order of magnitude. #

Theorem 3 suggests SOR is always preferred for cases of very slow convergence. For any $\alpha < 1$ we may replace δ by $\delta\alpha$, since $||V||_r \leqslant \alpha||V||_r$ because any norm monotonically increases with the absolute value of its elements. Similarly we have $||D_\alpha||_r = \alpha||D||_r$ and so $\gamma\alpha$ for γ. We have now

Theorem 4:

For any $\alpha < 1$.

(i) $|\lambda' - \lambda|$ may be made arbitrarily small with a sufficiently small $0 < \alpha < 1$;

(ii) $|\mu' - \mu|$ may be similarly arbitrarily small (hence $\rho(V_\alpha) \leqslant \eta_i$ i=1,2 confirming Theorem 3 by combining (i) and (ii));

(iii) α_1, α_2 of corollary 2, may be located by finding the α value for which $|\lambda' - \mu'| \simeq 0$ for some JOR, or if α_o of corollary 4 has been found then for some $\hat{\alpha}_1$ $\hat{\alpha}_2$, we can locate α_3 α_4 similarly (important because α^* for SOR will be imprecise and should be checked to be in such an interval).

---

1. Note this theorem could be used to replace the last step in the proof of Theorem 8.10.

(iv) That α for which $|\lambda' - 1| \simeq 0$ is the maximum permitted α, corresponding to β of JOR. (This is the only general test we have, and may be done on λ, or on λ' calculated from any α small enough to imply convergence, using the bounds given, to find a more suitable α).

(v) Where $\alpha < 1$, (10), (12), (13) or (14) should be further tightened by multiplication with α. #

We now have a complete comparison theory. It is impossible to develop it further because Theorem 2 may not be further generalised without relationships between SOR and JOR for other classes of matrices beyond those of Theorem 1. Comparison tests have been devised but since they are sufficient conditions their effectiveness may be uneven. They can be improved in a number of ways but considerable emphasis has been given to computational simplicity. All the tests mentioned have some computational saving over a direct comparison of spectral radii for optimal (or nearly so) α values (the logical alternative). The savings depend on the stringency of the test and the sophistication of the bound used. For $B^+ \geq 0$ it is considerable, (the computation of n - 1 eigenvalues, and α^* of JOR, is avoided). For B^+ p-cyclic it may be more (a saving of n eigenvalues, and α^* for JOR) when testing against Jacobi; or rather less if ρ^* of JOR is the alternative. The computation of the bounds is negligable in comparison if $\frac{\sqrt{n}}{\delta} = ||V||_2$, is used - and only slightly more if the norms are actually calculated in a complete test.

9.5 Extensions of SOR: calculating approximations for α^* and reordering rules.

The next problem of interest, in providing general working rules for SOR methods, is how can the estimates of α^* be improved? How may the computation of Chapter 8.5 be simplified? And what are the consequences in making errors in the estimation of α^*? We have seen there the exact calculation of α^* is extraordinarily difficult, especially when n is large, and it will never be

achieved in practice without a complete search (e.g. Bodin (1974)). We noticed that $\rho(R_\alpha)$ is a polynomial of order $\frac{n-2}{n-1}$ in α and will have possibly several local minima in the permitted α range. The approximations of Chapter 8.5 come from smoothing the function ϕ^{-1} and taking the minimum of the result. This is one possible approach. Others, more ad hoc, are considered. Their justification is strictly in computational simplicity, although they have some theoretical rationale as well. We may try:

(i) α^* using (8.23) with λ_i, the eigenvalues of R, in place of μ_i;

(ii) a mixed p-cyclic and approximate α^* (from smoothing or elsewhere) scheme;

(iii) using α^* of p-cyclic case, if B "nearly" p-cyclic;

(iv) using α^* of JOR;

(v) using the limit of optimal nonstationary parameters for semi-iterative methods;

(vi) using iterated vectors of SOR to search a suitable interval of α space for α^*.

The details of the semi-iterative limit method will be discussed in the next chapter when those methods are introduced. Rigorous comparisons among these suggestions are impossible as all involve differing degrees of approximation and "ad hocery". Some rules are however proposed. The six suggestions are not intended to be exhaustive of the possibilities, but they do seem the simplest strategies beside the formal smoothing results.

9.5.1

The demonstration of how $\rho(R_\alpha)$ varies with α is best conducted graphically. Following Varga (1962), consider first the case B is p-cyclic. We wanted to choose α such that for (8.23a), where $\bar{\mu} = \phi_1'(\alpha,B)$, if

$$\lambda^* = [-1+1/\alpha+\psi_1(\alpha,B)] = h(\alpha), \text{ we get } \min_\alpha h(\alpha). \qquad (16)$$

$\bar{\mu}$ may be written as a function of λ^* only, as in (8.24) and hence of α alone as in (8.26). In the special case of B p-cyclic the relation (8.3.4) implies $\bar{\mu} = \pm\lambda^{*\frac{p-1}{p}}\mu$

$$\text{i.e.} \qquad \frac{\lambda^*_i + \alpha - 1}{\alpha} = \pm\lambda^{*\frac{p-1}{p}}_i \mu_i \qquad (17)$$

holds, $2 \leq p \leq n$, $i = 1 \ldots n$.

Let $g_\alpha(\lambda) = \dfrac{\lambda^*_i + \alpha - 1}{\alpha}$ for $0 < \alpha < 2$ and $m(\lambda) = \pm\lambda^{*\frac{p-1}{p}}_i \mu_i$.

$g_\alpha(\lambda)$ is a straight line through (1,1) with gradient (constrained by α) between $[\frac{1}{2},\infty]$, and monotonically decreasing with increasing α. $m(\lambda)$ is symmetric about $m(\lambda) = 0$, and viewing $m(\lambda) > 0$ only, it is monotonically increasing in λ^*_i for μ_i fixed; also monotonically increasing in μ_i since $\frac{\partial m(\lambda)}{\partial \lambda^*_i} \geq 0$ all $\lambda_i \geq 0$ and $\frac{\partial m(\lambda)}{\partial \mu_i} \geq 0$ all $\mu_i \geq 0$. Hence we have

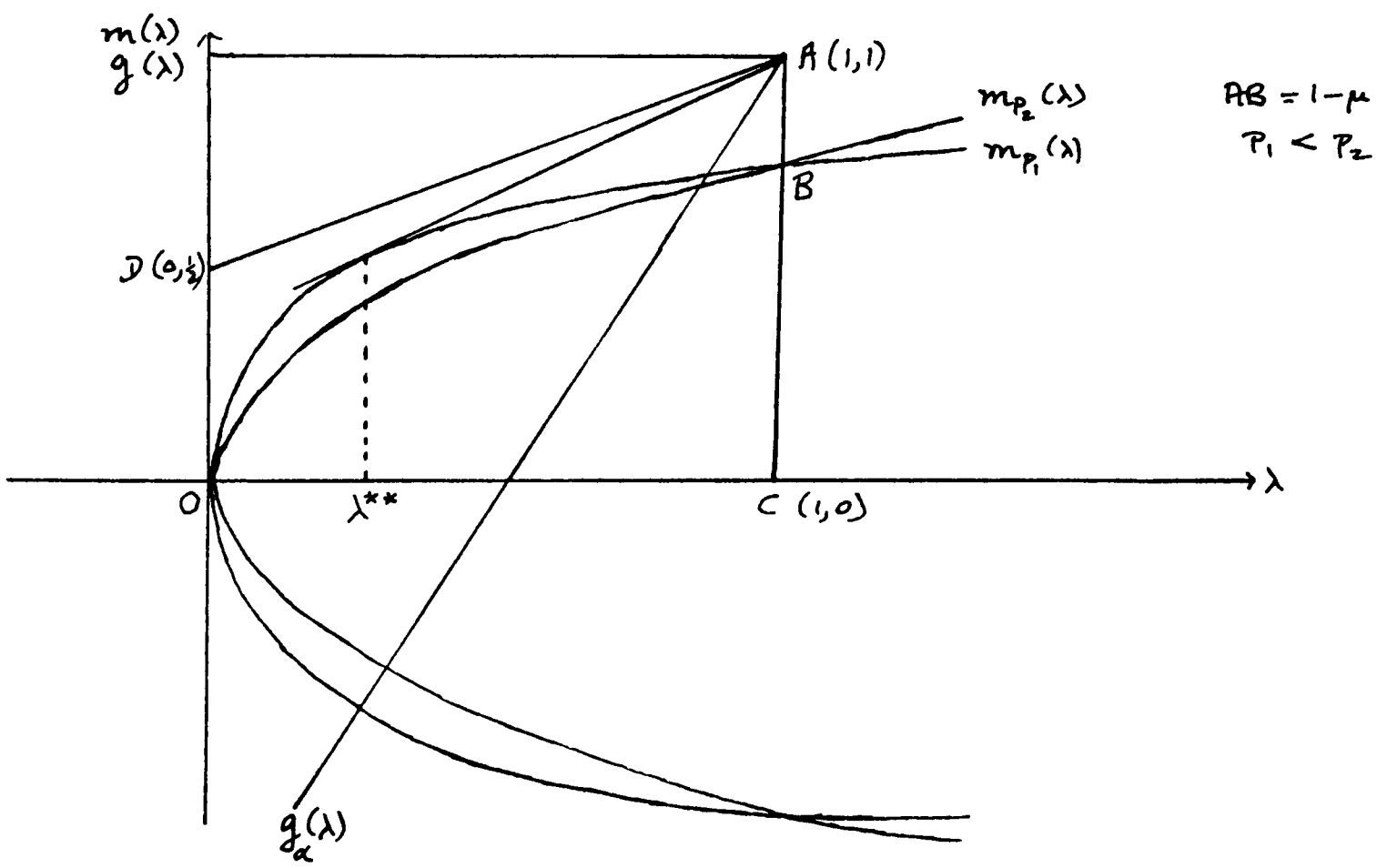


Figure 1: The determination of α^* and λ^{**}

Clearly the α^* which minimises λ^*_i is that which fits $g_{\alpha^*}(\lambda)$ as a tangent to $m(\lambda)$ - this α^* may be obtained with a knowledge only of μ from (8.35)

Three important points emerge. (1) Because of the monotonicity of $m(\lambda)$ in μ_i, $m(\lambda)$ with $\mu = \rho(B)$ is an envelope for other functions $m_i(\lambda)$ with $|\mu_i| \leqslant \mu$. We need only use $m(\lambda)$ with μ to calculate α^* in SOR; a knowledge of all μ_i is unnecessary in constrast to JOR. (2) because of the monotonicity of $m(\lambda)$ in λ, there is only one point of tangency so α^* is unqiue and represents a global minimum. (3) $m(\lambda)$ is written in this case as a function of λ with a fractional power of λ, $\therefore \frac{\partial m(\lambda)}{\partial \lambda_i} \to \infty$ as $\lambda_i \to 0$. The existence of a minimum λ^* is assured.

Continuing, clearly $\alpha < \alpha^* \Rightarrow \lambda > \lambda^{**}$. λ is real at the intersection $g_\alpha(\lambda) = m(\lambda)$; and $\lambda \to \lambda^{**}$ at a rate $\frac{\partial m(\lambda)}{\partial \lambda} = \left(\frac{p-1}{p}\right)\lambda^{-1/p}\mu$ as $\alpha \to \alpha^*$ from above. If $\alpha > \alpha^*$ then $\lambda > \lambda^{**}$ and λ is complex. $\lambda \to \lambda^{**}$ at the constant rate $|\alpha - 1|$, as $\alpha \to \alpha^*$ from below, being the absolute value of the complex roots $(\alpha, \mu \neq 0)$. By the monotonicity of (8.35) an over-estimate of $\mu \Rightarrow$ an over-estimate of α^*, and visa versa, any p. If $p_1 < p_2$ then for given B, $m_{p_1}(\lambda) \geqslant m_{p_2}(\lambda)$ if $\lambda < 1$ and vice versa if $\lambda > 1$. We are only here concerned with convergent systems so we need only analyse the quadralateral ACOD. In that area $m_p(\lambda)$ is bounded above by $m_2(\lambda) = \pm\lambda^{\frac{1}{2}}$ and below by $m_n(\lambda) = \pm\lambda^{\frac{n-1}{n}}\mu$. As $n \to \infty$, $p \to \infty$ is possible and $m(\lambda) \to \lambda\mu$ a straight line. Evidently as p, $n \to \infty$, $\lambda^{**} \to 0$ and $\alpha^* \to 1+$ rapidly. This agrees with Varga (Theorem 4.5 and corollaries) where $\lambda^{**} = (\alpha^* - 1)(p - 1)$ but $R_\infty(R_{\alpha^*}) \simeq \left(\frac{2}{p-1}\right)^{\frac{1}{2}} [R_\infty(R)]^{\frac{1}{2}}$ and $R_\infty(R_{\alpha^*}) \simeq \left(\frac{2p^2}{p-1}\right)^{\frac{1}{2}} (R_\infty(B))^{\frac{1}{2}}$ imply $R_\infty(R_{\alpha^*})$ increases with p monotonically, but the gain in optimal SOR over Gauss-Seidel is maximised when p = 2. λ^{**} decreases with rising p but at a decreasing rate. Most of the results in this analysis of the p-cyclic case appear in both Youngand Varga for p = 2 and $\lambda_i \geqslant 0$ i = 1 .... n only. The generalisations to other p and λ values appear to be new.

We may summarise the effects on $\rho(R_\alpha)$ as α varies as follows

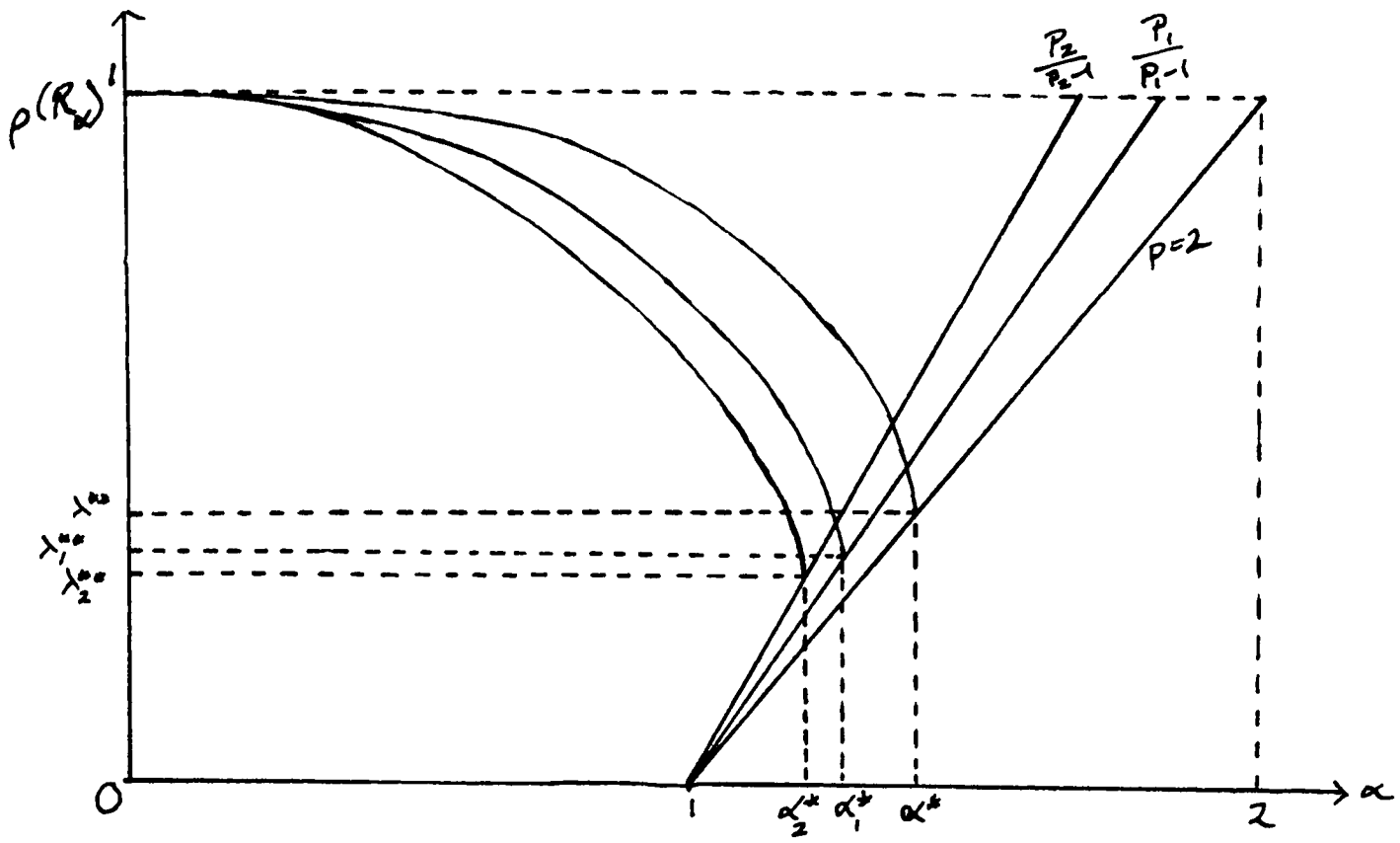


Figure 2: $\rho(R_\alpha)$ as a function of α. The effect of different p values.

$p_2 > p_1 > 2$ for B fixed and p-cyclic, $\mu_i \geq 0$ $i = 1 \ldots n$.

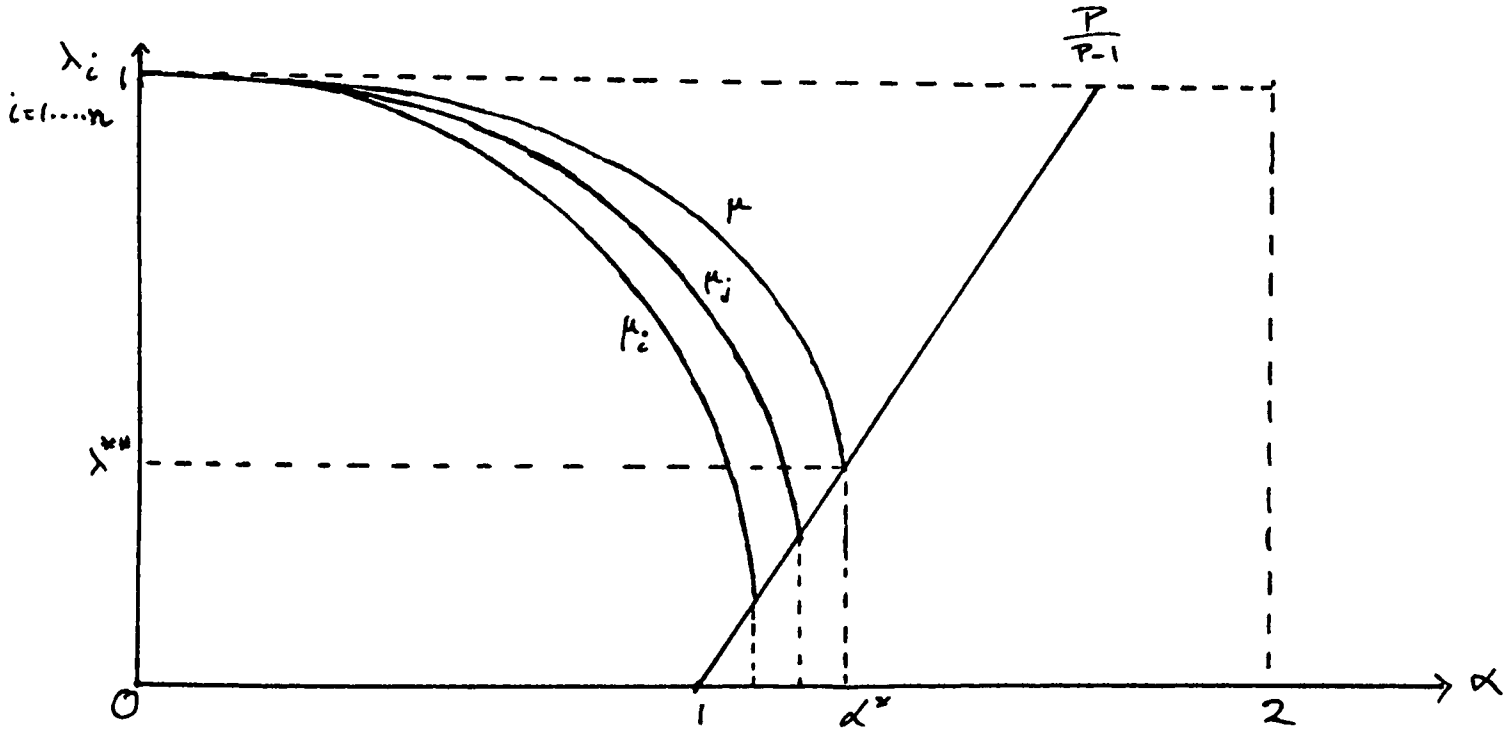


Figure 3: Eigenvalues of R_α as functions of α.

$\mu = \rho(B) > \mu_j > \mu_i$ any two other eigenvalues.

B p-cyclic, p fixed.

9.5.2 The general case:

We have proved for any B;

Theorem 5:

λ^{**} corresponds to a unique and global minimum iff $m(\lambda)$ has

(a) the envelope property; that all other lines, μ_i, μ_j, lie no higher for all $0 < \alpha < \alpha^*$;

(b) monotonicity in λ; ensuring no turning points of any kind in $m(\lambda)$ over the range $0 < \alpha < \alpha^*$;

(c) it is a polynomial in λ, or α, of order ≤ 1; guaranteeing the existence of a minimum (if it is of order 1, $\lambda^{**} = 0$; we assume a convergent B). #

Comment:

These three properties are necessary and sufficient conditions for theorem 5. If (a) or (b) do not hold there will in general be a global minimum but no unique minimum and searching for the global one is, as we have seen, a very burdensome process.

Corollary 6:

When B p-cyclic, an over estimate of α^* is less serious than an under-estimate (the opposite for JOR) but that as p increases the imbalance of severity operates over an increasing interval about α^*. However for a restricted interval around α^* the penalty for mis-estimating α^* is less than miscalculating p. #

Theorem 5 works for arbitrary B provided that $m(\lambda) \leq 1$ for $\lambda \leq 1$. So far we have considered only where $m(\lambda)$ is written in terms of λ. Since $\bar{\mu}$ is the root of (8.24), it is a function of λ^* which varies with α. However we can still apply the above analysis in terms of α. We have $\bar{\mu} = m(\alpha,\lambda) = \psi_1(\alpha,B)$ is (by (8.26) the root of a $\left(\frac{n-2}{n-1}\right)$ - order polynomial in either λ^* or α i.e.

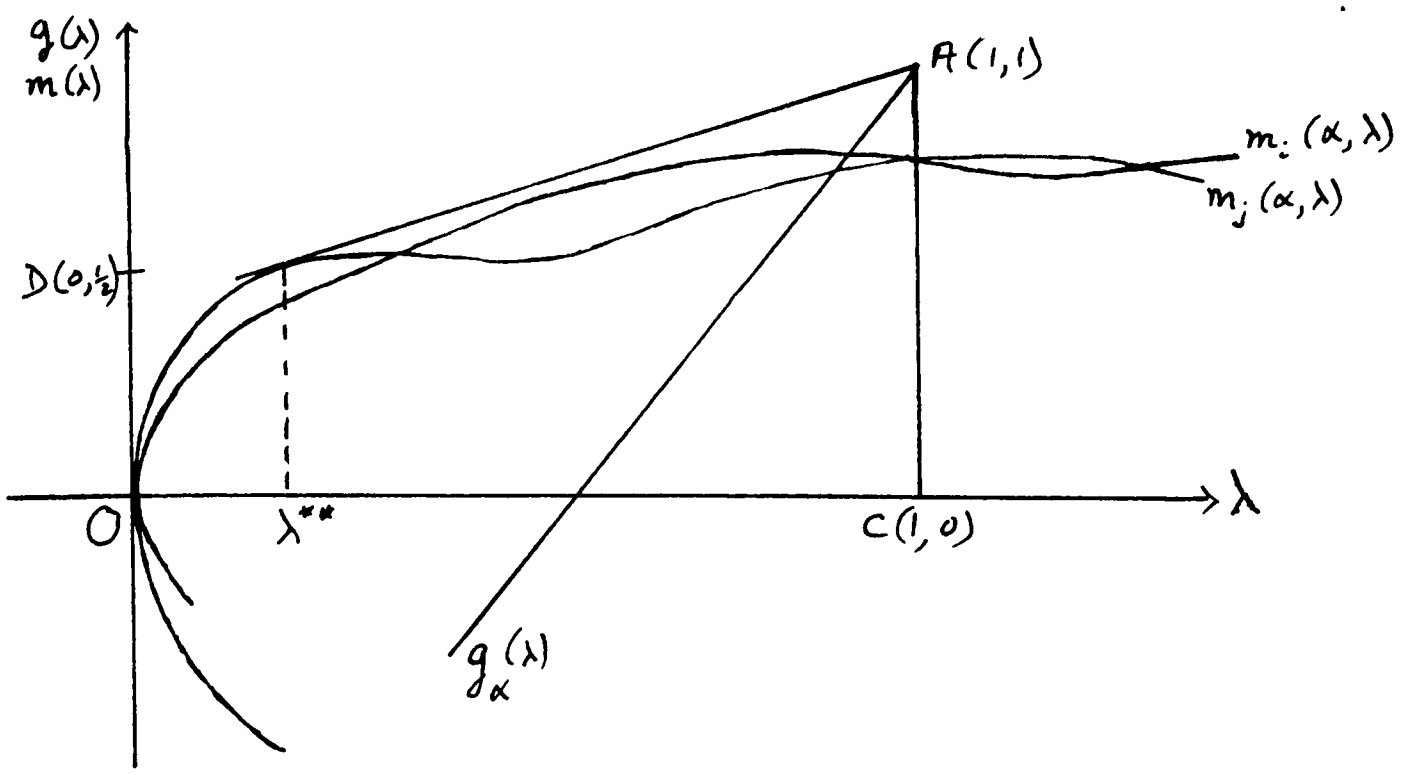


Figure 4: The determination of α^* and λ^{**}, the general case.

However we still want to choose α such that $g_\alpha(\lambda) = m(\alpha,\lambda)$ implies λ is at a minimum; so we could apply exactly the same procedure. For example,

Lemma 2:

for arbitrary B, $\exists$ a global minimum λ^{**} with respect to α and hence an optimal α^*.

Proof:

By (8.26) $m(\alpha,\lambda)$ is a $\frac{n-2}{n-1}$ order polynomial in α and by Theorem 5(c) this lemma follows. In particular $\alpha \to 0$ implies $m_i(\alpha,\lambda) \to 0$ at a rate $\alpha^{\frac{-1}{n-1}} \to \infty$, $i = 1 \ldots n$ (chapter 8.5). Similarly $\lambda \to 0 \Rightarrow m_i(\alpha,\lambda) \to 0$ at the rate $\lambda^{\frac{-1}{n-1}} \to \infty$. Theorem 5 (a) and (b) can be similarly reworked in terms of α. #

In the general case, however, Theorem 5(a) and (b) is not satisfied. Indeed there is no reason to suppose $m_i(\alpha,\lambda)$ all pass below (1,1). Without theorem 5(a) any of $m_i(\alpha,\lambda)$ may have the largest value for a given α, and hence unlike the p-cyclic case each of these $m_i(\alpha,\lambda)$ must be investigated in computing α^*. It may mean local minima at change-over points; from $m_i(\alpha,\lambda)$ to $m_j(\alpha,\lambda)$ being dominant. Without theorem 5 (b) any $m_i(\alpha,\lambda)$ may have turning points in its own path. Local minima are possible for each and hence the

dominant one. Each m_i (α,λ) will be a polynomial of the same order in α (or λ^*) and so will have the same number of turning points but they will not necessarily be in phase. For even quite small values of n there will be a large number of local minima to search through when theorem 5 (a) or 5 (b), or worse both, fail. Notice also that (8.26) contains

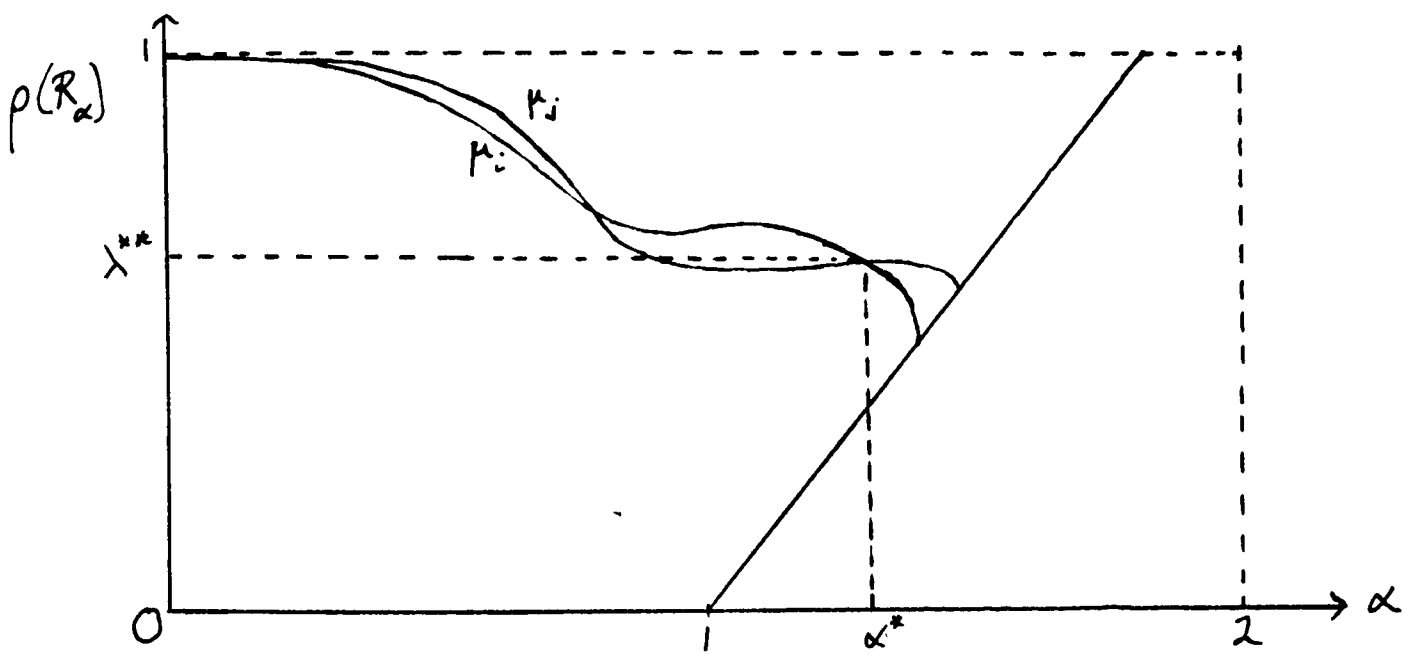


Figure 5: $\rho(R_\alpha)$ as a function of α: the general case.

all the terms of the characterisitc equation of B, so to search the α space is as burdonsome as to calculate to the dominant root of R_α for each α. Because of the local nature of the minima there is no finite strategy - virtually the whole space must be searched. The obvious way to reduce the computation is to smooth the m_i (α,λ), and hence the envelope $m(\alpha,\lambda)$, by approximations. The global minimum of that smoothed envelope is determined, with some effort, in Chapter 8.5.

Another possibility is to establish classes for matrices which satisfy Theorem 5(a) and (b). Then a rational search procedure could easily and rapidly locate the unique global minimum (by branch and bound techniques). Obvious candidates are $B \geq 0$ and B positive definite. In either case in (8.26), $K_1 \geq 0$, $K_{n-1} \geq 0$ and if $\alpha < 1$, $K_2 \geq 0$; but at least some $C_{ji} \leq 0$, $j \geq 2$ and hence $K_i \leq 0$, $i \neq 1, 2, n-1$ is possible. There will be subclasses which have the property $\frac{\partial m(\alpha,\lambda)}{\partial \alpha} \geq 0$, $0 < \alpha < 2$, but they appear to be unduly restrictive

and very difficult to test. This possibility is therefore not a helpful one to pursue even for B "nearly" in the subclasses. $B \geq 0$ or B positive definite gives no help in calculating α^*, despite their special comparison and convergence properties. The envlope and monotonicity properties of $m(\alpha,\lambda)$ depend entirely on the size and signs of elements of B and its sparseness and we cannot generalise further.

9.6 Approximate α^* strategies:

9.6.1

We investigate alternative approximations for $m(\alpha,\lambda)$ which are simpler than smoothing. For $g_\alpha(\lambda) = \frac{\lambda+\alpha-1}{\alpha}$, we might use

$$\bar{\mu} = m_i(\alpha,\lambda) = \left\{ \begin{array}{ll} \mu_i \text{ as in JOR;} & \text{suggestion (iv) above} \\ \lambda\frac{p-1}{p}\,.\ \mu \text{ as if B p-cyclic;} & \text{suggestion (iii) above} \\ \lambda_i \text{ for suggestion (i) above} & \end{array} \right\} i = 1 \ldots n \qquad (18)$$

We know $\bar{\mu}$ is generally rather more complicated than the suggestions (18) imply. However they may be used to estimate α^* if it is considered $m(\alpha,\lambda)$ is adequately locally approximated by one of those simple functions within a neightbourhood of α^*.

Theorem 6: (strategy (iv))

$\lim_{\lambda^*\to 1}$ α^* for SOR is given by (8.16).

Proof:

By (8.20) $\bar{\mu}$ is spectral radius of $(U + \lambda^*L)$. As $\lambda^* \to 1$, $U + \lambda^*L \to B$, so $\bar{\mu} \to \mu$. By Ostrowski, $U + \lambda^*L = B + (1 - \lambda^*)L \Rightarrow |\bar{\mu} - \mu| < (n+2)b(1 - \lambda^*)^{1/n}$ for $b = \max_{i,j} |L_{ij}|$. #

Comments:

The accuracy of the approximation, and hence the interval for which the

approximation is adequate, may be estimated by using the bounds (14) or (15); and it also varies with the sparseness of L. If we assume, or test, that B is convergent we need only be concerned with $\lambda^* \to 1$ -. Secondly, two rules of thumb for this situation are: permute B so that L is as sparse as possible but the absolute size of the elements is as small as possible. If L is normalised so b = 1 (as was L_D earlier) then the priority should be for small elements (i.e. reducing b) since then $(1 - \lambda^*) < 1$ will appear in the bounds (14), (15) in place of δ; whereas $1 \leqslant ||L||_r \leqslant n$. Thirdly, we can always use this α^*, without calculating λ^*_i, if we take a pessimistic view of the work to be done. It is a useful ad hoc "fail safe" procedure. The interval for which $\lambda^* \simeq 1$ is a reasonable approximation is expanded when the ordering rules implied are used. Fourthly, how can we know $\lambda^* \simeq 1$ if (8.16) and μ_i only are calculated? It is possible for some α_o, $\lambda^* \geqslant 1$ when $\mu < 1$. But we can prove

Lemma 3:

If for some α_o, $\lambda^* \leqslant \mu$ where $\mu = \rho(B) < 1$ then for all $0 < \alpha \leqslant \alpha_o$ $\mu \to 1 \Rightarrow \lambda^{**} \to 1$ and conversely; and then Theorem 6 applies with $\mu \to 1$.

Proof:

Here $0 < \dfrac{||R^m{}_\alpha||^{1/m}}{||B^m||^{1/m}} = \sigma_m \leqslant 1$, since both have linear asymptotic convergence rates (Ortega and Rheinboldt (1970)). Also $\sigma = \lim_{m\to\infty} \sigma_m$ is bounded above and below because of the root convergence factors of R_α, B, (assuming convergence and noting theorem 3) i.e.

$0 < r(R_\alpha) < r(B) < 1$; $r(B) = \lim_{m\to\infty} ||\varepsilon^{(m)}||^{1/m} \leqslant \lim_{m\to\infty} ||B^m||\ ||\varepsilon^{(0)}||$ etc.

are norm independent. For $m > 0$, $\exists\, \varepsilon > 0$ so that $\sigma + \varepsilon = 1$. But $\mu = \lim_{m\to\infty} \dfrac{\log ||B^m||}{m}$, $\lambda^* = \lim_{m\to\infty} \dfrac{\log ||R^m{}_\alpha||}{m}$ any α, so $\mu \to 1 \Rightarrow \lambda^{**} \to 1$ and conversely because σ is bounded. $\sigma_m > 0$, consequently the lemma is true for any m. #

Remark:

Clearly strategy (iv) provides the right α^* in models where convergence is difficult and slow.

It is often claimed optimal SOR is $2\sqrt{2}$ times faster than Jacobi (Westlake (1968) etc.). This is in fact true only for B 2-cyclic, but we may take it as a likely rough upper bound of improvements in practical cases. Thus if $\mu \geq .95$, $\lambda^{**} \geq .88$; $\mu \geq .95$ is a reasonable rule of thumb, based on the empirical evidence of relative performances, for applying Theorem 6 and using α^* from (8.16).

Strategy (iii):

If the polynomial $m(\alpha,\lambda)$ may be locally approximated by a fractional power of λ then we may use the α^* of (8.35) as the approximation. This strategy is difficult to use. By the continuity theorem for eigenvalues, the root of (8.35) becomes an arbitrarily good approximation to α^* as B becomes p-cyclic. This root should be used if B is nearly p-cyclic, i.e. the bounds (14) or (15) satisfy $|\lambda' - \lambda| < \varepsilon$ rather than $|\lambda' - \mu| < \varepsilon$ as before. However, the strategy may also be effective when B is not nearly p-cyclic but $m(\alpha,\lambda)$ happens to be well approximated by a fractional power of λ. It seems unlikely we would ever have a priori knowledge to justify the assumption, nor is it clear what that knowledge would be.

Strategy (i):

As $p, n \to \infty$ in (8.34), $m_i(\alpha,\lambda) \to \lambda_i\mu$. The approximation is rapid as $n \to \infty$. The third strategy of (18) is therefore equivalent to assuming B approximately is p-cyclic and sparse enough. (Where $\mu \neq 1$, we can actually use $\lambda_i\mu$ for λ_i.) This strategy is also useful when $\alpha \to 1$ because then $\lambda^* \to \lambda$. There will be at least one other value of α for which this is true but generally it will be just as difficult to determine that value(s) as α^* itself, so unless it is assumed that α^* is close to 1 (or the value(s) for which $\lambda^* \simeq \lambda$) this point is of little significance. The strategy is useful

for medium, large or very sparse systems - say $\sum_1^M r_i \geq 30$ with $n \geq d \sum_1^M r_i$; d = density $\sum_1^M r_i / M^2$; that is for most typical econometric models (2.5).

Remark:

A general approximation to $m(\alpha,\lambda)$ may be achieved with a Taylor's series expansion $m(\alpha,\lambda) \simeq m(\alpha_o,\lambda) + (\alpha - \alpha_o)\, m'(\alpha_o\lambda) + \ldots$. We have

$$\phi'(\bar{\mu}) = \sum_{i=2}^{n-1} \frac{\partial K_i}{\partial \alpha}\, \bar{\mu}^{n-1} \quad \text{from (8.36) where}$$

$$\frac{\partial K_i}{\partial \alpha} = \sum_{j=1}^{n=2} \frac{\partial C_{ji}}{\partial \alpha} = \sum_{j=1}^{n-2} \sum_{k=1}^{j} k\, \alpha^{k-1}\, d_{i,k} \qquad (19)$$

$$\therefore \quad \phi'(\bar{\mu}) = \sum_{i=2}^{n-1} \sum_{j=1}^{n-2} \sum_{k=1}^{j} k\, \alpha^{k-1}\, d_{ik} \text{ and } m'(\alpha,\lambda) \text{ is the } (n-1)^{th} \text{ root.}$$

Contenting ourselves with a linear approximation, from (16), $\lambda^* \simeq$ constant $+ \frac{1}{\alpha} + \alpha m(\alpha_o,\lambda)$ about α_o. Then $\alpha^* = +\sqrt{\frac{1}{m'(\alpha_o,\lambda)}}$ is the rule, noting this is only sensible provided $m'(\alpha_o) \geq 0$. This suggests an iterative search procedure from an arbitrary start α_o. The difficulty is to obtain $m'(\alpha,\lambda)$ from $\phi'(\bar{\mu})$; for large n, it is no easier than to get $m(\alpha,\lambda)$ from $\phi(\bar{\mu})$. It could easily be done in special cases, i.e. $\alpha = 0$, $\alpha = 1$ but we already have approximations for those cases. Further convergence to α^* might be a problem if the approximation was bad.

9.6.2

It will be shown in the next chapter that the semi-iterative SOR method (Strategy (v)) has parameters tending to a limit, where they are optimally computed, as the number of iterations increase. Moreover this limit may be determined at the start. The limit might be used as a stationary α. In fact the limit turns out to be α^* of B 2-cyclic (or p-cyclic for more sophisticated semi-iterative versions) and so it is actually the same strategy as (iii).

Strategies (ii) and (vi) are in fact nonstationary methods (as defined in Chapter 10) but are best noted here.

Suppose B nearly p-cylcic. Then optimally solve $A^+ Y = b$ by SOR, where A^+ is p-cyclic, in place of (8. 5) on a weak criterion to give an approximate solution $\hat{Y}$. Then solve $A\hat{Y} = b$ to be desired accuracy with the same or some other approximate α^*. This strategy may be used where B is near any special property, but the likely p-cyclic case is the only one where we can get a precise α^* for A^+.

Lemma 4:

(a) The convergence conditions for linear (8.10) are not altered. By Young (1971) (Theorem 9.1.1) a composite method is convergent iff its components are convergent - and if B is not convergent then the second stage will not converge.

(b) However, in the nonlinear case,[1] when B^+ is within its neighbourhood of convergence but B is not, we will still get convergence provided the second stage start $\hat{Y}$ lies within the neighbourhood of convergence of B. #

The strategy is therefore particularly powerful in nonlinear iterations but may also be beneficial with linear equations.

Lemma 5:

Asymptotic convergence rates are not improved.

Proof:

The total iteration is $Y^{(s+q)} = R^s_{\alpha_2^*} (R^+_{\alpha_1^*})^q Y^{(0)}$; $R^+_{\alpha_1^*}$ from B^+, α_1^* by (45). In the linear case the asymptotic rate of convergence $= \rho(R^s_{\alpha_2^*}(R^+_{\alpha_1^*})^q) = \lim_{s+q\to\infty} \frac{\log}{s+q} ||R^s_{\alpha_2^*} (R^+_{\alpha_1^*})^q|| = \rho(R_{\alpha_2^*})$, for q fixed, is invariant to the start vector. #

Let $\rho_1 = \rho(R^+_{\alpha_1^*}) = (\alpha_1^* - 1)(p - 1)$ and $\rho_2 = \rho(R_{\alpha_2^*})$.

Theorem 7:

A mixed α^* strategy has a faster average rate of convergence than a simple strategy if $\rho_1 < \rho_2 < 1$ for some α_2^*, when $s \leq q \left(\frac{\log \rho_1}{\log \rho_2} - 1\right)$.

---

1. This observation is important for application to recursive iterative estimators whose iteration is always nonlinear. Strategy (ii) is therefore useful in iterative estimation for setting α^*.

Proof:

By Theorem A1 $\rho(R^s_{\alpha_2^*}(R^+_{\alpha_1^*})^q) \lesssim \min\{\rho_2^s, \rho_1^q\}$. But $\rho(R^{(s+q)}_{\alpha_2^*}) = \rho_2^{s+q}$ and $\rho_1^q < \rho_2^{s+q}$ iff $\rho_1 < \rho_2$, ($\rho_2^s < \rho_2^{s+q}$ is impossible for $q > 0$ ρ_1 and $\rho_2 \lesssim 1$). We require at least

$$[(\alpha_1^* - 1)(p - 1)/\rho_2]^q \lesssim \rho_2^s \tag{20}$$

Write $\rho_1 + \varepsilon = \rho_2$, $\varepsilon > 0$. We need $q(\log \rho_1 - \log(\rho_1 + \varepsilon)) \lesssim s \log(\rho_1 + \varepsilon)$. Note q is fixed by the tolerance τ_1 on $\hat{Y}$. But since $0 < \rho_1 < \rho_2 < 1$ $\log \rho_1 < \log \rho_2 < 0$ so we set s such that

$$1 > \frac{q}{q+s} \gtrsim \frac{\log \rho_2}{\log \rho_1} \quad \text{or} \quad 0 < s \lesssim q\left[\frac{\log \rho_1}{\log \rho_2} - 1\right]. \tag{21}$$

Further for the overall tolerance τ_2 on Y, let m_1 = number of iterations to convergence for the mixed scheme and m_2 = that for the simple α_2^* strategy, then

$$m_1 = s+q = \frac{\log \tau_2}{\log \rho'(R^s_{\alpha_2^*}(R^+_{\alpha_1^*})^n)} < \frac{\log \tau_2}{\log \quad _2} = m_2. \quad \text{Let } t = \left[\frac{||\varepsilon^{(m)}||}{||\varepsilon^{(0)}||}\right]^{1/m}$$

for any norm where $\varepsilon^{(m)} = Y^{(m)} - Y^*$ the error vector. $t \lesssim (||A^m||)^{1/m} = e$ for the iteration matrix A[$R(A^m)$ = average convergence rate]. t is the average error reduction per iteration. Now $m_1 = [R(R^s_{\alpha_2^*}(R^+_{\alpha_1^*})^q)]^{-1}$, $m_1 = s+q \Rightarrow t^{m_1} \lesssim e^{-1}$, for a suitable τ_2, and $m_1 < m_2$, gives a measure of the iterations necessary to reduce the norm of the error vector by a factor e. #

As long as s is a smaller precentage of q than the percentage speed increase of ρ_1 over ρ_2, the mixed strategy is superior in average rates of convergence. In practice $\tau_2 > 0$ so we are actually only concerned with average rates, although m_1 may be very large and the average rate arbitrarily

close to the asymptotic one. Hence $m_1 < m_2$ is always true for any finite q,s if (21) holds. The mixed strategy is superior for all finite m_1 but the advantage diminishes as m_1 increases. One procedure is to set τ_1 and iterate q steps with B^+ and then with B as far as the limit on s. If convergence has not been achieved return to B^+ and repeat.

Strategy (vi):

We could apply von Mise's theorem. Let R_{α_o} have eigenvalues $\lambda_1 \geq \lambda_2 \ldots$ n for some value α_o. We have for real $|\lambda_1| \geq |\lambda_2|$, $\lambda_1 = \lim_{i\to\infty} \left(\frac{Y_1^{(i)}}{Y_1^{(i-1)}} \right)$ since $Y^{(i)} = R^i_{\alpha_o} Y^{(0)}$ for arbitrary $Y^{(0)}$.

If λ_1 is complex we observe irregularly alternating signs, and find i for which $X_i = [Y^{(i)} : Y^{(i+1)} : Y^{(i+2)}]$ has rank 2. Solve for c,d in $\lambda^2 + c\lambda + d = 0$ noting $c = -2a$, $d = r^2 = a^2 + b^2$ for $\lambda_1 = a + ib$. For some α_o arbitrarily selected we can estimate $\lambda_1^{(0)}$ as we solve (8.5).

The simplest plan is to select α_o by one of the other α^* approximations of this chapter, or as α^* from smoothing; then estimate $\lambda_1^{(0)}$ by iterated vectors $Y^{(i)}$. Then another value α_1 should be tried and $\lambda_1^{(0)}$ estimated. Continuing, a sequence $\lambda_1^{(i)}$ may be generated to search the interval $0 < \alpha_i < 2 \quad i = 1, 2, \ldots .$ We want α_{i+1} such that $|\lambda_1^{(i+1)}| < |\lambda_1^{(i)}|$ for each i, but unless we know something of the form of Theorem 5 we must first search [0,2] with a wide grid and then progressively reduce it within a useful subinterval. There is no procedure to guarantee the inequality for each i.

There seems to be no short cut to searching most of the interval, but it may be possible to reduce it. For example if B p-cyclic $[0, \frac{p}{p-1}]$, so if it is nearly p-cyclic that will be a good approximation. If B is diagonally dominant [0,β]; if we apply Theorem 4 we may calculate an upper limit; if $|\mu|$ is large we need only search smaller α values; we may reject very small and large α, say outside [.25, 1.75], as giving $|\lambda_1|$ too close to 1 to be of

any interest. However, it is best to calculate α^* by some approximation mentioned earlier and search only a small interval about that value.

Writing g as the bound (13) or (15) when $\alpha = 1$, and g_α when $\alpha \neq 1$, Theorem 4(v) states $g_\alpha \leqslant \alpha g$ all $\alpha \leqslant 1$. The search strategy; find $\alpha_{i+2} \leqslant 1$ to maximise $|\lambda_1^{(i)}| - |\lambda_1^{(i+1)} + \alpha_{i+2}\, g|$ after competing stage $(i + 1)$, will rapidly complete the search of [0,1]. We should start with maximising $|\mu| - |\lambda_1^{(0)} + \alpha_1\, g|$, $\alpha_1 \leqslant 1$, for $\lambda_1^{(0)}$ with an approximate α^*. A grid search must then be applied for $\alpha > 1$. $\lambda_1^{(i)}$ must be reestimated after an arbitrarily set number of iterations Z. Suppose the grid width were set at ε and were reduced by a factor of 10 at every pass over a subinterval of length ℓ (also therefore progressively reduced) then $Z \leqslant \dfrac{\log \tau_2 \cdot \varepsilon}{\log \lambda_1^{(0)} \ell}$ seems a sensible rule to reduce the actual number of iterations.

The major disadvantage of this strategy is its unreliability. If $\lambda_1 \simeq \lambda_2$ .... we rapidly get a crude estimate of $\lambda_1^{(i)}$ but may slowly get a good estimate of it. In cases where $|\lambda_1^{(i)}| \simeq |\lambda_1^{(i+1)}|$ any inaccuracy in the estimate, if Z is not large enough, may easily cause us to reject α_{i+1} as an improvement when for larger Z we would accept it. In particular if $\lambda_1^{(i)} = a + ib$, X_Z may not have rank 2 after only Z iterations and since we are only interested here in $\sqrt{d}$, we must choose $\hat{c},\hat{d}$ such that $Y^{(Z)} + \hat{c}Y^{(Z-1)} + \hat{d}Y^{(Z-2)}$ is minimised in sum of squares, i.e. $\hat{d}$ from $\begin{pmatrix}\hat{c}\\ \hat{d}\end{pmatrix} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^{(Z)}$ with $X^* = Y^{(Z-1)} : Y^{(Z-2)}$. $\hat{d}$ is a consistent estimate, with "sample size" Z, in the sense that $Y^{(Z-1)}$, $Y^{(Z-2)}$ are predetermined, or lagged dependent, variables but the "disturbance term" becomes orthogonal as $Z \to \infty$. Thus $\hat{d}$ may be rather inaccurate if Z is not very large. The problems of unreliability are likely to be very real; Bodewig (1956) suggests $|\lambda_1/\lambda_2| > 3$ is required for rapid estimation of λ_1, which we will almost certainly never find. It is extremely unlikely that the starting vector $Y^{(0)}$ should be

nearly orthogonal to the eigenvector of λ_1 (which would make this method impossible) but to guard against bad starting vectors each group of iterations should be started with the current approximation $Y^{(Z)}$ to the solution of (8.5). The method then solves and searches for α^* concurrently. (The well known problems of extracting other eigenvalues (Bodewig) are irrelevant.)

In view of the large amount of searching and the inevitable unreliability of $\lambda_1^{(i)}$ this strategy can only be recommended as an addition to other methods: specifically to search a small interval about α^* approximated. Similarly the mixed α^* strategy must, by its nature, be regarded as an addition to the other methods. We conclude we may treat strategies (i) - (vi) in two groups. The first group, Section 6.1, including smoothing, provide α^* as approximations to that of the global minimum and are suitable in various different circumstances. The second group, Section 6.2, supplement the first and may be used to improve the α^* from a first group method.

9.7 Permuted matrices:

In addition to the rules obtained in Chapter 8.7 for permuting the matrix B to produce a good reordering and renormalisation, Theorem 6 indicated another rule. That rule applies only to the triangular matrix L so it should be made complementary to those results.

The qeustion may arise, if B is nearly p-cyclic, which of two possible $p_1 < p_2$ should be taken? p_1 will make D more sparse but if p_2 tends to make $|d_{ij}|$ $i,j = 1 \ldots n$ smaller there may be a trade off of smaller elements for a sparser matrix. The objective would be to minimise $(\alpha^* - 1)(p - 1) + \varepsilon_{\alpha^*}$, where ε_{α^*} is the bound (13) or (15) with respect to p, which involves minimising $||V_{\alpha^*}||_r$ with respect to p. Formally this is an unpleasant nonlinear integer programming problem. Simple rules will give improved solutions without the difficulty of the optimisation. By Theorem 6 we should permute B so

that L is sparse and has the smaller elements, priority given to the smaller elements. This causes $\alpha^* \rightarrow$ the solution of (8.23) and also will fit in with making B nearly p-cyclic. We should then choose p_2; we cannot estimate α^* precisely but can get close to it so we apply the corollary 6 - a safety first policy. By (6), (we have already got the smallest elements in L_D, L) we want maximum sparseness in D to make the elements in V small by annihilation in the summed products. Then the bounds (13), (15) will be small via small δ, rather than $1 \leq ||V||_r \leq n$ itself being small. Hence p_1 and p_2 are the values which are nearest to maximising the sparseness of D. p_2 also reduces $(\alpha^* - 1)(p - 1)$ independently of the bounds; we forego less of the possible improvements.

A computer programme is presented by Young(1971) to test if the elements of the matrix B may be permuted to give an iteration matrix which is p-cyclic. B matrix is utilised by the programme which then prints out the required permutation matrix. We therefore use this programme to test if B may be p_2-cyclic, storing those elements, if any, which violate this proposition. By the above we want then to permute B so that the fewest, and more importantly the smallest, elements are in positions violating the p_2-cyclic form. The programme could be rerun on the same lines to test for further improvements; and then repeated till no further improvements in the "near" p_2-cyclic structure are found. This adaption may not be the most effective in terms of reducing the eigenvalues of the iteration matrix but it should produce a suitable structure very quickly.

CHAPTER 10: COMPLETE MODEL SOLUTION II: NONSTATIONARY METHODS

10.1 Introduction

The analysis of the previous chapter has dealt with the better known iterative methods of solving large systems of simultaneous equations. In this chapter we widen our scope to consider other iterative methods. We need to find a method which will converge in any circumstances - especially in those when SOR or JOR will not. We must investigate if there may be other accelerations faster than JOR or SOR. We need also to consider how these methods function in nonlinear equation systems and compare them with methods specifically designed for the nonlinear case. The nonlinear analysis is particularly important since this is the likely situation in econometric models but also it certainly will be the case when we apply these results to iterative estimation. After comparisons we can make some recommendations on which method should be used.

10.2 Richardson's Stationary Methods:

One simple method - a first order linear stationary method - not considered so far, but shown by Young to be completely consistent, is Richardson's stationary method. For system (8.5) we propose the iteration

$$Y^{(s)} = Y^{(s+1)} + P(AY^{(s-1)} - b) \qquad (1)$$

where P is any nonsingular matrix. The method has iteration matrix $G = I + PA = I + P - PB$. If $P = pI$, p a scalar, Young refers to the stationary Richardson method and if $P = \{\text{diag } P_i\}$ he refers to the generalised Richardson method (GRF). There is however no reason why P should not be an arbitrary nonsingular matrix and because it will prove convenient later in the

analysis of iterative estimators, as well as for the benefit of generality, I shall broaden the definition of GRF to include this case. It will turn out that it is impractical to derive results for what would be a "good" general P matrix in any system. Convergence conditions depend on the eigenvalues of P so no improvements in results follow from this extra generality beyond the similarity transform, (a diagonal P matrix) - and even then a unique optimal P does not exist. The generalisation is of academic interest for model solution and is therefore only pursued for use in classifying iterative estimators.

10.2.1 Richardson's Stationary Method:

Theorem 1:

Let P = pI, and B have eigenvalues $\mu_j = a_j + ib_j$ $j = 1 \ldots n$. Then $\exists$ a p such that $\rho(G) < 1$ if and only if either $a_j < 1$ or $a_j > 1$ for all $j = 1 \ldots n$.

Proof:

From (1) $Y^{(s)} = G^s Y^{(0)} - \sum_{k=1}^{s-1} G^k Pb$ converges iff $\rho(G) < 1$.

But $\rho(G) = \max_j \{|1 + p - p\mu_j|\} = \max_j \{|1 + p(1 - a_j) - pib_j|\}$. The squared j^{th} root radius of the characteristic equation for G is therefore $r_j^2 = [1 + p(1 - a_j)]^2 + p^2 b_j^2$ which is less than unity if

$$p[2(1 - a_j) + p((1 - a_j)^2 + b_j^2)] < 0 \qquad (2)$$

Since $(1 - a_j)^2 + b_j^2 \geq 0$ we can satisfy (2) when $a_j < 1$ with $p < 0$ and

$|p| < \dfrac{2(1 - a_j)}{(1-a_j)^2+b_j^2}$ but never with $p > 0$; or when $a_j > 1$ with $p > 0$ and

$p < \left|\dfrac{2(1 - a_j)}{(1-a_j)^2+b_j^2}\right|$ but never with $p < 0$. Thus for either $a_j < 1$ or $a_j > 1$ for all $j = 1 \ldots n$ we can find some p such that

$$|p| < \min_j \left| \frac{2(1 - a_j)}{(1-a_j)^2 + b_j^2} \right| \qquad (3)$$

with the sign chosen so that $r_j^2 < 1$ all $j = 1 \ldots n$. But if say $a_k < 1$ and $a_m > 1$ we can never choose a single p such that both $r_k^2 < 1$ and $r_m^2 < 1$ simultaneously. #

It may appear that Theorem 1 represents an advance over JOR and SOR methods. However if $a_j > 1$ $j = 1 \ldots n$ then by Theorems 1 or 8.10 the solution of $-AY = b$ is possible using either JOR or SOR. The negative of the solution obtained is the solution to (8. 5) which we require. This procedure is open to us without Richardson's stationary method and is equivalent to using JOR or SOR with relaxation factor $-\alpha$ in place of α. Indeed the relationship is even closer as is suggested by the similarity of (3) to (8.13).

Lemma 1:

Richardson's stationary method is identical to JOR with $p = -\alpha$ if $a_j < 1$ $j = 1 \ldots n$ or the euqivalent negative JOR procedure mentioned above if $a_j > 1$ $j = 1 \ldots n$.

Proof:

The iteration matrix $G = I + pI - pB$ but JOR has $B_\alpha = \alpha B + (I - \alpha)I$ $\therefore$ $p = -\alpha$. Since JOR has $0 < \alpha < \min_j \left| \frac{2(1-a_j)}{(1-a_j)^2+b_j^2} \right|$ if $a_j < 1$, $j = 1 \ldots n$ (by the corollary to Theorem 8.3) it is identical to Richardson's stationary method in Theorem 1 above. Similarly if $a_j > 1$, the solution would be obtained as the negative of that to JOR with $-\alpha$ (by theorem 8.5) which is identical to Richardson's stationary method in Theorem 1 above with $p = -\alpha$. Note $p = -1$ gives the Jacobi method immiedately. #

Corollary 1:

The bounds and optimal values of p follow immediately and are identical to those of $-\alpha$ in JOR theory of the previous chapter. #

Corollary 2:

The Richardson stationary method may also be accelerated recursively to obtain the iteration $Y^{(s)} = (1+pL)^{-1}\left((1+p)I-pU\right)Y^{(s-1)} - pb$ from (1) (or from Lemma 8.2 noting $B_{-\alpha} = G$) i.e. SOR with $p = -\alpha$. Consequently the bounds and optimal values of p are those for $-\alpha$ in SOR theory. The Richardson stationary method may also be accelerated by other methods (notably the semi-iterative ones) where the properties of p will be apparent by analogy with JOR or SOR under the same accelerations. #

10.2.2 GRF methods

One obvious generalisation of JOR, or SOR, that has not been considered is to let relaxation factors to vary from equation to equation. That is we write $\bar{\alpha} = \{\text{diag } \alpha_i\}$ $i = 1 \ldots n$ for α in B_α or R_α.

Lemma 2:

The generalisation of JOR or SOR with relaxation factors varying over equations is identical to the use of GRF with $P = -\bar{\alpha}$.

Proof:

from Lemma 1 and Corollary 2 above. #

Since P is arbitrary, with the exception of $P = 0$, we can always consider first transforming (8. 5) into another equivalent system. Such a device would be particularly useful if convergence using A proved difficult. The transformation, noted by Young, is to scale the system by a matrix E and letting $W = E^{-1}Y$ we have in place of (8. **5**)

$$E\ A\ E\ W = Eb \qquad (4)$$

Lemma 3:

The GRF method is equivalent to the Richardson method on a prescaled system.

Proof:

The two approaches are equivalent if they have identical eigenvalues. Consider scaling (8.5) with $P^{\frac{1}{2}} = E$ in (4). We have then $P^{\frac{1}{2}}AP^{\frac{1}{2}}W = P^{\frac{1}{2}}b$ which if solved with Richardson's method and $p = 1$ means iteration $W^{(s)} = (I + P^{\frac{1}{2}}AP^{\frac{1}{2}})\, W^{(s-1)} = (P^{-\frac{1}{2}} + P^{\frac{1}{2}}A)\, Y^{(s-1)}$. Premultiplying by $P^{\frac{1}{2}}$ yields $Y^{(s)} = (I + PA)Y^{(s-1)}$ which is exactly GRF iteration on (8.5). #

Obviously there is no unique correspondence between the GRF method with P and the Richardson method. If we had used $E = aP^{\frac{1}{2}}$, for some scalar a, then $p = a^{-2}$ would give the same GRF.

Theorem 2:

For an arbitrary matrix A, the GRF method will converge when applied to (8.5) for a suitable matrix P.

Proof:

Consider the GRF method as Richardson's method on a prescaled system (8.5) with arbitrary A. We seek a scaling matrix E such that $\rho(EAE) < 1$ for use in theorem 1. If $\rho(A) < 1$ then by Lemma 3, $E = -I$ implies, for $p = 1$, convergence of GRF with iteration matrix $I - A$ (i.e. the Jacobi method). Suppose $\rho(A) > 1$. Then by Theorems A1 and A2 we can always choose E real and nonzero where $\rho(E) < 1$ such that $\rho(AE) < 1$. If A has eigenvalues $1 - \mu_j$ then by Stein's Theorem[1] if $\rho(E) < 1$, for some H, $H = EHE' = I$; so the choice of such an E implying

$$0 < h_j \left(1-(1-\mu_j)^{-1}(1-\bar{\mu}_j)^{-1}\right) < 1, \quad j = 1 \ldots n. \tag{5}$$

where $\bar{\mu}_j$ is the complex conjugate of μ_j and H has eigenvalues $h_j \geq 1$, would give $\rho(AE) < 1$. Further by Theorem A1 the product of two convergent matrices is itself convergent; hence $\rho(EAE) < 1$ for this E. Now if $\rho(EAE) < 1$, by Theorem 1, Richardson's stationary method applied to the scaled system (4) will be convergent. By Lemma 3 this prescaling and Richardson procedure is always equivalent to some GRF. Thus a suitable $P = p^{-2}E^2$ may always be

---

1. Equally E may be taken as complex but this seems an unnecessary complication here.

chosen such that GRF converges for any equation system (8.5). #

Corollary 3:

Recursive acceleration of GRF may also always be made to converge with a suitable P; by the same argument in the proof of Theorem 2 and Corollary 2. #

As will become apparent later GRF is not the only method of last resort - any semi iterative acceleration is also convergent for any equation system. There are a number of disadvantages in the GRF procedure which make it difficult to use in practice. First we have to have aprecise knowledge of all the eigenvalues μ_j of B. This we may not have, say, if a Meyer-Spasche test had been performed on B. If the intention, however, was to calcualte α^*, the μ_j will be known. We will be alerted to the possibility of GRF by the failure of the Meyer Spasche test(s), or a convergence failure. We also require to know the positions of the eigenvalues, which means obtaining them from a QR transform of B, in order that the corresponding bound for h_j may be obtained from the eigenvalues of A. The μ_j occur in a arbitrary order in the similarity transform of A (Peters & Wilkinson (1970)). There is substantial computation in obtaining a suitable P. However, we note

Corollary 4:

Theorem 2 holds when P is restricted to being diagonal.

Proof:

As for Theorem 2 above with E diagonal (see also Theorem 8 of Taussky (1964) for shortening the proof in this special case). Then $P = E^2$. #

Under this restriction, E may be obtained from H. If E is diagonal we may also take H to be (Theorem 5 in Taussky (1964)). By Stein's theorem (see theorem A2).

$$E = \text{diag}\left\{e_j = \pm\sqrt{\frac{h_j - 1}{h_j}}\right\} \qquad (6)$$

But also from (5)

$$h_j < \frac{1-\mu_j-\bar{\mu}_j+\bar{\mu}_j\mu_j}{-\mu_j-\bar{\mu}_j+\bar{\mu}_j\mu_j} \tag{7}$$

and if $\mu_j = a_j + ib_j$

$$\text{then} \quad e_j < \pm \sqrt{\frac{1}{(1-a_j)^2 + b_j^2}} \tag{8}$$

Thus

Theorem 3:

GRF convergent iff P has eigenvalues $|p_j| < \left|\frac{1}{(1-\mu_j)(1-\bar{\mu}_j)}\right|^{\frac{1}{2}}$

corresponding to μ_j of B, j = 1 .... n.

Proof:

The eigenvalues of A have negative real parts when $a_j < 1$. By Theorem 1 if E scales A to having eigenvalues within the interval [-2,0] then the iteration matrix I + EA is convergent. (8), therefore, by Theorem A2 is the bound on the absolute values of the eigenvalues of P if it is not diagonal - the unitary similarity transforms do not affect the eigenvalues. #

There remains the problem of the sign of p_j. In general either sign will not imply convergence. Having determined a suitable absolute value for p_j we should have to work out which combination of signs gave eigenvalues of PA ε [-2,0] - a long tedious business. But by Lemma 3 this difficulty is obviated if $E = P^{\frac{1}{2}}$ is chosen where $\rho(EA) < 1$. Then because of the double use of E, for at least p = 1, convergence is assured although it may not be very fast. Theorem 3 agrees with Young's Theorem 4.3.8 which applies only to the special case A symmetric and positive or negative definite.

The next problem is to find some method of obtaining an optimal, or failing that a good P for a given A. It appears that an expression for an optimal P is not generally available since no explicit relationship is available between the eigenvalues of PA and A as a function of the elements of P.

Bounds on the eigenvalues of PA may be obtained but they will only permit a "good" P to be calculated. Since the iteration matrix is I + PA we want P such that PA has eigenvalues clustered as closely as possible about -1. However, bounds localising the eigenvalues of a matrix such as Gerschgorin, Casini's Ovals, etc. do so only in terms of absolute values and for spectral radii and are therefore useless for designing P. We define: the spread of a matrix $s(B) = \max |\mu_i - \mu_j|$. Then, (Marcus and Minc (1964) Theorem 4.2.3), $s(B) \leq [2(1 - 1/n)(\text{tr } B)^2 - 4 \Sigma \Delta_i^{(2)}]^{\frac{1}{2}}$; $\Delta_i^{(2)} = i^{th}$ principle minor determinant of B, order 2. Thus $s(PA) \leq \sqrt{2}[(1-1/n)(\Sigma p_i)^2 - 2 \Sigma p_i p_{i+1}(1-\Delta^{(s)})]$ which is minimised when $(\Sigma p_i)^2 = \left(\frac{2-n}{n-1}\right) \Sigma p_i p_{i+1}(1-\Delta_i^{(2)})$. A good selection of P will have this condition satisfied.

A sensible rule suggested by the analysis in Theorems 2 and 3, appears to be take $p_j = -\frac{1}{1-\mu_j}$ if μ_j is real or $p_j = \frac{-1 \text{ (sign } a_j)}{\sqrt{a_j^2 + b_j^2 - 1}}$ if $\mu_j = a_j + ib_j$ is complex assuming the positions of the eigenvalues are known for A. This clearly works well in special cases and for some nondiagonal P.

Even if we could find an optimal P it would not be unique. Thus,

Corollary 5:

Although the method of different relaxation factors for each equation separately in JOR and SOR is quite general and always enables convergence for an appropriate P there are always different ways of choosing the weights so that convergence is just as good. #

Lemma 4:

For P diagonal, $\exists$ no unique P giving some particular $\rho(I + PA)$. Consequently an optimal P if it exists is not unique.

Proof:

Let the characteristic equiation of A be $\lambda^n + b_{n-1} \lambda^{n-1} + ... + b_o = 0$, with roots λ_i $i = 1 .... n$. Then

$$b_{n-1} = \sum_{1}^{n} \lambda_i = \sum_{1}^{n} a_{ii} = n; \quad b_{n-2} = \sum_{j>i=1}^{n} \lambda_i \lambda_j = \sum_{1}^{n} \Delta_i^{(2)}; \quad \ldots.$$

$$b_{n-r} = \sum_{r>\ldots>i=1}^{n} \lambda_i \lambda_j \ldots. \lambda_r = \sum_{1}^{n} \Delta_i^{(r)}; \ \ldots. \ ; \quad b_o = \prod_{i=1}^{n} \lambda_i = \det A,$$

Where $\Delta_i^{(j)}$ is i^{th} principle minor determinant of order j of A.

Therefore for PA we have $\lambda^n + b^*_{n-1} \lambda^{n-1} + \ldots. + b^*_o = 0$ and

$$b^*_{n-1} = \sum_{1}^{n} p_i; \quad b^*_{n-2} = \Sigma p_i p_{i+1} \Delta_i^{(2)}; \quad \ldots. \ b^*_{n-r} = \Sigma p_i p_{i+1} \ldots. p_{i+r} \Delta_i^{(r)}$$

and $b^*_o = \det A \prod_{1}^{n} p_i$. If we vary each p_i appropriately we can create every n^{th} order polynomial and hence every eigenvalue. If we restrict ourselves to the case where $|\lambda_{max}|$ remains fixed we may still vary the p_i by selecting various other values for the roots and solving the n relations for n p_i values, to create every polynomial whose remaining roots $|\lambda_j| \leq |\lambda_{max}|$. #

Corollary 6:

Lemma 4 also applies to SOR with different relaxation factors for each equation.

Proof:

Since it is true for the JOR case, recursive acceleration, which produces eigenvalues in a fixed but probably unknown relationship to those of the equivalent JOR, will not affect the result. #

This corollary agrees with Young's example for the special case result on MSOR (modified SOR where B is 2-cyclic and a separate relaxation factor is used for each block).

10.2.3 The use of GRF

Had we taken A to have a non-unit diagonal the Jacobi iteration matrix would have been $D^{-1}B$ where D = diag (a_{ii}) and B = D - A. Thus GRF is strictly only a scaling process with a standard solution algorithm and therefore may be regarded as using some new normalisation, for each equation, taken from a range of possibilities no longer restricted to the n! available from A itself.

There is then a compensation to the hard work and difficulty of designing a "good" P. From the earlier conjecture it appears likely that a well chosen P may well imply a faster iteration process than is possible with optimal JOR or SOR (and suitable reordering). In difficult cases then a careful construction of P may be beneficial despite the extra computation. But generally if the simpler optimal JOR or SOR has a reasonable spectral radius then there is little to be gained anyway - see figure 8.2 - due to the logarithmic relationship between convergence rate and spectral radius. In the empirical results, optimal SOR with simple reordering has almost always achieved most of the acceleration possible and it seems likely that the usefulness of GRF would be the exception rather than the rule if SOR was available. This conjecture is not at variance with, say, theorem 8.33 of Young - that MSOR with B 2-cyclic can be optimised with the same relaxation factor for each block of equations. It is shown below that this result is a consequence of group SOR iterative methods (section 6.3), i.e. refers to the relative acceleration of the blocks. However if each equation were allowed differing relaxation factors the result would not hold. This remark carries over also to semi-iterative methods. If there are separate blocks of equations their independent optimal acceleration should be superimposed on that of individual within block equations.

10.2.3.2

One case where the GRF method can be useful is when $\rho(B) > 1$ and the elements causing this may be localised (after any permutations) say to the k^{th} row. Suppose we intend to use JOR. If this row causes a large $\rho(B)$ such that a suitable α is so small as to imply $\rho(B_\alpha) \simeq 1$ or if it causes a slowly converging optimal JOR, whereas without the k^{th} row there is considerable potential for accelerating JOR, then it would be desirable to scale B. Such a situation is very common in econometric models especially those with a few nonlinearities. For example where there are a few products

or ratios of variables the partial derivatives when evaluated may be quite large. One solution is to use optimal JOR for B less the k^{th} row. A considerably smaller relaxation factor may be used on the k^{th} equation. Let α^{**} be the resulting diagonal matrix of relaxation factors. Let α^{*} be the diagonal matrix with the scalar optimal relaxation factor for B, ignoring row k, in each position. Thus

$$\alpha^{**} = \alpha^{*}(I + \varepsilon_k q) \qquad \text{where } q = \begin{pmatrix} 0 & & & & 0 \\ & \ddots & & & \\ & & 0 & & \\ & & & \ddots & \\ 0 & & & & 0 \end{pmatrix} \; ; \; q_{kk} = 1$$

and $\varepsilon_k = \alpha^{**}_k - \alpha^{*}_k$. Then we have the GRF method with $P = -\alpha^{**}$ (recall P nonsingular so $\varepsilon_k \neq 0$) and

$$B_{\alpha^{**}} = \alpha^{*}B + \varepsilon_k qB + I - \alpha^{*} - \varepsilon_k q$$

$$= B_{\alpha^{*}} + \varepsilon_k \hat{B} \qquad \text{where } \hat{B} = \begin{bmatrix} & 0 & \\ b_{k1} \ldots, & b_{kk}-1, & \ldots b_{nk} \\ & 0 & \end{bmatrix} \qquad (9)$$

We may use the ideas of the previous chapter to bound $\rho(B_{\alpha^{**}})$ and hence select ε_k so that $\rho(B_{\alpha^{**}})$ is close to that of optimal JOR ignoring the k^{th} equation. We renormalise $\hat{B}$ so $\gamma = \max_i\{|b_{ki}|,1\}$ - since $b_{kk} = 0$ and use the form (9.13), and replacing ε_k by $\varepsilon_k\gamma$ in (9), the absolute difference in eigenvalues is

$$\frac{||\hat{B}||_2^2 \, \varepsilon_k^2 \, \gamma^2}{|S_1 \quad S_2|} \qquad (10)$$

where S_i is defined as in (9.7) for the matrix of eigenvectors of B.

Lemma 5:

If only one element of P, $p_k \neq p_j \quad k \neq j = 1 \ldots n$ then $\rho(I + PA) = \rho(I + p_j A)$ and p_j is chosen optimally by Corollary 1.

Proof:

By an application of Ostrowski's Theorem (1960) quoted at (9.4) to the above case (9), to obtain a bound on $\rho(I + PA)$ similar to (10). #

When the "problem" elements of B cannot be localised into one row then the bound (10) must be used to determine a suitable ε_k. It might prove rather laborious to use (10) in some cases. A simple way of selecting values $\varepsilon_k \quad k = 1 \ldots \ell$, where ℓ is the number of nonzero rows in $\hat{B}$, is by means of Gerschgorin's theorem. Using both row and column sum

$$\rho(I+PA) \leq \min\{\max_i |1+p_i| + |p_i| \sum_{j\neq 1} |b_{ij}|; \quad \max_j |1+p_j| + \sum_{i\neq j} |p_i b_{ij}|\}.$$

$\therefore \quad \rho(I+PA) < 1$ if $|p_j| > \sum_{j\neq j} |p_i| \, |b_{ij}|$ or $\sum_{j\neq 1} |b_{ij}| < 1$ (diagonal dominance).

Hence we attempt to find ε_k such that $|\alpha^* + \varepsilon_k| > \sum_{i\neq k} |\alpha^* + \varepsilon_i| \, |b_{ik}|$ for $k = 1 \ldots \ell$ and $k \neq 1 \ldots n$. $\varepsilon_k \leq 0$ is likely. By definition of the problem diagonal dominance certainly does not exist for the index k.

10.3 Nonstationary methods: Aitken's doubel Seidel process

We have briefly introduced two nonstationary methods in Chapter 9.6 [methods (ii) and (vi)]. They were both periodic nonstationary methods. We define:

A periodic nonstationary method is an iterative method whose iteration matrix varies with the iteration number but is repeated in a set sequence of h iterations. h is called the sweep length. If $h \to \infty$ we have a (nonperiodic) nonstationary method.

We consider a well known periodic nonstationary method before considering nonstationary methods properly.

The Aitken double Seidel process attempts to reduce the effect of the ordering of equations on SOR and also to stablise the iterations in the sense

of constant step lengths and signs. The procedure is to iterate in two parts: (a) as SOR; (b) as SOR but taking the equations in reverse order. It is clearly possible to use different relaxation factors in each iteration part. The procedure has also been called SSOR (symmetric SOR) by Young when the relaxation factor is constant and USSOR (unsymmetric SOR) when different relaxation factors are used. Thus h = 2. Raising h is possible but will be dealt with later (Section 5).

The iteration matrix of SSOR is given by

$S_\alpha = (I - \alpha U)^{-1} (\alpha L + (1-\alpha)I) (I - \alpha L)^{-1} (\alpha U + (I-\alpha)I)$ and for USSOR

$S_{\alpha_1\alpha_2} = (I - \alpha_2 U)^{-1} (\alpha_2 L + (1-\alpha_2)I) (I - \alpha_1 L)^{-1} (\alpha_1 U + (1-\alpha_1)I)$.

Theorem 4:

SSOR converges if $0 < \alpha < 2$. On the other hand USSOR converges if $0 < \alpha_1, \alpha_2 < 2$.

The proof relies on Theorem 9.1.1 of Young: a periodic nonstationary method converges iff each component method of one sweep converges. Taking each component of SSOR or USSOR as an SOR method and applying Theorem 8.9 to each component we get Theorem 4. #

Further we have

Theorem 5:

USSOR and SSOR are convergent iff all eigenvalues of B have real part less than unity.

Proof:

By Theorem 9.1.1 Young (1971), USSOR and SSOR converge iff component parts are convergent. For SSOR the component parts are $(I - \alpha U)^{-1}(\alpha L + (1 - \alpha)I)$ and $(I - \alpha L)^{-1} (\alpha U + (1 - \alpha)I)$ which are both SOR iteration matrices for different orderings. We require

$$\rho \begin{pmatrix} (I-\alpha U)^{-1}(\alpha L+(1-\alpha)I) & \vdots & 0 \\ \cdots\cdots\cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ 0 & \vdots & (I-\alpha L)^{-1}(\alpha U-(I-\alpha)I) \end{pmatrix} < 1.$$

By corollary 3 to Theorem 8.10 SOR iteration matrices are convergent whatever ordering for some α if all eigenvalues have real part less than unity. The same argument applies to USSOR having component parts both SOR iteration matrices $(I-\alpha_2 U)^{-1}(\alpha_2 L+(1-\alpha_2)I)$ and $(I-\alpha_1 L)^{-1}(\alpha_1 U+(1-\alpha_1)I)$ are both being convergent for some α_1 and α_2. #

Corollary 7:

The upper permitted limit of α for SSOR is the minimum of the upper limits for the two components. For USSOR the upper limits of α_1 and α_2 may be used in the normal way on each component. #

The problems of speed of convergence and obtaining α^*, or α_1^* and α^*_2, as optimal parameters are just as difficult as for SOR itself. The case of B p-cyclic has been studied by Young. Generalisation does not appear possible except in so far as the results for SOR may be carried over. The computations will be much more laborious. However in view of Theorem A1 we have $\min_\alpha \rho(S_\alpha) \leq \min_\alpha [\rho(R_\alpha), \rho(\hat{R}_\alpha)]$; $\hat{R}_\alpha = (I-\alpha U)^{-1}(\alpha L+(1-\alpha)I)$. Some "good" α values suggest themselves. Although α^* to minimise $\rho(\hat{R}_\alpha)$ is no easier to calculate than that for R_α, if we had them both for SSOR we want to use the α which provides the lower of the two optimal spectral radii calculated separately. For USSOR we could take α_1^* and α_2^* calculated separately and use them both. On the basis of the bound given in theorem A.1, it is not possible to provide comparison theorems on the convergence rates of SSOR and USSOR. In either case we have improved on SOR if $\rho(S_{\alpha^*})$ or $\rho(S_{\alpha_1^* \alpha_2^*})<[\rho(R_{\alpha_o})]^2$ where α_o is optimal for SOR in terms of asymptotic rates of convergence.

It seems unlikely we will be in a position to undertake such a test. For SSOR if $\alpha^* = \alpha_o$ (i.e. the original ordering component) then the test seems most unlikely to be satisfied. Similarly for USSOR if these "good" rules are followed. Of course, if B is symmetric then $\rho(R_{\alpha_1}) = \rho(\hat{R}_{\alpha_1})$ and so $\alpha_1^* = \alpha_2^*$; USSOR becomes identical to SSOR but no improvement is possible over SOR.

10.4 Nonstationary Methods: Alternating Direction Implicit (ADI) Methods.

ADI methods have been most extensively studied by Young, but these results apply only to some special cases of A. It appears results for convergence conditions and optimal parameters are not available for a general matrix A. ADI methods may be considered as periodic nonstationary or nonstationary depending on the use of parameters.

We split $A = A_o + V_o + \theta$ where θ is nonnegative and diagonal. Then the iteration is in two parts

$$\begin{aligned} [H_o + \delta_1\theta + \alpha_{s+1}I]Y^{(s+\frac{1}{2})} &= b - [V_o + (1 - \delta_1) - \alpha_{s+1}I]Y^{(s)} \\ [V_o + \delta_2\theta + \alpha'_{s+1}I]Y^{(s+1)} &= b - [H_o + (1 - \delta_2)\theta - \alpha'_{s+1}I]Y^{(s+\frac{1}{2})} \end{aligned} \tag{11}$$

for parameters δ_1, δ_2 α_{s+1} α'_{s+1} $s = 1 \ldots\ldots$

(11) specifies the Peaceman-Rachford method. Following Young, if θ, is chosen to have constant elements we may without loss of generality take $\delta_1 = \delta_2 = \frac{1}{2}$. Let $H = H_o + \frac{1}{2}\theta$, $V = V_o + \frac{1}{2}\theta$ then the iteration matrix for Peaceman-Rachford is

$$T_{\alpha_s\alpha'_s} = [V + \alpha'_s I]^{-1}[H - \alpha'_s I][H + \alpha_s I]^{-1}[V - \alpha_s I] \tag{12}$$

If $\alpha_s = \alpha$ and $\alpha'_s = \alpha'$ or if $\alpha_s = \alpha'_s = \alpha$ we have a periodic nonstationary method h = 2 with iteration matrices $T_{\alpha\alpha'}$, and T respectively. By theorem 17.2.1 (Young (1971)) $\rho(T_{\alpha\alpha'}) < 1$ iff $(H + V)^{-1}(HV+\frac{1}{2}(\alpha-\alpha')(V-H)+\alpha\alpha'I)$ is N-stable. Similarly by Young's theorem 17.2.2, $\rho(T_\alpha) < 1$ iff $(H+V)^{-1}(HV+\alpha^2 I)$ is N-stable. Therefore in addition to Young's results we conclude convergence in either case is most easily tested by the Meyer-Spasche test.

The natural question is what values of α_s, α'_s are permissible, and optimal and how should H,V be selected? The problem of parameters is

studied by Young and also summarised by Westlake (1968). Optimal parameters have not been obtained but a "good" scheme has been derived for the case HV = VH. In fact two possibilities are suggested - the Peaceman-Rachford parameters and the Wachspress parameters. Both sets make (11) a periodic nonstationary method in which the sweep length h depends on the eigenvalues of H and V. Further it is shown by Young that if HV = VH then with either set of parameters (11) is convergent. Numerical evidence shows Wachspress parameters to be superior to Peaceman-Rachford and Young claims to have found them generally nearly as good as optimal ones except when h is small. Indeed it is known for Wachspress parameters (Young Chapter 17, Westlake),

$$\rho\left(\prod_{i=1}^{h-1} T_{\alpha_i\alpha_i'}\right) \leq \left|\frac{1 - C^{\frac{1}{2}(h-1)}}{1 + C^{\frac{1}{2}(h-1)}}\right|^2 \qquad (13)$$

where $C = \lambda\min/\lambda\max$ when H, V have eigenvalues λ_i μ_i respectively and $\lambda_{min} \leq \lambda_i, \mu_i \leq \lambda_{max}$ $i = 1 \ldots n$. Thus, although in the absence of precise theoretical results we can recommend the Waschspress parameters, the success of iteration (11) depends on the choice of H, V. With these parameters and comparing it with any other stationary method (JOR or SOR etc.) whose spectral radius is known to be ρ we have from (13) $\left[\frac{1-C^{\frac{1}{2}(h-1)}}{1+C^{\frac{1}{2}(h-1)}}\right]^2 \geq \rho^2$ since ADI methods are two part iterations. Hence (11) is superior if

$$1 \geq (\rho + 1)C^{\frac{1}{2(h-1)}} + \rho. \qquad (14)$$

(14) holds only for $C^{\frac{1}{2(h-1)}}$ small enough, i.e. $|C| < 1$ at least, and since $h \geq 2$ this implies the most favourable case is h = 2. If h = 2 (14) becomes

$$|C| \leq \left[\frac{1 - \rho}{1 + \rho}\right]^2 \leq 1 \qquad (15)$$

Noting $0 \lesssim \rho < 1$ assuming the alternative method convergent, this may always be tested in a given case.

Further, from (13), we note convergence is assured if $2 > C^{\frac{1}{2(h-1)}} > 0$[1]. The last two statements suggest rules for choosing H and V. They should be chosen so that the spread of their eigenvalues is approximately the same on each side of the origin for the maximum convergence rate by (13). But (13) and (15) present conflicting rules. By (15) it is unlikely in this case ADI methods are competitive. However in other cases, particuarly when H and V are positive or negative definite, ADI methods can be made to be considerably faster than alternatives by (15); although in absolute terms they may not be very fast. Empirical investigations relate mostly to the special case where H and V may be taken positive definite. For example work by Birkhoff, Varga and Young (1962) reported by Westlake showed a superiority over SOR that is quite marked. The main conclusion here is that superiority cannot be expected to carry over to the general case. It may not be possible to find H and V from A such that ADI methods are superior to, say, SOR. However it must be emphasised that these conclusions are based on a sufficient condition for superiority [by (15) and (13]. Failure of this condition does not necessarily mean inferiority although that is the conjecture. Superiority cannot in these circumstances be established whereas it can if (15) holds. Further analysis is not undertaken as everything depends on the particular H and V matrices. A more precise analysis and a resolution of the conflict of (13) and (15) waits upon an exact relation for the spectral radius to replace the inequality (13). The method is not recommended for the general case owing to the difficulty of constructing actual matrices H and V with certain desirable eigenvalues - or even of knowing such H and V matrices are available. All the results on convergence and the analysis following (13) depend on commutivity in H and V. Without that not even convergence can be

---

1. The suggestion that even values of h should be taken is automatically satisfied by the two part iteration. The actual value of h is set by Wachspress parameters.

proved and we must fall back on the Meyer Spasche test for N-stability of the matrices.

Another ADI method, the Douglas-Rachford method, is very similar to Peaceman-Rachford. A similar analysis could be applied since its spectral radius can be represented similarly to (13) (see Westlake).

A suggestion which is promoted by considering the SSOR, USSOR methods is to choose $H_o = U$ and $V_o = L$. However this would give $C \simeq 1$ and so, although convergent, it is unlikely to improve on SOR, SSOR, USSOR, etc.

10.5 Nonstionary Methods: First Order Methods

All the methods analysed so far for solving (8.5) have been linear first order methods and they all have used some parameter or relaxation factor which must be set. The most natural generalisation is to allow that parameter to vary from iteration to iteration - thus achieving a periodic non-stationary or, if $h \to \infty$, a nonstationary version of a stationary method already considered, e.g. of the form of Jacobi

$$Y^{(s)} = p_s(B)Y^{(0)} \quad \text{where } p_s(\cdot) \text{ is a } S^{th} \text{ order polynomial.} \tag{16}$$

If the solution to (8.5) is Y^* and we write the error $e^{(s)} = Y^{(s)} - Y^*$ then

$$||e^{(s)}|| \leq ||p_s(B)||\ ||e^{(0)}||. \tag{17}$$

We need to define new measures for the speed of convergence. Young offers six possibilities but to define the average rate of convergence over S iterations as $R(p_s(B)) = -\frac{1}{S} \log ||p_s(B)||$ is sufficient since $\rho(p_s(B))$ will also be available, and from which other measures may be obtained. The connection to the stationary case is obvious since then $p_s(B) = B^s$; in the following we fix $S = h$. If convergence is not obtained after S steps we repeat with another sweep; or we may let $S \to \infty$.

The first and most general representation of a first order nonstationary method was that of Richardson. We consider

$$Y^{(s+1)} = Y^{(s)} + \alpha_s (AY^{(s-1)} - b) \tag{18}$$

which is obviously similar to (1). The scalar α_s is dependent on S. Therefore

$$e^{(s)} = (I + \alpha_s A)e^{(s-1)} = \prod_{i=1}^{S} (I + \alpha_i A)e^{(0)} \tag{19}$$

for an arbitrary start $Y^{(0)}$. Therefore in the notation of Section 2, $p_s(G) = \prod_{i=1}^{S} (I + \alpha_i A)$ and (18) is convergent for an arbitrary start iff $\lim_{S\to\infty} p_s(G) = 0$. In view of Lemma 1 it might appear there was no advantage to such a scheme since the spectral radius of $p_1(G)$ can be minimised with an optimal choice of α^* and therefore the spectral radius of $\prod_{1}^{S} (I + \alpha_i A)$ will also be minimised with the same choice. However the possibilities of improving average rates of convergence and easier calculation of optimal α_i than α^* must not be ignored.

Note if $A = 0$, $p_s(1) = 1$. From (17) and (19) we must choose α_i to maximise the speed of convergence, i.e. choose α_i to achieve $\min_{p_s(1)=1} ||p_s(G)||$. But

$$||p_s(G)|| \geq \rho(p_s(G)) = \max_i |p_s(\lambda_i)|, \tag{20}$$

where λ_i are eigenvalues of G (Theorem 5.1 Varga (1962)).

If $a \leq \lambda_i \leq b \quad i = 1 \ldots n$ are real[1] then

$||p_s(G)|| = \max_{a\leq x\leq b} |p_s(x)| \geq \max_i |p_s(\lambda_i)|$ so we require α_i to achieve

$$\min_{p_s(1)=1} \left\{ \max_{a\leq x\leq b} |p_s(x)| \right\} \tag{21}$$

---

1. For the remainder of this chapter, b refers to the upper bound of these eigenvalues, not to the notation of (8.5)

Since $p_s(x)$ is real the solution is well known as Chebychev's polynomial

$$C_s(x) = \begin{cases} \cos\ (s\ \mathrm{Cos}^{-1}\ x) & -1 \leqslant x \leqslant 1 \qquad S \geq 0 \\ \cosh(s\ \cosh^{-1}\ x) & |x| \geqslant 1. \end{cases} \tag{22}$$

Then $C_o(x) = 1$, $C_1(x) = x$ and $C_{s+1}(x) = 2xC_s(K) - C_{s-1}(x)$ $S \geqslant 1$; and $\max\limits_{-1\leqslant x\leqslant 1} |C_s(x)| = 1$. We will need the following theorem:

Theorem 6:

For any real $z > 1$ and fixed integer $S \geqslant 1$, if $p_s(x) = C_s(x)/C_s(z)$, $C_s(\cdot)$ defined as above, then

(a) $p_s(z) = 1$

(b) $\max\limits_{-1\leqslant x\leqslant 1} |p_s(x)| = \dfrac{1}{C_s(z)}$

(c) if $q(x)$ is any polynomial of degree S such that $q(z) = 1$ and $\max\limits_{-1\leqslant x\leqslant 1} |q(x)| \leqslant \max\limits_{-1\leqslant x\leqslant 1} |p_s(x)|$ then $q(x) \equiv p_s(x)$.

Proof:

Theorem 9.3.1 Young or 5.1 Varga. #

Corollary 8:

$$\max_{-1\leqslant x\leqslant 1} |p_s(x)| < 1 \quad \text{if} \quad z > 1.$$

Proof:

From (b) of Theorem 6 and (22) since $\cosh^{-1}z > 0$ and $\cosh z > 1$. By (22) if $|z| < 1$ is maximum > 1 when $x = \cos(k\pi/n)$, $k = 0 \ldots n$. #

For the minimisation (21) consider therefore

$$p_s(x) = \frac{C_s\left[\dfrac{2x-(b+a)}{b-a}\right]}{C_s\left[\dfrac{2-(b+a)}{b-a}\right]} \tag{23}$$

which implies $p_s(1) = 1$ and $z = \dfrac{2-(b+a)}{b-a} > 1$ provided $b < 1$. Since $p_s(1) = 1$ and $z > 1$ if $b < 1$, we can apply Theorem 6 and its corollary directly to show

Richardson's method is convergent where α_i are chosen to yield $p_s(G)$ as defined by (23). Additionally these parameters are unique and optimal in view of part (c) of that theorem. Evidently if $q_s(y) = p_s\left(\frac{(b-a)y + b + a}{2}\right)$ then $p_s(x) = q_s\left(\frac{2x - (b+a)}{b - a}\right) = q_s(y)$, so

$$\max_{a \leqslant x \leqslant b} |p_s(x)| = \max_{-1 \leqslant y \leqslant 1} |q_s(y)| \qquad (24)$$

Application of Theorem 6 confirms (23) as the optimal choice to minimise (24) and hence by (20), (21) provide optimal convergence. Notice that y provides a 1-1 mapping of the real interval $a \leqslant x \leqslant b$ into $-1 \leqslant y \leqslant 1$. By the inequality of (21) we need to choose b as the maximum eigenvalue and a as the minimum eigenvalue of G. Thus the average rate of convergence is maxmised for any S which also minimises the upper bound on the spectral radius if $S \to \infty$, and the 1-1 mapping into [-1, 1] is preserved. Finally $C_s(z)$ is a constant in (23), for any fixed z, which adjusts $C_s(x)$ to fit $p_s(x)$ to the point $p_s(1) = 1$.

From (23) we may use $p_s(G) = C_s\left(\frac{2G}{b-a} - \left(\frac{b+a}{b-a}\right) I\right) / C_s(z)$ by (21) and from the recurrence relation in Chebychev polynomials deduce the optimal parameters α_i $i = 0, \ldots S - 1$ for any S value - as in Westlake.[1]

Two generalisations of this analysis are necessary. Firstly for the case some or all λ_i are complex.

Lemma 6:

Theorem 6 holds unaltered if x is a complex variable; other definitions remaining as in Theorem 6.

Proof:

(a) Obvious.

(b) From the definitions of the functions cosh x and sinh x and the Euler formulae, it follows if x = c + id then cos x = cos c cosh d + i sin c sinh d, and cos (c + id) = cosh (d-ic). The Chebychev polynomial of an arbitrary

---

1. Most authors work with a, b in terms of the eigenvalues of A instead of those of G. The translation may easily be made.

complex number may be written either way in (22) - we take the cosine function. Hence $|\cos(c + id)| < 1$ if $\cos^2 c \cosh^2 d + \sin^2 c \sinh^2 d < 1$ i.e. if $\sinh^2 c < \sin^2 d$. This last condition is satisfied when $(e^d - e^{-d})^2 < -(e^{ic} - e^{-ic})^2$ which implies $\cosh 2d + \cos 2c < 2$. Moreover $\cos(c + id) = 0$ iff simultaneously $\cos c \cosh d = \sin c \sinh d = 0$ which follows iff $d = 0$ and $c = k\pi/2$ for $k = \ldots -1, 1, 3, \ldots$. Thus $C_s(c + id)$ achieves its maximum whenever $x = \cos\left(\frac{k(c'+id')}{S}\right)$ and its minimum whenever $x = \cos\left(\frac{2k-1}{2S}\right)\pi$ and $d = 0$, (where $\cosh 2d + \cos 2c' = 2$) for $k = 1 \ldots S$. The zeroes of $C_s(\cdot)$ are naturally the same as the real case. But we require $\max_{c,d} |C_s(x)| = 1$ to prove part (b) and the corollary 8. For any $c + id$ in the closed set $\{|c + id| \leq 1\}$; $x = \cos\left(\frac{c+id}{S}\right) \Rightarrow |C_s(x)| \leq 1$. Thus part (b) follows iff a, b are chosen to uniquely map the closed complex set $\{x = c'+id' \mid a \leq x \leq b\}$ into the unit circle $\{|y| = |c + id| \leq 1\}$. Rules for this are shown in Lemma 7 below.

(c) The uniqueness and optimality of $p_s(x)$ is proved identically as for Theorem 6. That is, if $R(x) = Q(x) - p_s(x)$ and allowing x to range over all points of the unit circle we can construct at least n distinct zeroes of $R(x)$ at points along the real axis in the interval $[-1,1]$. Since also $R(z) = 0$ we have n + 1 zeroes and hence $R(x) \equiv 0$. (Theorem 9.3.1 Young) for $z > 1$. #

Lemma 7:

Let G have eigenvalues $\lambda_j = c_j + id_j$. Then $y_j = \frac{2\lambda_j - e}{f}$ with $e = \frac{1}{n}\Sigma c_j$ and $f = \max_j \{4(|\lambda_j|^2 - c_j e) + e^2\}^{\frac{1}{2}}$ provides a 1-1 mapping of the ellipse $\max_j \left\{ \frac{c_j^2}{a^2} + \frac{d_j^2}{b^2} \right\} = 1$ into the unit circle. Hence we can use (23) and Theorem 6 unchanged with $b = \frac{e + f}{2}$ and $a = \frac{e - f}{2}$ if $\frac{e + f}{2} < 1$.

Proof:

Complex roots in conjugate pairs $\Rightarrow$ the ellipse is symmetric about the real axis. Substracting e from each real part makes it symmetric about the

imaginary axis. Then $|y| < 1$ if $(2c_j - e)^2 + 2d_j^2 < f^2$

i.e. $4(|\lambda_j|^2 - c_j e) + e^2 < f^2$ for $j = 1 \ldots n$.

The use of (23) to which Theorem 6 applies, is then appropriate by Lemma 6, provided $z > 1$ i.e. $b = \frac{e + f}{2} < 1$. Optimality in convergence rates using this a and b follows from Lemma 6 part (b). #

One of the benefits of Richardson's method by Theorem 6 was that only the two extreme eigenvalues of G are needed. In Lemma 7 a knowledge of all is necessary.

Corollary 9:

A sufficient rule to preserve this mapping into the unit circle is to choose $b = \rho(G)$ and $a = -b$ which may be easily verified. But the condition for $z > 1$ is now $\rho(G) < 1$ which is more stringent than $\frac{e + f}{2} < 1$. #

We now provide the second generalisation - the case where $z < 1$. Richardson's method converges here too, so it is quite adequate in practice to use $b = \rho(G)$ and $a = -b$.

Lemma 8:[1]

Theorem 6 holds unchanged if $z < 1$; other definitions the same.

Proof:

Nonsingularity of I - G is required for a unique solution to (8.5). Hence $\lambda_i \neq 1$,[2] $i = 1 \ldots n$. Assume some λ_i lie with and some without the unit complex circle Then $a' \leqslant b' \leqslant 1$ can always be chosen such that every λ_i is in one of the two complex circles, centre $\left(\frac{a' + b'}{2}, 0\right)$ radius $\frac{b' - a'}{2}$ or centre $\left(\frac{4 - (a' + b')}{2}, 0\right)$ radius $\frac{b' - a'}{2}$. Assume further that a' and b' are chosen so that the two extreme eigenvalues lie on the boundaries of the two circles. The matrix $G(2I - G)$ has eigenvalues $\lambda_i(2 - \lambda_i)$ therefore in

---

1. Corollary 8 is only a sufficient condition for convergence as shown by the general convergence theorem 12.1 (also theorem 8.1) and (20), (21) above.

2. Having $\lambda_i = 1$ only implies $\rho(G(2I - G)) = 1$ below which is weakly convergent but as usual we avoid this complication here.

and on the boundary of the complex circle,[1] centre $\left(\frac{a'(2 - a') + b'(2 - b')}{2} = a' + b' - \frac{a'^2 + b'^2}{2}, 0\right)$ and radius $a'-b'+\frac{b'^2-a'^2}{2}$. This radius is less than unity if

$$(b' - 1)^2 < (a' - 1)^2 + 2 \qquad (25)$$

which by construction of $a' \leq b' < 1$ always holds. Let $2e = a'(2-a')+b'(2-b')$ and $f = a' - b' + \frac{b'^2 - a'^2}{2}$. Then $y_j = \frac{2\lambda_j - e}{f}$ provides a 1-1 mapping of the complex circle containing the roots of $G(2I - G)$ to a complex circle contained within the unit circle. A sufficient, but nonoptimal, rule guaranteeing convergence of Richardson's method with the iteration matrix $G(2I - G)$ is therefore given in Lemma 7. Suppose that method uses iteration parameters γ_i $i = 1 \ldots S$ for (19) and we define θ_j by

$$I + \sum_{i=1}^{S} \gamma_i [G(2I - G)]^i = I + \sum_{j=1}^{2S} \theta_j G^j \qquad (26).$$

The sequence on the left of (26) is convergent and so therefore is that on the right. One inefficient (nonoptimal) Richardson method on G, guaranteeing convergence, is to set $\alpha_i = \gamma_i + \sqrt{\gamma_i(\gamma_i + 1)}$ and $\alpha_{i+1} = \gamma_i - \sqrt{\gamma_i(\gamma_i+1)}$ $i=1....S$. The optimal parameters for Richardson's method on G imply convergence at least as fast as that, requiring, as normal, a and b of Lemma 7 where λ_j are the roots of G. $C_s(x)$ is that optimal polynomial since the uniqueness of $p_s(x)$ with $\min_{p_s(1)=1} \left\{ \max_{a \leq x \leq b} |p_s(x)| \right\}$ is unaffected by the value of Z (from Lemma 6 part (c)) and $C_s(z)$ is merely a normalising constant. #

We conclude therefore in summary:

Theorem 7:

Richardson's method is always convergent for any real A and optimal

---

1. See also Young (1971).

according to criterion (21).

Proof:

Theorem 6 with Lemmas 6, 7 and 8. #

Theorem 7 opens the way for comparison theorems with stationary methods. Such theorems are not attempted as there appears to be some difficulties in using Richardson's method in practice. However the above analysis will be useful in proving an equivalent to Theorem 7 for semi-iterative methods.

From numerical experiments Richardson's method is very sensitive to rounding errors and the sensitivity increases with increasing S (Young (1971)). We can get just as good theoretical properties by semi-iterative methods - indeed we can even design a semi-iterative method to produce the same result as Richardson's method - with no problems of sensitivity to rounding errors. A theoretical difficulty is there is no criterion for the best value of S = h. The performance is affected by our choice but apart from the larger n the larger potentially is S, we have no further knowledge. This problem is overcome in semi-iterative methods which retains the benefits of requiring only the extreme eigenvalues of G and having optimal average rates of convergence.

We note if $\alpha_i = -\alpha$ $i = 1 \ldots$. we obtain the JOR method, by Lemma 1. Thus Richardson's method with $-\alpha_i$ is the simple nonstationary JOR method. If we recursively accelerate the JOR method we obtain the nonstationary SOR method. Theorem 7 would then prove general convergence of both methods and optimal parameters for nonstationary JOR - those for nonstationary SOR are not so easy to derive. Other nonstationary methods are possible (e.g. for GRF). It is important to note that any of these methods mentioned is a special case of a first degree Chebychev acceleration if the parameters are obtained from Theorem 6. We have considered Richardson's method with underlying iteration matrix $I + \alpha_i A$. But we can apply this type of acceleration to any linear stationary method, e.g. to Jacobi, optimal JOR, Gauss-Seidel, SOR if the iteration matrix G is replaced by $\alpha_i B + (1 - \alpha_i)I$, $\alpha_i B_{\alpha^*} + (1 - \alpha_i)I$,

$(I - \alpha_i L)^{-1}(\alpha_i U + (1 - \alpha_i)I)$, $\alpha_i R_{\alpha^*} + (1 - \alpha_i)I$ respectively and a, b of (23) adjusted to the extreme eigenvalues of that matrix or as Lemma 7.

10.6 Semi-iterative Methods:

10.6.1 Convergence

The problems of the nonstationary first order Chebychev accelerations are overcome if it is replaced by a second order acceleration, with the only attendant disadvantage of a small increase in computer storage. As in the first order case, a second order Chebychev acceleration may be superimposed on any linear stationary method considered so far which may or may not have already been accelerated optimally by its relaxation parameter, recursively or by reordering and renormalisation. The general form of the second order acceleration on a basic first order method with iteration matrix G is

$$Y^{(s+1)} = \frac{\alpha_{s+1}}{2 - (b+a)}[2G - (b+a)I]Y^{(s)} + (1 - \alpha_{s+1})Y^{(s-1)} \qquad (27)$$

in place of (18).

The iteration is derived in detail by Young (equation 11.2.28); with a and b as in Richardson's method. (27) is defined as the semi-iterative method on matrix G. Clearly we can construct a semi-iterative version of Jacobi, optimal JOR, stationary Richardson, SOR by replacing G in (27) by B, B_{α^*}, I + pA, R_α and adjusting a and b to contain the extreme eigenvalues of the given matrix. However in a second order iteration as (28), the SOR version is ambiguous. Lemma 8.1 no longer applies. Semi-iterative SOR is taken as (28) with $G = R_\alpha$. The alternative is to recursively accelerate (27) with G = B, i.e.

$$Y^{(s+1)} = \left[I - \frac{2\alpha_{s+1}}{2-(b+a)} L\right]^{-1} \left\{\left[\frac{2\alpha_{s+1}}{2-(b+a)} U - (b+a)I\right]Y^{(s)} + (1-\alpha_{s+1})Y^{(s-1)}\right\} \qquad (28)$$

To determine the optimal value of α_{s+1} in (28) presents a problem since it is not even a nonstationary SOR version of (27). Further analysis is not undertaken.

The optimal parameters in (27) are easily obtained given b, a[1] and are given by Young as

$$\alpha_1 = 1,\ \alpha_2 = \frac{2Z^2}{2Z^2-1},\quad \alpha_{s+1} = [1 - \frac{1}{4Z^2}\alpha_s]^{-1};\quad S \geq 2 \tag{29}$$

for $Z = \frac{2 - (b+a)}{b - a}$. If $b = -a = \rho(G)$ then $Z = 1/\rho(G)$.

There are a number of different ways of deriving (27), for the general stationary method with matrix G, applied to (8.5) and some new interpretations are given below. Young's approach is to observe the convergence of the sequence $Y^{(s)}$ $S = 1, \ldots$ may be accelerated by the Cesaro summation $V^{(s)} = \sum_{k=0}^{S} \beta_{s,k} Y^{(k)}$ where $\sum_{k=0}^{S} \beta_{s,k} = 1$. It is well known that the sequence $V^{(s)}$ has the same convergence point. If the sequence $Y^{(s)}$ is generated by $Y^{(s)} = GY^{(s-1)}$ then in the notation of (17) $e^{(s)} = G^k e^{(0)}$. Let $\eta^{(s)} = V^{(s)} - Y^*$ then

$$\eta^{(s)} = \Sigma\beta_{s,k} Y^{(k)} - Y^* = \Sigma\beta_{s,k} Y^{(k)} - \Sigma\beta_{s,k} Y^* = \Sigma\beta_{s,k} e^{(0)}$$

$$\therefore \quad \eta^{(s)} = p_s(G)e^{(0)} = p_s(G)\eta^{(0)} \tag{30}$$

since $Y^{(0)} = V^{(0)}$ if $\beta_{0,0} = 1$, where $p_s(G) = \Sigma\beta_{s,k} G^k$.

Now corresponding to (17)

$$||\eta^{(s)}|| \leq ||p_s(G)||\ ||e^{(0)}|| \tag{31}$$

Thus the optimal values of $\beta_{s,k}$, $k = 1 \ldots S$ for fixed S, would be obtained

---

1. The determination of a,b may not be easy for R_α as noted in Chapter 8 but see below in Chapter 14.2.2.

by the inequalities and minimisation (20) and (21). Once again the solution is provided by Chebychev polynomials with

$$p_s(G) = \frac{C_s\left[\frac{2G-(b+a)I}{b-a}\right]}{C_s\left|\frac{2-(b+a)}{b-a}\right|} \qquad (32)$$

The fact that the minimisation of $\max_{a\leq x\leq b} |p_s(x)|$ with $p_s(x)$ as in (23) implies that (32) in (30) will yield the semi-iterative method on G; and that it may then be rewritten as (27), with parameters α_{s+1} in place of $\beta_{s,k}$, may be checked in Young Chapter 11. But we may now prove:

Theorem 8:

The semi-iterative method on G is convergent for all matrices G. The implied parameters $\grave{\alpha}_{s+1}$ are optimal in the sense of minimising $||p_s(G)||$ for any fixed S.

Proof:

By application of Theorem 6 to (31). If G has complex eigenvalues by Theorem 6 plus Lemmas 6 and 7. If G has eigenvalues such that Z < 1 by Theorem 6 and Lemmas 6, 7, and 8 amended as follows. Since (27) is a second order iteration, (26) may be replaced by

$$\sum_{k=0}^{S} \gamma_{s,k}\,[G(2I - G)]^k = \sum_{k=0}^{2S} \delta_{2s,k}\, G^k \qquad (33)$$

Since $\Sigma\gamma_{s,k} = 1$ then $\Sigma\delta_{2s,k} = 1$ and hence we could actually use (inefficiently)

$$\left.\begin{aligned} \delta_{s,k} &= \gamma_{s,k} \quad k = 0 \ldots S && S \text{ even} \\ &= \delta_{s-1,k} \quad k = 0 \ldots S-1 \quad \delta_{s,s} = 0 && S \text{ odd} \end{aligned}\right\} \text{ for each value of S.}$$

The proof is completed identically to Lemma 8 for the optimal polynomial. #

10.6.2 Comparison Theorems:

Theorem 8 parallels Theorem 7 and makes a semi-iterative method available in any circumstances. It also contains the suggestion that the average rate

of convergence is minimised for each S value. The asymptotic rate of convergence is obtained, as in the stationary case, as $\lim_{S\to\infty} R(p_s(G))$. Formally therefore we can state:

Theorem 9:

The semi-iterative method on G has the following convergence characteristics:

(a) the average rate of convergence in uniquely maximised among all normalised methods, (i.e. where $||p_s(I)|| = 1$) for every $S \geq 2$;

(b) α_{s+1} are monotonically decreasing with increasing S, and converge on $\lim_{S\to\infty} \alpha_s = \left\{1 + \left(\frac{1}{Z + (Z^2 - 1)^{\frac{1}{2}}}\right)^2\right\} (2 - (b+a))$

(c) it has precisely the same asymptotic rate as the optimal block SOR method for the iteration $V^{(k+1)} = \hat{G}\, V^{(k)}$ where $\hat{G} = \begin{bmatrix} G & O \\ O & G \end{bmatrix}$ and $V^{(k)} = \begin{pmatrix} Y^{(k)} \\ Y^{(k+\frac{1}{2})} \end{pmatrix}$

i.e. $\lim_{S\to\infty} R(p_s(G)) = \lim_{S\to\infty} R(\hat{G}_{w^*}) = -\log_e \left\{\frac{\rho(G)}{1+(1-(\rho(G))^2)}\right\}$ where in (32) $b = -a = \rho(G)$ is chosen.

Proof:

(a) The semi-iterative method on G is set up such that by (21) $\max_{a\leq x\leq b} |p_s(x)| = ||p_s(G)||$ is minimised by the choice of $p_s(G)$ in (32). Hence $p_s(1) = 1$. Therefore by Theorem 6 (and Lemmas 6, 7 and 8) $p_s(G)$ uniquely solves this minimisation problem and for any other polynomial $q_s(G) \neq p_s(G)$, $q_s(1) = 1$, $||p_s(G)|| < ||q_s(G)||$ or $R(p_s(G)) > R(q_s(G))$ for every $S \geq 2$. For $S = 0,\ 1$; $e^{(0)} = \eta$ and $e^{(1)} = \eta^{(1)} = Ge^{(0)}$ hold for Chebychev as well as other polynomials including stationary Jacobi methods.

(b) (27) may be written, using $b + a = 0$,

$$Y^{(s+1)} = w_{s+1} \{G\, Y^{(s)} - Y^{(s-1)}\} + Y^{(s-1)} \qquad (34)$$

---

1. note (a) it does not coincide with 2-cyclic SOR,
   (b) its speed of convergence may be increased beyond this if Lemma 7 is followed. We concentrate on this evaluation of a and b hereafter.

so $w_{s+1} = \dfrac{\alpha_{s+1}}{2-(b+a)}$.

The recurrence relation in Chebychev polynomials with (32) gives

$$C_{s+1}(z)\, p_{s+1}(G) = z2GC_s(z)p_s(G) - C_{s-1}(z)\, p_{s-1}(G) \tag{35}$$

Now using $\eta^{(s)} = p_s(G)\, e^{(0)}$ and writing $\eta^{(s)} = Y^{(s)} - Y^*$ we have the iteration implied by (35) as

$$c_{s+1}(z)Y^{(s+1)} = z2C_s(z)\,[GY^{(s)} - Y^{(s-1)}] + Y^{(s-1)} \tag{36}$$

Thus $w_{s+1} = \dfrac{z2C_s(z)}{c_{s+1}(z)} = 1 + \dfrac{c_{s-1}(z)}{c_{s+1}(z)}$. Now if $Z > 1$ from (22), let

$\sigma = \cosh^{-1}(z)$ and $w_{s+1} = 1 + \dfrac{(e^{(s-1)\sigma} + e^{-(s-1)\sigma})}{(e^{(s+1)\sigma} + e^{-(s+1)\sigma})} = 1 + e^{-2\sigma}\left[\dfrac{1+e^{-2(s-1)\sigma}}{1+e^{-2(s+1)\sigma}}\right]$ (37)

But if $Z \lesseqgtr 1$ let $\sigma = \cos^{-1}(z)$ and then

$$w_{s+1} = 1 + \frac{(e^{i(s-1)\sigma} + e^{-i(s-1)\sigma})}{(e^{i(s+1)\sigma} + e^{-i(s+1)\sigma})} = 1 + e^{-2i\sigma}\left[\frac{1+e^{-2i(s-1)\sigma}}{1+e^{-2i(s+1)\sigma}}\right] \tag{38}$$

Moreover when $Z > 1$ $\sigma = \log_e[Z + (Z^2-1)^{\frac{1}{2}}]$ by definition and when $Z \lesseqgtr 1$ since[1] for any real $\theta \lesseqgtr 1$, $\cos\theta = \cosh i\theta$ we have $\cosh^{-1}z = i\sigma$ so $i\sigma = \log_e[Z+(Z^2-1)^{\frac{1}{2}}]$.

Therefore

$$e^{-k\sigma} = \left[\frac{1}{Z+(Z^2-1)}\right]^{-k} \text{ when } Z > 1 \text{ and } e^{-ik\sigma} = \left[\frac{1}{Z+(Z^2-1)^{\frac{1}{2}}}\right]^{-k} \text{ if } Z \lesseqgtr 1, \tag{39}$$

Evidently in both (37) and (38) w_{s+1} is monotonically decreasing as $S\to\infty$ for fixed Z as the bracketed term is monotonically decreasing $\to 1$ as $S \to \infty$.

Substituting (39) in either (37) or (38) we therefore get $\lim_{S\to\infty} w_{s+1} = 1 + \left[\dfrac{1}{Z+(Z^2-1)^{\frac{1}{2}}}\right]^2$.

Comparison of α_{s+1} and w_{s+1} in (34) and (28) completes the proof of part (b).

---

1. We use Euler's formula $\cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$

(c) By Theorem 6 for Z where $p_S(x)$ is given by (32) $\max_{-1\leq x\leq 1} |p_S(z)| = \frac{1}{C_S(z)}$; so by (21) we have $||p_S(G)|| = \frac{1}{C_S(Z)}$ when $b = -a = \rho(G)$. Using (39) in the definition (22) of $C_S(Z)$ we get for $Z > 1$ or $Z \leq 1$.

$$C_S(Z)=\tfrac{1}{2}[(Z+(Z^2-1)^{\frac{1}{2}})^{-S}+(Z+(Z^2-1)^{\frac{1}{2}})^{S}] =\tfrac{1}{2}\ [Z+(Z^2-1)^{\frac{1}{2}}]^{S}[1+(Z+(Z^2-1)^{\frac{1}{2}})^{-2S}]$$

hence $R(p_S(G))=-\frac{1}{S}\{\log_e 2-S\ \log_e[Z+(Z^2-1)^{\frac{1}{2}}]-\log_e[1+(Z+(Z^2-1)^{\frac{1}{2}})^{-2S}]\}$

and $\lim_{S\to\infty} R(p_S(G))=-\ \log_e\{(Z+(Z^2-1)^{\frac{1}{2}})^{-1}\}$ where $Z = \frac{2-(b+a)}{b-a} = \frac{1}{\rho(G)}$

Thus $\lim_{S\to\infty} R(p_S(G)) = -\ \log_e \left\{\frac{\rho(G)}{1+1-[\rho(G)]^2)^{\frac{1}{2}}}\right\}$.

The block SOR method[1] with relaxation factor w on $\hat{G}$ is

$$\begin{pmatrix} Y^{(k+\frac{1}{2})} \\ Y^{(k+1)} \end{pmatrix} = w\hat{G}\begin{pmatrix} Y^{(k)} \\ Y^{(k+\frac{1}{2})} \end{pmatrix} + (1-w)\begin{pmatrix} Y^{(k-\frac{1}{2})} \\ Y^{(k)} \end{pmatrix} \qquad (40)$$

which may be treated as

$$Y^{(k+1)} = w\ G\ Y^{(k+\frac{1}{2})} + (1 - w)\ Y^{(k)} \qquad (41)$$

by Golub (1959). (41) is precisely a stationary version of the semi-iterative method (34). Further $\hat{G}$ is 2-cyclic and $\rho(\hat{G}) = \rho(G)$. We may use the results of Chapter 8.6 specifically

$w^* = 1+ \left\{\frac{\rho(G)}{1+(1-[\rho(G)]^2)}\right\}^2$ and $\rho(\hat{G}_{w^*}) = \mu = |w^* - 1|$. Let $q_S(G)$ be the

---

1. Based on Varga Theorem 5.3

polynomial of order S implied by (41). If $e^{(k)} = Y^{(k)} - \dot{Y}^*$ then (41) with

$$w^* \text{ implies } q_{S+1}(G) = w^* G\, q_S(G) + (1 - w^*)\, q_{S-1}(G) \tag{42}$$

renumbering by S = 2k-1 and noting $e^{(k)} = q_k(G)e^{(0)}$ where $e^{(0)}$ and $e^{(1)} = Ge^{(0)}$ are the initial error vectors for both (40) and (34). (42) has roots $|\phi| = |q_S(G)|$ satisfying $\phi^2 - w^*\lambda\phi + (1-w^*) = 0$, $\lambda = \rho(G)$, giving $|\phi| = \max_i |q_S(\lambda_i)|$ But since $\hat{G}$ is 2-cyclic μ is given by $\mu + w^* - 1 = w^*\lambda\mu^{\frac{1}{2}}$ so $\mu = |w^* - 1| = |\phi|^2$. Hence $|\phi| = |w^* - 1|^{\frac{1}{2}}$ and since $R(q_S(G)) - \log|\phi|$ we conclude

$$\lim_{S\to\infty} R(q_S(G)) = -\log_e\left\{\frac{\rho(G)}{1+(1-[\rho(G)]^2)^{\frac{1}{2}}}\right\} = \lim_{S\to\infty} R(p_S(G)) \tag{43}$$ #

Theorem 9(c) may be immediately used to produce comparison theorems. We compare each semi-iterative method with its stationary counterpart as well as with various other semi-iterative methods for relative convergence speeds.

Theorem 10:

The semi-iterative method bas on G_1 is iteratively faster on both asymptotic and average rates than that on G_2 iff $\rho(G_1) < \rho(G_2) < 1$ or $1>\rho(G_2)>\rho(G_1)$; a and b as in Theorem 9.

Proof:

From (43) if $\rho(G) \leq 1$ then $\lim_{S\to\infty} R(p_S(G))$ is monotonically decreasing with $\rho(G)$ since

$$\xi = \rho(G)[1+(1-[\ (G)]^2)^{\frac{1}{2}}]^{-1} \tag{44}$$

monotonically increases with $\rho(G)$. If $\rho(G) > 1$ (44) is complex but (43)

may also be written

$$\lim_{S\to\infty} R(p_S(G)) = \log_e\left\{{}^1\!/_{\rho(G)}+\left[{}^1\!/_{\rho(G)} - 1\right]^{\frac{1}{2}}\right\} = \cosh^{-1}({}^1\!/_{\rho}(G)),$$ and since

$\cosh i\theta = \cos\theta$ we may use $\lim_{S\to\infty} R(p_S(G)) = \cos^{-1}({}^1\!/_{\rho(G)})$ here. For

$1 < \rho(G) < \infty$, $\cos^{-1}({}^1\!/_{\rho(G)})$ is monotonically increasing in $\rho(G)$. If say

$\rho(G_1) < 1 < \rho(G_2)$ since $\lim_{S\to\infty} R(p_S(G)) = \lim_{S\to\infty} \max|p_S(G)| = |C_S({}^1\!/_{\rho(G)})|^{-1}$

then G_1 is iteratively faster if $\cosh^{-1}({}^1\!/_{\rho(G_1)}) > \cos^{-1}({}^1\!/_{\rho(G_2)})$.

$$R(p_S(G_1)) - R(p_S(G_2)) = {}^{-1}\!/_{S}\left\{S(\log \xi_1 - \log \xi_2)-\left(\log(1+\xi_1^{-2S})-\log(1+\xi_2^{-2S})\right)\right\} \quad (45)$$

where ξ_1 is (44) with $\rho(G_1)$ etc. If $\rho(G_1) < \rho(G_2) < 1$ the second bracketed term is clearly of opposite sign to the first. By above if $1<\rho(G_2) < \rho(G_1)$, $\log \xi_1$ and $\log \xi_2$ are inverse cosine functions giving the first term negative and hence the second positive. Thus either way $R(p_S(G_1)) - R(p_S(G_2)) \geq 0$. We can test the asymptotic speeds of convergence and clearly if $\lim_{S\to\infty} R(p_S(G_1)) > \lim_{S\to\infty} R(p_S(G_2))$ then it holds for average rates also even in $\rho(G_1) < 1 < \rho(G_2)$ because the two bracketed terms in (45) will ways be of opposite signs. #

Corollary 10:

Of any two matrices G_1 and G_2, the faster semi-iterative method has $|\rho(G_1) - 1| > |\rho(G_2) - 1|$. Specifically the tests of relative speed of convergence on the basic stationary methods is also the test of relative speed of convergence of its semi-iterative method. In particular semi -iterative JOR is best done with $G = B_{\alpha^*}$, semi-iterative SOR best done with $G = R_{\alpha^*}$; and semi-iterative SOR is better than that on JOR if B is p-cyclic or $B \geq 0$ or the tests of Chapter 9 are satisfied. #

Semi-iterative Richardson's method, i.e. $G = I + pA$ p a scalar may be treated just as semi-iterative JOR. We note Theorem 11.4.1 of Young (1971): for a given Richardson's method a semi-iterative method with $G = I + pA$ can always be constructed to yield exactly the same iterates for some start $Y^{(0)}$. Conversely for any given semi-iterative method with $G = I + pA$ there is a Richardson's method which has the same n^{th} iterate.

We may also use Theorem 9 to compare semi-iterative and stationary methods for a given matrix G.

Theorem 11:

The semi-iterative method on G is iteratively at least as fast asymptotically and faster on average rates of convergence than the first order stationary method on G; a and b as in Theorem 9.

Proof:

The stationary method implied by G, $Y^{(S)} = G\, Y^{(S-1)}$ has $q_S(G) = G^S$. The semi-iterative version has $p_S(G)$ as in (32). Therefore $p_S(1) = 1$, $q_S(1) = 1$ and by Theorems 9(a) and 6 $p_S(G)$ represents a unique min $|p_S(G)|$ for any G with finite $S \geq 2$. Thus $R(p_S(G)) > R(q_S(G))$ for $S = 2, \ldots$ and also $\lim_{S\to\infty} R(p_S(G)) \geq \lim_{S\to\infty} R(q_S(G))$. In fact only when

$$|\phi| = |\rho|^{\frac{1}{2}} = \lim_{S\to\infty} |C_S(1/\rho(G))|^{\frac{1}{2}} = \frac{\rho(G)}{1+(1-[\rho(G)]^2)^{\frac{1}{2}}} \qquad (46)$$

where ρ is the spectral radius of the iteration matrix of the first order stationary method, will the two methods have equal asymptotic rates. Since $||p_S(G)|| < ||q_S(G)||$ for $p_S(G) \neq q_S(G)$ the semi-iterative average rate is always better. #

Theorem 11 implies we can always improve the average rate of convergence of any first order iterative method by superimposing the semi-iterative acceleration. It is very probable we can improve the asymptotic rate. In fact, we compare only semi-iterative methods with convergent alternatives; $\rho(G) < 1$.

Corollary 11:

If a and b are chosen as in Theorem 9, the semi-iterative version of any JOR is iteratively faster than JOR itself.

Proof:

The statement follows by Theorem 11 iff $\rho^{-\frac{1}{2}} > \left[\frac{1}{1 + (1 - \rho^2)}\right]$ from (46) which holds for any $\rho < 1$; $\rho = \rho(B_\alpha)$. The corollary applies to optimal JOR, and Richardson's stationary method (Lemma 1) but is qualified below (Theorem 12 following). #

Corollary 12:

If B is 2-cyclic then the semi-iterative method on B has a higher average rate of convergence but the same asymptotic one as optimal SOR.

Proof:

From (46) and (8.34) or Theorem 9 parts (a) and (b) viewing $\lim_{S\to\infty} w_{S+1}$. #

In other circumstances the speed of convergence of SOR, optimal or not for G, may be tested against the 2-cyclic case[1] as a proxy for the asymptotic speed of the semi-iterative method on G. Corollary 12 makes the determination of the preferable method easy for a particular SOR case but it is rather weak because a simple calculation of optimal SOR still eludes us. It would be attractive to prove that the simple scheme of semi-iterative optimal JOR is preferable to optimal SOR (stationary) but it is not possible. It will frequently be the case, but in practice must be tested directly in (46). We can make no statements on the comparison of semi-iterative and optimal SOR without extracting the spectral radius $\rho(R_{\alpha^*})$ which is unlikely to be worthwhile because typically semi-iterative methods will achieve the same order of acceleration as optimal SOR over Jacobi with only the calculation of the dominant eigenvalue. Instead we offer

---

1. This remark appears to contradict the conclusions of Young which were based on Zarantanello's theorem: P(X) an m^{th} order polynomial with p(1)=1 has for any $0 \leq r \leq 1$, max $P(X) \geq r^m$. A counter example is $X^2+.1X-.1$. Optimal SOR is not necessarily always as good as semi-iterative methods based on G.

Corollary 13:

The semi-iterative method on $G = R_\alpha$, where α is chosen as α^* or a "good" value for SOR as in the previous chapter, is iteratively faster than that SOR.

Proof:

Identical to Corollary 11. #

The sensible strategy, therefore, is to perform the tests on the iteration matrices of stationary methods - with optimal relaxation parameters - and set that which has lowest spectral radius as G. By this means, the best available stationary method, within whatever practical computation constraints are implied by minimum overall computation, is accelerated both in average and asymptotic rates. Again, as in Chapter 9.6.2 the acceleration in average rates is of particular importance for the finite tolerance of convergence which will be used implies we will never actually reach asymptotic rates. For looser tolerances this point should be borne in mind.

So far we have only considered the rule $b = -a = \rho(G)$.

Theorem 12:

The semi-iterative method on G is iteratively faster if a and b are chosen as in Lemma 7 than if $b = -a = \rho(G)$.

Proof:

Let $b = -a + \varepsilon$; $\varepsilon \neq 0$. Then $Z = \frac{2 - \varepsilon}{2b - \varepsilon} = \frac{2 - \varepsilon}{\varepsilon - 2a}$.

When $\rho(G) = b$, $\varepsilon > 0$ and $Z > {}^{1}/_{\rho(G)} = \frac{1}{b}$ for $0 < \rho(G) < \infty$.

When $\rho(G) = a$, $\varepsilon < 0$ and $Z > {}^{1}/_{\rho(G)} = \frac{1}{-a}$; $0 < \rho(G) < \infty$.

Since we can write $||p_S(G_1)|| = |c_S(Z)|^{-1}$ and $||p_S(G_2)|| = |c_S({}^{1}/_{\rho(G)}|^{-1}$,

say, the theorem follows from Theorem 10 for any integer $2 \leq S \leq \infty$. #

Thus if a,b are chosen as Lemma 7 the imposition of a semi-iterative acceleration always improves the asymptotic as well as average rates of convergence over the underlying stationary method - including the case of B 2-cyclic being the only case when the imposition made no improvement on optimal SOR if $b = -a = \rho(G)$.

Corollary 14:

If B has real eigenvalues, semi-iterative methods on B_{α^*} coincide with semi-iterative methods on B (see Young theorem 11.3.1).

Proof:

$Z = \frac{2 - (b + a)}{b - a} = \frac{2 - (b' + a')}{b' - a'}$ where B has eigenvalues λ_i, B_{α^*} has λ'_i. If $0 \leq \lambda_i \leq b$ we have $a' = \alpha^* a + 1 - \alpha^* \leq \lambda'_i \leq \alpha^* b + 1 - \alpha^* = b'$. But if some λ_i are complex it is not generally true that a,a' and b,b' have the relation given. #

The conclusion of this section is that semi-iterative acceleration should be used as a powerful auxiliary to the best simply computed stationary method. Its effectiveness may only be accurately assessed in particular cases, but it seems likely that the speed of convergence will be increased to compensate for any lack of optimality in the stationary method. As an auxiliary acceleration it has the benefits of maximising the average rate of convergence and of being extremely simple computationally.

10.6.3 An Extension of Semi-iterative Accelerations:

It is useful to complete the semi-iterative theory by listing the several different interpretations it may be given. A semi-iterative method on G may be obtained by accelerating the sequence of Jacobi iterates generated by G by means of Cesaro sums. This was Young's approach. We could also regard nonstationary methods as those which maximise each average rate of convergence - patently stationary methods cannot do this and are consequently set to maximise only the asymptotic convergence rate. The second order nonstationary methods are those which overcome the rounding error problems and the problem

We could also develop semi-iterative methods from second order stationary methods. For example the iteration

$$Y^{(S+1)} = G\,Y^{(S)} + H\,Y^{(S-1)} \tag{47}$$

can be rewritten

$$V^{(S+1)} = \hat{G}\,V^{(S)}; \quad \hat{G} = \begin{bmatrix} 0 & I \\ H & G \end{bmatrix} \text{ and } V^{(S)} = \begin{pmatrix} Y^{(S-1)} \\ Y^{(S)} \end{pmatrix} \tag{48}$$

If the associated stationary first order method was

$$Y^{(S+1)} = G_1\,Y^{(S)}$$

then the second order method may be written

$$Y^{(S+1)} = Y^{(S)} + d(Y^{(S)} - Y^{(S-1)}) + e\,[G_1\,Y^{(S)} - Y^{(S)}].$$

This compared to (28) shows the semi-iterative method merely as a nonstationary version of the second order stationary method with $d = \alpha_{S+1} - 1$; $e = 2\alpha_{S+1}(2 - (a+b))^{-1}$; $G = (1 - e + d)I + e\,G_1$ and $H = -dI$. Further the optimal choice of d,e (to minimise $\rho(\hat{G})$) are shown by Young to be $d = w^* - 1$; and $e = \dfrac{2w^*}{2 - (b+a)}$ and thus $\rho(\hat{G}) = |w^* - 1|$, where $w^* = 1 + \left\{\dfrac{\rho(G)}{1+(1-[\rho(G)]^2)^{\frac{1}{2}}}\right\}^2$

Theorem 13:

Therefore the optimal second order method coincides with the semi-iterative method on G_1 as $S \to \infty$. The semi-iterative acceleration maximises average rates of convergence for second order methods - a welcome improvement since finite tolerances will be used. #

Theorem 14:

We conclude (a) a second order stationary method can always be found which converges, (b) we can construct a second order method corresponding to any first order method by variation of G_1 given the system (8 . 5). #

The semi-iterative method when G = B is 2-cyclic, known as the Cyclic Chebychev semi-iterative method was introduced by Golub and Varga (1961-2) It was developed as the MSOR method in which not only were different relaxation

parameters allowed for the two blocks of equations, but also they could vary with the iteration number. Norms other than the spectral norm were found to be improved. Young develops exactly the same method as the semi-iterative Jacobi method (which starts from (40)).

An extension of semi-iterative methods is suggested by (48). Any system of n differential or difference equations with a maximum order p may be treated as a system of np equations of the first order by the definition of a new variable, as $V^{(S)}$ in terms of $Y^{(S)}$ above. We could then consider a p^{th} order iteration to replace (47) based on the same general first order iterative method $Y^{(S+1)} = G_1 Y^{(S)}$. Therefore we may extend semi-iterative (p) methods as the nonstationary iteration corresponding to this p^{th} order iteration; where the parameters are chosen to maximise the average rate of convergence at every iteration. Such a p^{th} order method could, as in (40), be treated as a block p-cyclic SOR.

$$\begin{pmatrix} Y^{(S)} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y^{(S+p)} \end{pmatrix} = w \begin{bmatrix} 0 & G_1 & & & \\ \cdot & \cdot & & & 0 \\ \cdot & & \cdot & & \\ \cdot & & & \cdot & \\ \cdot & & & & \cdot \\ \cdot & 0 & & & \\ \cdot & & & & G_1 \\ G_1 & 0 & \cdots & & 0 \end{bmatrix} \begin{pmatrix} Y^{(S+p-1)} \\ Y^{(S-1)} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y^{(S+p-2)} \end{pmatrix} + (1-w) \begin{pmatrix} Y^{(S+p-2)} \\ Y^{(S-2)} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y^{(S+p-3)} \end{pmatrix}$$

which would have w^* as the solution α^* of (8.34) and asymptotic rate of convergence $\lim_{S\to\infty} R(q_S(G_1)) = \frac{-1}{p}\log_e |w^* - 1|$ in the notation of (43). Thus the semi-iterative (p) method would coincide with the p^{th} order stationary method as $S \to \infty$; and would have an asymptotic rate of convergence equal to optimal SOR when B is p-cyclic. A formal proof of this extension is exactly similar to theorem 9 (c). However this extension of semi-iterative theory will not be used and so is not pursued although further accelerations might be achieved. It is not clear how to calculate the acceleration parameters for each step, nor how much computation would be involved.

10.6.4 Nonlinear Equation Systems

We have four methods of equation solution which converge for any iteration matrix - GRF, Richardson, semi-iterative and second order stationary methods. Naturally in a nonlinear equation system, plural solutions are generally possible so we investigate local convergence conditions. The contraction mapping theorem of Ostrowski (1960) - quoted as theorem 12.1 indicates that convergence is assessed exactly by the methods of Chapters 8 - 10 on (8.5) where B is replaced by $B^{(s)} = F'(Y^{(s)})$ and $F'(\cdot) = \partial F/\partial Y^{(s)}$ and the system is $Y^{(s+1)} = F(Y^{(s)})$. We may quote convergence criteria for each method as before with exactly the same replacement of B - obviously the final and overriding criteria using $B = F'(Y^*)$. The optimal acceleration parameters should be reestimated at each step.

Naturally for any of the four methods mentioned we always have local convergence for all equation systems provided the acceleration parameters are continuously reestimated. How this is achieved for the favoured semi-iterative emthod in the nonlinear systems that concern us is specified in Chapter 14. In the context of some iterative estimators we can go beyond local convergence. It is therefore irrelevant to consider more complicated iterations available only for nonlinear equation systems.

CHAPTER 11: A SURVEY OF ITERATIVE ESTIMATORS AND THEIR FAMILY STRUCTURE

11.1 Introduction

Following the analysis of Chapters 5 and 7, we can obviously treat any iterative algorithm that produces, as its iterates, linear combinations of predetermined variables which may be used as instrument sets $\hat{N}$ in either (5.1) or (5.10) to provide structural parameter estimates as an approximation to the maximum likelihood estimator. We can now rank estimators by asymptotic efficiency, classify them as various approximations to FIML, and deduce the asymptotic properties, all by reference to the established hierarchy of maximum likelihood estimators.

The whole theory of iterative estimators therefore hangs on the representation of maximum likelihood estimators given in (5.1) and (5.10). The problem is neither representation indicates any form of the iterative algorithm to produce the instrument set iterates. To preserve asymptotic equivalence to maximum likelihood we need only an algorithm which produces the linear combinations $Zp^{(s)}$ in $\hat{N}^{(s)} = [Zp^{(s)} \vdots Z]$ where plim $p^{(s)} = \Pi$. Thus a new member of the FP and of the IIV families is created everytime we pair a different iterative algorithm with the estimators (5.1) or (5.10). We may choose any algorithm from the whole range of those which produce approximations to the reduced form of the model.

A good many such pairings have already been made. Jacobi iteration gives FP and IIV, JOR gives FFP and FIIV, SOR gives RFP, RIIV, etc. A number of new estimators are suggested by the analysis of the last three chapters. These are specified precisely in Section 2. It is of immediate interest to investigate the structure of both estimation families and classify the various iterative estimators relative to one another according to their various properties. Although it is possible to fit all members into a scheme of classification by

iterative algorithm - including parametric and reduced form versions - it is shown that this is unhelpful. It would not correctly describe the structure of the family; it would obscure the range of decisions that must be taken which independently affect the computational burden of the estimator chosen; and it would prevent an analysis of the computational aspects and thus the selection of a computationally efficient estimator. The structure is analysed in Sections 3 - 5.

11.2 The family structure of an iterative estimator:

11.2.1

Any estimator which follows the iterative schemes (5.1) or (5.10) will be asymptotically equivalent to the maximum likelihood estimator of its information class; and for identical starts identical estimates result. Clearly there are no distinctions to be made between family members on the basis of asymptotic properties. It is further argued (Chapter 16) there is nothing of any generality to be deduced from an analysis of small sample properties which will facilitate comparisons between any estimators. Therefore failing to achieve discrimination on statistical grounds, we must aim for discrimination on computational, availability and "ease of use" grounds. Since all iterative estimators in the FP and IIV families have the same optimal criterion function and the same asymptotic statistical properties the sole possibility of a rational preference ordering is by computational and convergence analysis. This is the first such analysis undertaken for econometric estimators. The classification of the iterative estimators is also new. From its building blocks all members of a family may be constructed; some already well known and some new. However the theory of iterative estimation which underlies their construction, and the identification of the various steps involved, is both new and comprehensive.

11.2.2

We have observed that any iterative estimator may be constructed by marrying an estimation phase with some iterative simultaneous equation solution method into one iteration process. Explicit demonstrations are given by Mosbaek and Wold (1970) and Bodin (1974). Such an analysis presents a wide range of alternatives but comparisons would be uninformative for the following reasons:-

(a) The aim is to provide a computational analysis of each iterative estimation process. Hence each estimator must be constructed from a series of computationally independent steps. We define a step computationally independent if it involves a decision between various alternatives each of which has different consequences for the computational burden of the estimator as a whole, but the decision itself has no consequence for the computational burden imposed by a previous or subsequent decision.

A series of independent decisions enables the design of an estimator with overall minimal computation. Moreover the computational burden may then be analysed from an analysis of the component decisions. Computational analysis would be very difficult treating each estimation iteration as a complete alternative.

(b) Any basic iterative process may be accelerated by methods which reduce the number of iterations or by methods which reduce the computation per iteration, or may be by those which do both. Any comparisons or classification of estimators for computational properties must separate the "within" and "between" iteration computational burdens since the second is an infinite process whereas the first is not. This distinction is completely ignored by a classification treating each iterative process as a complete alternative and so prevents the analysis of computation saving devices.

(c) Parametric and reduced form methods e.g. PFP, PIIV, RFFP, RFIIV may be treated as (5.1) or (5.10) with GRF iteration. It is shown below that

these parametric and reduced form methods in the FP family have iteration matrices (in the notation of (2.5)) with expected value $I - \beta$ and $-(I - \beta)\beta^{-1}$ respectively. In the GRF iteration $I + P\beta$ this implies $P = -I$ and $-[I + (I - \beta)\beta^{-1}]\beta^{-1}$ neither of which are illuminating. The purpose of both methods is not clear, the computational advantages of either are obscured; as is the fact that both may be used in conjunction with other methods or with each other.

(d) Computational analysis of accelerations and calculation of relaxation parameters, orderings, etc. will not be possible without the results of previous chapters. However accelerations, relaxation parameters, etc. may be applied to any matrix A in (8.5). Hence the decisions affecting A must be made before any further acceleration analysis is undertaken. We show, in this scheme, A is always the same matrix in a given model.

(e) It is not clear in (5.10) if structural parameters should be obtained from equations taken seriatim or together.

(f) Analysis by independent decisions permits a vast simplification not only in computational analysis. A huge number of estimators are possible and now we can give them a framework which is comprehensive in that any possible estimator is quickly generated and analysed. Modifications such as full information or autoregressive versions are simple to construct and their convergence analysis carries over from this scheme.

It is obvious that practice of treating each iterative algorithm as a complete alternative cuts across the independence of decisions and creates artificial distinctions between estimators. Not only does it obscure the family structure of estimators but it prevents a tractable computational analysis and obscures the potential of various iteration methods. It is therefore abandoned.

11.2.3

In a formal analysis the concept of independent decisions implies that the expression for the computational burden of any estimator may be written as a function of the computational burdens due to each decision such that the function is weakly separable with respect to variables representing those decision states. We now propose a family structure embodying these independent decisions.

In Chapter 13, where computational burdens are analysed by operation counts, computational burden functions of each estimator are found to display this weak separability. However they are not necessarily weakly separable with alternative definitions of the computational burden. Weak separability is the formal justification of the structure below but we proceed meanwhile with less formality.

11.3 The independent decisions

The decisions in constructing an iterative estimator may be identified as follows:

1. normal equation form,
2. iteration mode; linearity,
3. iteration mode; iteration variable,
4. instrument formation,
5. acceleration,
6. accelerations in instrument formation within each iteration,
7. acceleration superimposed to influence average rates of convergence.

The sequence of decisions may be illustrated in a decision tree indicating all the iterative estimators that may be constructed from the algorithms discussed in this study.

11.3.1 The normal equation decision:

Decision 1:

Since both the FP and IIV families imply asymptotic equivalence with

maximum likelihood estimates, the choice is between instrumental variable normal equations as (5.1) or (5.10) or least square ones [i.e. $X^{(s)}$ becomes a regressor in (5.10); or (5.1) replaced by

$$\hat{\delta}^{(s+1)} = (\hat{X}^{(s)\prime}(\hat{\Sigma}^{(s)} \bar{\otimes} I)^{-1} \hat{X}^{(s)})^{-1} \hat{X}^{(s)\prime} (\hat{\Sigma}^{(s)} \bar{\otimes} I)^{-1} y].$$

Clearly the same amount of computation is involved in either case, at each estimation phase. Moreover since neither normal equation implies any conditions on the iterative process to produce the instrument sets, both the complete range of options on suitable algorithms and the amount of computation to produce each instrument set within any algorithm remain uninfluenced by the normal equation choice. Thus, within the definition given, the normal equation decision is independent of other decisions. The normal equation choice may in fact only influence the total number of iterations necessary for convergence. It is then more efficient to estimate equations seriatim.

11.3.2 The iteration mode decision:

This decision must be taken in two parts. The two parts cannot be treated as two separate decisions since there is no necessity to take any of the options under the first part. Indeed their existence is not general. One of the options of the second part must be taken, at least implicitly.

Iterative estimators are conventionally defined as two part mixed iterations. First instruments are obtained, second the estimates. Then the procedure is repeated.

Decision 2: the linearity decision.

The nonlinearity of interest is in the instrument formation partnering the normal equations selected at Decision 1. It therefore is a question of nonlinearity in model specification, not in the estimator nor in the estimation iteration which is known, e.g.(5.1) and (5.10) to be nonlinear in any case.

If the model is specified to be linear in parameters, or the iteration partnering the estimation is transformed linear then clearly the iteration may

be performed on the reduced form parameters only, since they vary with the iteration number but their variables do not. Thus if the iteration is at step S linear with respect to S,

$$Y^{(s+1)} = Y^{(s)}(I - \hat{\beta}^{(s+1)}) + Z\hat{\Gamma}^{(s+1)} \tag{9}$$

we have substituting $Y^{(s)} = Z\hat{\Pi}^{(s)}$

$$Z[\hat{\Pi}^{(s+1)} - \hat{\Pi}^{(s)}(I - \hat{\beta}^{(s+1)}) - \hat{\Gamma}^{(s+1)}] = 0 \tag{10}$$

Since neither Z nor $[\hat{\Pi}^{(s+1)} - \hat{\Pi}^{(s)}(I - \hat{\beta}^{(s+1)}) - \hat{\Gamma}^{(s+1)}]$ are square, and $Z \neq 0$, a sufficient condition for (10) is

$$\hat{\Pi}^{(s+1)} = \hat{\Pi}^{(s)}(I - \hat{\beta}^{(s+1)}) + \hat{\Gamma}^{(s+1)} \tag{11}$$

Either normal equation system may be used with (11). This is the well known parametric iteration (5.23).

Clearly if the model specification is nonlinear, parametric iteration is not possible. However iterates $Y^{(s)}$ as systematic parts of the model specification still approximate the reduced form dependent variables and a precise representation is achieved at the fixed point $Y^* = F(Y^*,Z)$. In this case a nonlinear iteration could be used, i.e. Newton, Secant or Conjugate gradient methods. The theory has been reviewed and the conclusion drawn that their use was not recommended. A mixed iteration is possible and depending on the incidence of nonlinearities endogenous iteration, explained below, may also be available. Nonlinear iterations contain their own optimal parameters so after the instrument formation decision, following, the estimator construction is complete if this option were taken. Any of the methods of Chapters 8-10 could be used instead - as specified below.

Decision 3: iteration variables.

Whether the model is linear or nonlinear we may ignore part 1 of this decision and merely iterate as (9) or $Y^{(s+1)} = F(Y^{(s)},Z)$. In the case of (9), and possibly in a nonlinear model, we may consider on which variables we iterate. We define a mixed iteration as iterating on both $Y^{(s)}$ and $\hat{\beta}^{(s+1)}$, $\hat{\Gamma}^{(s+1)}$ in a two part procedure as described, e.g. (9) with (5.10). However we could also interpret (11) as parametric iteration, i.e. iterating only on parameters, the variables having been substituted out. There may be computational savings here (see Chapter 13). Another possibility is to substitute the parameters out; that is iterate on the endogenous variables only (Section 4 below). The result is a new iterative scheme, and is in fact a one part iteration. We note (11) still implies a two part iteration so this new estimator will have some further computation savings. However parametric is in an ambiguous position since it is also possible to eliminate the predetermined part and reduce (11) to a one part iteration. Moreover since it is only available for linear models, it is more sensible to classify parametric iteration as part of the linearity decision.

11.4 A new iterative scheme: a digression within the iteration mode decision

There is another possibility which will prove most fruitful. We could attempt to eliminate the estimated parameters at step S. The range of iteration modes would then be complete - mixed, parametric, and endogenous (only $Y^{(s)}$ varies with S among the set of variables). Consider the i^{th} normal equation from (5.10).

$$N_i^{(s)\prime} y_i = N_i^{(s)} X_i \hat{\delta}^{(s+1)}; \quad N_i^{(s)} = [Y_i^{(s)} : Z_i] \text{ at step S} \qquad (12)$$

From the second vector equation within (12)

$$\hat{\gamma}_i^{(s+1)} = (Z_i'Z_i)^{-1} Z_i' (y_i - Y_i \hat{\beta}_i^{(s+1)}) \qquad (13)$$

which when substituted in the first vector equation yields

$$Y_i^{(s)'} y_i = Y_i^{(s)'} Y_i \hat{\beta}_i^{(s+1)} + Y_i^{(s)'} Z_i (Z_i'Z_i)^{-1} Z_i' (y_i - Y_i \hat{\beta}_i^{(s+1)})$$

$$= Y_i^{(s)'} M_i Y_i \hat{\beta}_i^{(s+1)} + Y_i^{(s)'} (I - M_i) y_i \qquad (14)$$

where

$$M_i = I - Z_i (Z_i'Z_i)^{-1} Z_i'$$

$$\therefore \quad \hat{\beta}_i^{(s+1)} = (Y_i^{(s)'} M_i Y_i)^{-1} (Y_i^{(s)'} M_i y_i) \qquad (15)$$

The equivalent of the i^{th} relation in (9), but using the notation (4.30b), is

$$y_i^{(s+1)} = Y_i^{(s)} \hat{\beta}_i^{(s+1)} + Z_i \hat{\gamma}_i^{(s+1)} \qquad (15a)$$

so using (13) and (15) we get the iteration which is purely in endogenous variables

$$y_i^{(s+1)} = (I - M_i) y_i + [Y_i^{(s)} - (I - M_i) Y_i] (Y_i^{(s)'} M_i Y_i)^{-1} (Y_i^{(s)'} M_i y_i) \qquad (16)$$

The iteration (16) on endogenous variables appears to be new and is offered to complete the range of iteration modes under Decision 3. The iteration is called endogenous iteration - (16) is endogenous IIV (EIIV). It is easily verified that had we taken a least squares normal equation in place of (12) with $N_i^{(s)}$ as regressors then we obtain endogenous FP (EFP)

$$\text{i.e.} \quad \beta_i^{(s+1)} = (Y_i^{(s)'} M_i Y_i^{(s)})^{-1} Y_i^{(s)'} M_i y_i \qquad (18)$$

in place of (15) and

$$y_i^{(s+1)} = (I - M_i) y_i + M_i Y_i^{(s)} (Y_i^{(s)'} M_i Y_i^{(s)})^{-1} Y_i^{(s)'} M_i y_i \quad i = 1 \ldots M \qquad (19)$$

replacing (16). Edogenous iteration has the advantage, shared with PFP, PIIV, that one part of the iteration is redundant, but they are superior in that no predetermined variable coefficients need be calculated until after convergence. Endogenous iteration saves on both counts over mixed iteration and on the second over parametric iteration. We may combine the advantages of endogenous and parametric iterations as follows. Let $Y_i^{(s)} = Z\,(\hat{\Pi}^{(s)})_i$ and $y_i^{(s)} = Z\hat{\Pi}_i^{(s)}$, which may be substituted into (15) and (18). We may write (16) and (19) as

$$Z\hat{\Pi}_i^{(s+1)} = (I - M_i)y_i + M_i Z(\hat{\Pi}^{(s)})_i\ \hat{\beta}_i^{(s+1)} \tag{19a}$$

if (15) or (18) respectively are used for $\hat{\beta}_i^{(s+1)}$, with $Y^{(s)} = Z\hat{\Pi}^{(s)}$.

Alternatively we may write out one equation of (11) as

$$\hat{\Pi}_i^{(s+1)} = (\hat{\Pi}^{(s)})_i\ \hat{\beta}^*{}_i^{(s+1)} + \hat{\gamma}^*{}_i^{(s+1)} \tag{19b}$$

where

$\hat{\beta}^{*(s+1)}$ and $\hat{\gamma}_i^{*(s+1)}$ have all parameters of the i^{th} columns of $\hat{\beta}^{(s+1)}$ $\hat{\Gamma}^{(s+1)}$ respectively including the zero ones. Write C_i^* as $(Z_i'Z_i)^{-1}Z_i'$ but with added rows of zeros corresponding to $\gamma_{ji} = 0$ $j = 1 \ldots K$, for this i. C_i^* then is K x T. By (13) we have

$$\hat{\Pi}_i^{(s+1)} = (\hat{\Pi}^{(s)})_i\ \hat{\beta}_i^{(s+1)} + C_i^* y_i - C_i^* Z(\hat{\Pi}^{(s)})_i\ \hat{\beta}_i^{(s+1)}$$

$$= C_i^* y_i + I_i^*(\hat{\Pi}^{(s)})_i\ \hat{\beta}_i^{(s+1)} \tag{19c}$$

where I_i^* has $I^*_{ijj} = 1$ if $\gamma_{ji} = 0$; $I^*_{ijj} = 0$ if $\gamma_{ji} \neq 0$; $I^*_{ijk} = 0$ $j \neq k$ — from consideration[1] of $C_i^* Z$, $k, j = 1 \ldots K$

---

1. note $I_i^* = I - M_i Z$ iff $z_i'z_j = 0$ all $i \neq j$ (unlikely).

The difference between the FP and IIV versions of (19c) is again in the new forms of (15) and (18) that we suggest be used for $\hat{\beta}_i^{(s+1)}$. (19c) is easier to use and computationally preferable to (19a)[1], and is therefore recommended.

It is not surprising that we find for all practical purposes endogenous itera tion in superior for reducing the computation - a subject taken up in Chapter 13.

Obviously we can specify endogenous iteration for the full information case (5.1). In an IV context the normal equations are at step S

$$\hat{X}^{(s)'}\ \phi^{(s)-1}\ \hat{X}\ \hat{\delta}^{(s+1)} = \hat{X}^{(s)'}\ \phi^{(s)-1}\ y \tag{20}$$

with

$$\phi^{(s)} = \frac{1}{T}(y - \hat{X}\hat{\delta}^{(s)})'(y - \hat{X}\hat{\delta}^{(s)}) = \Sigma^{(s)} \otimes I_T \tag{21}$$

or an auto regressive version with $\phi^{(s)} = T^{-1}$ diag $\{(y_i - X_i\hat{\delta}_i^{(s)})(y_i - X_i\hat{\delta}_i^{(s)})'\}$ or $\phi^{(s)}$ a mixture of the two as in Section 7.6. In this connection it is convenient to amend the notation of (2.5) to separate the γ and β coefficients. Let $\hat{X}^{(s)} = [\text{diag}\ (Y_i^{(s)}) \vdots \text{diag}\ Z_i]$ and $\hat{\delta}^{(s)} = \begin{bmatrix} \text{vec}\ (\hat{\beta}_i^{(s)}) = \hat{\beta}_0^{(s)} \\ \text{vec}\ (\hat{\gamma}_i^{(s)}) = \hat{\gamma}_0^{(s)} \end{bmatrix}$.

Thus if diag $(Y_i^{(s)}) = \hat{Y}^{(s)}$ and diag $Z_i = \hat{Z}$, (20) becomes

$$\begin{bmatrix} \hat{Y}^{(s)'}\ \phi^{(s)-1}\ \hat{Y} & \hat{Y}^{(s)'}\ \phi^{(s)-1}\ \hat{Z} \\ \hat{Z}'\ \phi^{(s)-1}\ \hat{Y} & \hat{Z}'\ \phi^{(s)-1}\ \hat{Z} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0^{(s+1)} \\ \hat{\gamma}_0^{(s+1)} \end{bmatrix} = \begin{bmatrix} \hat{Y}^{(s)'}\ \phi^{(s)-1} \\ \hat{Z}'\ \phi^{(s)-1} \end{bmatrix} y \tag{22}$$

As before taking the second equation $\hat{\gamma}_0^{(s+1)} = (Z'\ \phi^{(s)-1}\hat{Z})^{-1}\ \hat{Z}'\ \phi^{(s)-1}(y - \hat{Y}\hat{\beta}_0^{(s+1)})$; and substituting into the first we get with manipulation

$$\begin{aligned} \hat{\beta}_0^{(s+1)} &= \{\hat{Y}^{(s)'}\phi^{(s)-1}\hat{Y} - \hat{Y}^{(s)'}\phi^{(s)-1}\hat{Z}(\hat{Z}'\phi^{(s)-1}\hat{Z})^{-1}\hat{Z}'\phi^{(s)-1}Y\}^{-1}\{\hat{Y}^{(s)'}\phi^{(s)-1} \\ &\qquad - \hat{Y}^{(s)'}\phi^{(s)-1}\hat{Z}(\hat{Z}'\phi^{(s)-1}\hat{Z})^{-1}\hat{Z}'\phi^{(s)-1}\}y \\ &= [\hat{Y}^{(s)'}M^{*}(s)\hat{Y}]^{-1}[\hat{Y}^{(s)'}M^{*}(s)y] \end{aligned} \tag{23}$$

---

1. Lemma 13.5

where $$M^{*(s)} = \phi^{(s)^{-1}} [I - \hat{Z}(\hat{Z}'\phi^{(s)^{-1}}\hat{Z})^{-1}\hat{Z}' \phi^{(s)^{-1}}] \quad (24)$$

If we employ the same iteration as in the limited information case above, i.e. $y^{(s+1)} = \hat{Y}^{(s)} \hat{\beta}_0^{(s+1)} + \hat{\hat{Z}}\hat{\gamma}_0^{(s+1)}$ then

$$y^{(s+1)} = [I-\phi^{(s)}M^{*(s)}]y + [\hat{Y}^{(s)}-(I-\phi^{(s)}M^{*(s)})\hat{Y}](\hat{Y}^{(s)'}M^{*(s)}\hat{Y})^{-1}\hat{Y}^{(s)'}M^{*(s)}y \quad (25)$$

which is similar to (16).

At (25) we have reduced a three part iteration - on $Y^{(s)}$, $\delta^{(s)}$ and $\phi^{(s)}$ - to a two part scheme. By (21) $\phi^{(s)}$ is a function of $\delta^{(s)}$; $\delta^{(s)}$ has not been fully eliminated in (25). We can also write (25) as

$$y^{(s+1)} = y + \phi^{(s)}M^{*(s)}(\hat{Y} \hat{\beta}_0^{(s+1)} - y) + (\hat{Y}^{(s)} - \hat{Y}) \hat{\beta}_0^{(s+1)} .$$

From (21), (24), $y-\hat{\hat{Y}}\hat{\beta}_0^{(s+1)}-\hat{\hat{Z}}\hat{\gamma}_0^{(s+1)} = y-\hat{\hat{Y}}\hat{\beta}_0^{(s+1)}-(I-\phi^{(s-1)}M^{*(s-1)})(y-\hat{\hat{Y}}\hat{\beta}_0^{(s+1)})$

$$= \phi^{(s)}M^{*(s)}(y-\hat{\hat{Y}}\hat{\beta}_0^{(s+1)})$$

Thus

$$\phi^{(s)} = \frac{1}{T}(y^{(s)}-y+\hat{V}^{(s)}\hat{\beta}_0^{(s)})'(y^{(s)}-y+\hat{V}^{(s)}\hat{\beta}_0^{(s)}); \quad \hat{V}^{(s)} = (\hat{Y}^{(s)}-\hat{Y}) \quad (26)$$

The iterative process (25) can therefore be reduced to iterations in endogenous variables alone by substituting for $\hat{\beta}_0^{(s)}$ in (26) using (23), and for $\phi^{(s)}$ in $M^{*(s)}$ by (26) throughout. The formulae become vastly more complicated and it is computationally more efficient to proceed with the two part iteration. The purpose of (26) is merely to demonstrate that even for (5.1) iteration purely within endogenous variables is possible. It is a general proposition, investigated in Section 5 below, that any iterative algorithm, iterating in any number of parts, may always be reduced to an iteration in a single unknown exactly as the underlying iteration appropriate to (5.1). Indeed the speed of convergence is unaffected, by (12.23) below, although the work per full information iteration is increased. In place of (26), and writing

$U^{(s)} = V^{(s)} - V^{(s-1)}(I - \beta^{(s)})$ from (6.5) we have $\Sigma^{(s)} = T^{-1}\, U^{(s)\prime} U^{(s)}$ and $\phi^{(s)-1} = \Sigma^{(s)} \otimes I_T$. Not only are the matrix inversions reduced from order MT to order M which is preferable to (26) but $\hat{\beta}^{(s)}$ is calculated automatically in the process of obtaining $y^{(s)}$ and therefore involves no extra computation. It is suggested the greatest economies in computation are achieved by pursuing a partial endogenous iteration obtaining $y^{(s+1)}$ and $\Sigma^{(s)}$ separately, rather than substituting (26) and (23) into (25), in the full information version.

Once again it is easy to derive the equivalent formulae for the least squares normal equation case:

$$\hat{\beta}_O^{(s+1)} = (\hat{Y}^{(s)\prime} M^*(s)\, \hat{Y}^{(s)})^{-1}\, (\hat{Y}^{(s)\prime}\, M^*(s) y) \qquad (27)$$

and

$$y^{(s+1)} = [I - \phi^{(s)} M^*(s)]y + \phi^{(s)} M^*(s) \hat{Y}^{(s)} (\hat{Y}^{(s)\prime} M^*(s) \hat{Y}^{(s)})^{-1} \hat{Y}^{(s)\prime} M^*(s) y \qquad (28)$$

and $\phi^{(s)}$ in (26) remains the same. The parametric versions are similarly straightforward and follow (25) or (28) in the form (19c).

11.5 The independent decisions (cont.)

We return to the discussion of the decisions taken in building up the iterative estimator. It is clear the iteration mode decision is independent of both the normal equation decision and the decisions on the precise form of the instrument iteration. Both parametric and endogenous iteration follow from the iterative scheme proposed at (9) and (15a). However these schemes were chosen arbitrarily and many other available iterations, accelerated in any suitable manner, might have been used. (16), (19), (25) and (28) would then be amended in an obvious manner. We may take mixed iteration as the basis and consider parametric and endogenous iterations and their combination as dependent on that; differing only by which unknowns are substituted out of the iteration as redundant. But mixed iteration itself is no decision for

(9) and (15a); no particular iteration is implied. In fact it will turn out in Chapter 12 that each iteration mode has the same iteration matrix which is itself governed by the iteration form, (e.g. as taken at (9) or (15a)) for the corresponding mixed iteration.

Decision 4:

It is obvious that the iteration mode decision is one part of the process of achieving within iteration computational savings. But, apart from deciding in which unknowns it is best to iterate, we might consider how much work is best done to produce each instrument set $N^{(s)}$. In (9) and (15a) we consider only one step in a potential iteration

$$y^{(s+1,k+1)} = \hat{Y}^{(s+1,k)} \hat{\beta}_0^{(s+1)} + \hat{Z} \hat{\gamma}_0^{(s+1)} \qquad (29)$$

$k = 0, \ldots ;$ S fixed, and $Y^{(s+1,0)} = Y^{(s)}$.

We have, for S = 1 .... , stopped at k = 1. However we could have iterated beyond that with respect to k, in the hope that $\hat{N}^{(s+1)} = [\hat{Y}^{(s+1,k+1)} : \hat{Z}]$ will reduce the value of S necessary for the convergence of the estimator enough to offset the extra work due to higher values of k. Iterations in the index S are called estimation iterations, and those in k instrument iterations. Either or both may be accelerated in the conventional ways noted below.

An obvious possibility is to iterate (29) to convergence with respect to k obtaining

$$y^{(s+1,k+1)} = \hat{Y}^{(s+1,k+1)} \hat{\beta}_0^{(s+1)} + \hat{Z} \hat{\gamma}_0^{(s+1)} \qquad (30)$$

(29) with max k = 1 is method 2 whereas (30) is method 1 of Maddala (1970). Naturally (30) in a linear model gives RFFP or RFIIV and is expressed in the more compact notation of (4.36) and (5.22). However it is emphasised that the model does not have to be linear for RFFP or RFIIV to be used. We can, without any linearising approximations, arbitrarily well approximate the

restricted reduced form of the system by instrument iterations of the type (30). The more familar notation

$$Y^{(s+1,k+1)} = Y^{(s+1,k+1)} (I - \hat{\beta}^{(s+1)}) + Z\Gamma^{(s+1)} \qquad (31)$$

involves exactly the same computation as (30), and is that proposed as LIVE and FIVE by Brundy and Jorgenson. However applying the results of Chapter 13.5.1 we conclude that the form (4.36) when it is available will not be computationally superior to (31) except for very small models, at any value of S.

Instrument formation (29), with max k = 1, is called direct instrument formation and (30) reduced form instrument formation. Clearly both are independent of the normal equation decision. Similarly if they are valid for mixed iteration, they also are for parametric, endogenous, their combination, and nonlinear iterations. If we substitute (29) for (9) and (15a) for some k we get

$$\hat{\Pi}^{(s+1,k+1)} = \hat{\Pi}^{(s+1,k)} (I - \hat{\beta}^{(s+1)}) + \hat{\Gamma}^{(s+1)} \qquad (32)$$

with $\hat{\Pi}^{(s+1,0)} = \hat{\Pi}^{(s)}$ to replace (11). Similarly (16) becomes

$$y_i^{(s+1)} = (I - M_i)y_i + [Y_i^{(s+1)} - (I - M_i)Y_i] (Y_i^{(s)'}M_iY_i)^{-1}Y_i^{(s)'}M_iy_i \qquad (33)$$

when RF instruments (31) are used together with (12) - (15) which are unchanged. (33) is RFEIIV. We can obtain RFEFP, using (31) and (18), to replace (19) as

$$y_i^{(s+1)}=(I-M_i)y_i + [Y_i^{(s+1)}-(I-M_i)Y_i^{(s)}](Y_i^{(s)'}M_iY_i^{(s)})^{-1}Y_i^{(s)'}M_iy_i \qquad (33a)$$

of course versions of (29) are also available, e.g. (33a) becomes

$$y_i^{(s+1,k+1)} = (I-M_i)y_i+(Y_i^{(s+1,k)}-(I-M_i)Y_i^{(s)})(Y_i^{(s)'}M_iY_i^{(s)})^{-1}(Y_i^{(s)'}M_iy_i) \tag{33b}$$

with $Y_i^{(s+1,0)} = Y^{(s)}$ $i = 1 \ldots M$; and (19), (25), (28) may all be amended in the same way, and $\Sigma^{(s+1)}$ will automatically be adjusted.

Obviously if the instrument formation decision is solely concerned with what value of max (k) we use in (29), it is independent of any decision on the form of the iteration (29) to produce instruments which is set by the "fixed" functions of S, $\hat{\beta}_0^{(s+1)}$ and $\hat{\gamma}_0^{(s+1)}$; and in particular what accelerations are used either with respect to k or S. We could profitably, in full information cases, revert to notation where $U^{(s)} = V^{(s)}\hat{\beta}^{(s)}$; that is $V^{(s)} = Y^{(s)} - Y$ and $\beta^{(s)}$ is the M x M matrix obtained from (23) both of which are available from the previous iteration. Then $\Sigma^{(s)} = T^{-1}\hat{\beta}^{(s)'}V^{(s)'}V^{(s)}\hat{\beta}^{(s)}$ and $\phi^{(s)-1} = \Sigma^{(s)^{-1}} \otimes I_T$. Again the inversion involves less computation and $\hat{\beta}^{(s)}$ is already known from obtaining $y^{(s)}$.

There is the immediate generalisation from (29) to allow max (k) as any integer in the interval $[1,k_0]$; where $k = k_0$ satisfies (30). Such an instrument formation procedure will be called quasi-reduced form (QRF) and it will always be stated what rule is used to select max (k). Evidently obtaining QRF instruments with some max (k) is formally equivalent to relaxing the tolerance acceptable, in the instrument formation (30), by some amount to be determined. There are three options either $1 < \max(k) < k_0$ a constant with respect to S; or max (k) chosen to vary with S such that the work done in iterative estimation as a whole is minimised (which may be a complicated calculation); or max (k) chosen at k_0 for S = 1 and steadily reduced as S increases. We return to these options in Chapter 12.

Instrument formation in a linear or nonlinear model by means of (29) allows direct, reduced form or QRF instruments. The choice is in no way diminished if we had elected to use a nonlinear iteration method at the previous decision Max (k) may be set as above in the interval $[1,k_0]$ and those types of instruments be produced.

Decision 5:

Decisions 2 and 3 cover all the possibilities for reducing the work done within the step S to S + 1. We now consider how we may reduce the value of S required to produce convergence of the iterative estimator. This may be considered independently since, apart from the caveat on non-linear iterations, we have retained the same iteration (29) with respect to S itself for each decision so far. Thus, assuming we have made suitable decisions on normal equations, iteration mode and instrument formation, we may investigate the possibilities of changing that iteration by various accelerations. If in (29) with max (k) fixed, $\hat{\beta}_0^{(s+1)}$ was fixed with respect to S we could apply the results of the previous three chapters directly. However, since they alter with S the analysis is very much more complicated.[1] In principle there would be no problem in applying the acceleration results if we did the appropriate calculations at every S, so optimal parameters, methods, orderings, etc. - or at least suitable ones - were used at each stage. However in practice this would imply an outrageous computational burden even for the simpler calculations. We might try selecting parameters, methods and orderings that are suitable for all, or a group of iterations, in the index S on the argument $\hat{\beta}_0^{(s+1)}$ changes sufficiently little with S (which is certainly true for larger values of S).

The range of options for accelerating (29) with respect to S, however they are used in view of changing $\hat{\beta}_0^{(s+1)}$, may be divided into two groups. The first group, considered here, is accelerations which affect the asymptotic rate of convergence with respect to S. (29) is a Jacobi iteration; by means of applying

$$y^{(s+1)} = \alpha y^{(s+1,k_{s+1})} + (1 - \alpha)\, y^{(s,k_s)} \qquad (34)$$

for a suitable scalar α, k_{s+1} = max (k) at step s + 1, we get JOR at (29) and

---

1. Chapter 14 shows two solutions to this problem.

hence fractional estimators, e.g. FFP, FEFP, FRFEFP, FRFPFP and so on. If (29) was made recursive between equations, we have Gauss-Seidel iterations and Gauss-Seidel estimators. If (34) were applied where $y^{(s+1,k_{s+1})}$ and $y^{(s,k_s)}$ were obtained from Gauss-Seidel iterations in (29) we have SOR iterations; and if also the components of β,γ are re-estimated recursively also, recursive estimators, e.g. RFP, REFP, RRFEFP, RRFPFP, etc. Clearly Richardson's stationary and GRF iterations may be applied to (29), as can any of the many amendments introduced in Chapter 9 to improve the calculation of α^* or SOR. Any of these accelerations may be used to reduce the spectral radius of the estimation iteration - if the appropriate iteration matrix of the estimator implied by the iteration mode, instrument formation method and normal equations chosen, is known - by means of a suitable ordering and stationary α parameter from the results of the previous three chapters where B is that iteration matrix. The asymptotic rate of convergence of estimation iteration is thus increased.

If we widen our range to include nonstationary accelerations we may also maximise the average rate of convergence at each S - but this as we have seen in Chapter 10 requires the specification of the underlying linear iteration matrix and hence is an independent decision - the seventh. Moreover Decision 5 is independent of Decision 6 since they refer to separate iterations in different indices. There is no necessity beyond convenience in using the same accelerations in both iterations and there is no computational effect on one caused by the other. Similarly accelerating (29) or (15a) or (9) will change the overall iteration with respect to S, but not the within iteration work, as already established.

Decision 6:

If the results of the previous three chapters may only be applied imprecisely in practice to accelerating the estimation iteration due to $\hat{\beta}_0^{(s+1)}$ altering with S and the computational problems, they may be applied exactly to accelerate the instrument iteration; k_s specified each S.

We refer to (31). The iteration matrix is then $(I-\hat{\beta}^{(s+1)})$ fixed for all $1 \leq k \leq \max(k)$. For $(I - \hat{\beta}^{(s+1)}) = B$ we may accelerate the instrument iteration optimally by JOR, SOR, GRF, etc. as in Chapters 8 - 10 and also by improved equation orderings. The aim of such acceleration is simply to reduce the work done in obtaining RF instruments and thus the total computation in any estimator using those instruments. The results are likely to be dramatic improvements in convergence rates in instrument iteration. Several empirical examples are provided in the next chapter. Generally the increased work in obtaining such reduced form instruments is less than proportional to any reductions in estimation iterations. We may aid these improvements by relaxing the tolerance on instrument formation or using QRF instruments according to some rule.

The superimposition of semi-iterative acceleration to maximise the average rates of convergence assumes a particular importance here. Maximising average rates on a weak convergence criterion at each value of S may make a significant improvement overall to instrument iteration, and hence work done within estimation iteration, rather than optimising only on asymptotic rates. It is rather more easy to use semi-iterative methods than optimal JOR, SOR since then only two extreme eigenvalues need be calculated at each step S. Further convergence is obtained in (31) provided the sequence $\lambda^{(s+1)} = \rho(I - \hat{\beta}^{(s+1)}) < 1$ all S. However, we have improvements by theorems 8.5 and 8.10 which may require considerable effort to find a suitable α. More important still is

Lemma 1:

By Theorem 10.8 (also Theorem 10.2 but this is inconvenient to implement) iteration (31) is always convergent if semi-iterative acceleration is used. #

We note this lemma applies to linear or nonlinear systems. It is obvious that when the model specification is nonlinear this lemma holds, as it does for RF instruments by (31). RF instruments are therefore available in all circumstances.

Decision 7:

Finally there is the decision on accelerations affecting the average rates of convergence in estimation iteration. This decision is in a somewhat ambiguous position since it is dependent on the previous decisions having been taken and so must come last. However in view of the analysis of semi-iterative methods of the previous chapter if accelerations to optimise average rates are superimposed they must be done by semi-iterative methods. Thus freedom to do so or not is maintained provided the previous decisions have been made. The computational burdens imposed by previous decisions will be in no way affected if semi-iterative estimators are used. The importance of this decision is clear from Theorem 14.1 and the discussion following in Section 14.3.

11.6.1 The availability criterion

The likelihood function (2.35) requires the nonsingularity of $D'X'XD = T^{-1}\Sigma$ i.e. $T > K + M$ is required for the existence of both FIML and LIML estimates. However any of the iterative estimators work (subject to convergence criteria) provided

$$T > \max \{\max_i (r_i + s_i), M\} \qquad \text{or } T > \max_i \{r_i + s_i\} \tag{35}$$

However for endogenous iteration $(Z_i'Z_i)^{-1}$ and $(Y_i^{(s)}M_iY_i^{(s)})^{-1}$ must exist so (35) may be relaxed to

$$T > \max_i (\max (r_i, s_i) \tag{36}$$

which can be somewhat less. Note $\max_i r_i < M$ so for FI versions $T < M$ is no doubt effective.

Corollary to Chapter 5:

An IV interpretation of ML estimators always exists and the fixed point of (5.1), (5.10) is always available given (36). The reverse

interpretation of maximum likelihood for iterative estimators is approximate unless T > K + M; giving equivalent estimates if T < K + M. #

We note this corollary refutes the claim by Wold (1965) that all conventional simultaneous estimators break down if T ≤ K,M.

11.6.2 Relaxation of assumptions:

Chapter 2.2.4 detailed the relaxations possible in the assumptions made for model (2.5). In particular the relaxation of the sample size required for the estimator to be available is of interest. The absence of collinear relations in the predetermined variables is also relaxed - $(Z_i'Z_i)^{-1}$ is assumed to exist, rather than $(Z'Z)^{-1}$, for i = 1 .... M. (36) improves on the availability criterion.

11.7 A convention on naming estimators:

Having completed the analysis of the construction of iterative estimators in terms of 7 independent decisions, it is obvious that potentially there are a vast number of such estimators. It would seem sensible to have a convention on naming them, which includes those estimators already proposed and well known in the literature, so that not every estimator need be cited. More important the naming scheme should reflect the construction of the estimator. In listing the options under each decision, I propose the following letters should denote the taking of that option at that decision in the construction process.

Decision 1:

Least squares normal equations - FP, instrumental variables -IIV, if the limited information versions are used. If full information versions are used FIFP and FIIV respectively, if autoregressive version the prefix AR be attached.

Decisions 2 and 3:

Mixed iteration no prefix, parametric prefix P and endogenous prefix E. Nonlinear iterations prefixes N for Newton, S for Secant and CG for conjugate gradient. Letters from Decision 3 to precede those from Decision 2.

Decision 4:

Direct instrument formation no prefix, reduced form instruments RF prefix, and QRF to have prefix QRF and an explanation of the rule of k_s, e.g. QRFEFIFP with $k_s = k_0 - S$, say.

Decision 5:

A JOR acceleration in (29) the prefix F, recursive acceleration the prefix GS, a combination of both the prefix R_1, whereas if the parameters of each equation are re-estimated recursively also the prefix R (the distinction is Bodin's). Richardson's stationary and GRF methods would be mentioned by names as would SSOR and DI methods. For F, R_1 or R α^* is assumed unless (α = c) follows the prefix, for c = some constant.

Decision 6:

Any accelerations here would have the same letters as under Decision 5 but they are put in brackets between the prefixes of Decisions 5 and 7.

Decision 7:

The prefix SI is used if the semi-iterative accelerations are used.

The letters should be put together from right to left, as in Arabic script, with the decisions taken in the order specified. We note that elsewhere the prefix S has been used to indicate the fractional version of IIV (Bergstrom). This prefix is not consistent with FP nomenclature and is dropped to prevent confusion. Also IIV has been used to indicate what, in this scheme, is RFIIV. It is important to distinguish the two since the classification scheme is incomplete otherwise. Indeed the two methods have radically different convergence properties. We therefore need separate names for the two methods.

CHAPTER 12: CONVERGENCE ANALYSIS OF ITERATIVE ESTIMATORS

12.1 Introduction

This chapter is concerned with the very important problem of convergence conditions for iterative estimators. The theory of iterative estimation demands a knowledge of convergence conditions for each estimator. Indeed their proper use turns on that knowledge. Convergence is determined by factors which are a priori unknown, which is a stumbling block to their optimal use - but we can prove that certain iterative estimators are always convergent. Here we consider convergence conditions of various estimators to a given fixed point. Plural fixed points are analysed later. Thus iterative estimators are always available, and constitute a real and general alternative to conventional methods. Finally we extend these results to autoregressive, full information and autoregressive - full information versions of each iterative estimator.

In the subsequent chapter we will then be able to compare any iterative estimator to any other and a family of iterative estimators as a whole to conventional methods. Recommendations will be possible on which estimator is preferable in any given model.

12.2 A Framework for general convergence analysis

12.2.1

The convergence analysis of a general iterative process is well documented. We have already noted that both (5.1) and (5.10) imply nonlinear iterative processes for any model. The iterative process works in the variable indicated by the iteration mode, e.g. $Y^{(s)}$ for endogenous or $\hat{\Pi}^{(s)}$ for parametric. It is obvious that $\hat{\delta}^{(s)}$ converges iff $Y^{(s)}$ converges, and thus since in linear models $Y^{(s)} = X\,\Pi^{(s)}$ convergence analysis may be conducted either in terms of $Y^{(s)}$, or $\Pi^{(s)}$.

The general nonlinear iteration in the T x M matrix $Y^{(s)}$, with respect to S only, corresponding to (11.9), (11.15a), (11.39) but in the familiar notation of (11.31) is

$$Y^{(s+1)} = F(Y^{(s)}, Z) \tag{1}$$

$F(\cdot)$ is a T x (M + K) matrix valued nonlinear functional for the $(s+1)^{th}$ iteration. The particular form of $F(\cdot)$ depends on the specification of the model in question, the instrument formation method (the state of k_s in (11.31) and the iteration mode (e.g. endogenous iteration implies Z is not an argument of $F(\cdot)$). $F(\cdot)$ is a nonlinear mapping in R^M with domain D; $F: D \varepsilon R^M \rightarrow R^M$. It is clear that $F(\cdot)$ will have stochastic values and for the form of $F(\cdot)$ is likely to be stochastic itself. Therefore we use the following definitions:

Definition 1:

A global convergence criterion is a condition for the convergence of (1) to a fixed point Y* where Y* = F(Y*,Z) for any arbitrary start $Y^{(0)}$.

Definition 2:

A local convergence criterion is a condition for the convergence of (1) for any arbitrary start $Y^{(0)}$ restricted to be within some neighbourhood of the fixed point Y*.

The well known theorem of Ostrowski (1960) is the basis for the convergence analysis of (1).

Theorem 1:

If $\exists$ a fixed point Y* in the interior of D corresponding to $F(\cdot)$ of (1) and $F(\cdot)$ is continuous and differentiable at Y* then (1) will converge to Y* for an arbitrary $Y^{(0)}$ start within a neighbourhood of Y* if $\rho(F'(Y^*)) < 1$ where $\partial F/\partial Y = F'(Y)$ and $J = F'(Y^*)$ is the Jacobian of $F(\cdot)$ evaluated at Y*. #

J is a square matrix of order MT so for even the smallest models it is impractical to use Theorem 1 as it stands. A major objective in the analysis is to transform J to a matrix of a size which makes it feasible to use the criterion of theorem 1. We note that theorem 1 implies an approximate criterion since if we write $Y^{(s)} = Y^* + E^{(s)}$, then we can approximate (1) by

$$Y^{(s+1)} = Y^* + J|_{Y^*}(E^{(s)}) + ... \qquad (2)$$

with $E^{(s)} = (Y^{(s)} - Y^*)$, from a Taylor's expansion about Y^*. Thus for $Y^{(s)}$ sufficiently close to Y^* the higher order terms in (2) may be ignored. Theorem 1 is a local condition for convergence where $Y^{(0)}$ is sufficiently close to Y^* for (2) to be an approximation to (1) (thus defining the neighbourhood as the set of such $Y^{(0)}$). The local nature of theorem 1 consists of two parts. If the iteration matrix is not independent of S, convergence then requires $\rho(F'(Y^*)) < 1$. In addition a simple convergence criterion results if the first order approximation to (1) in (2) is taken, and for $Y^{(p)}$ is sufficiently close to Y^*, to make that approximation reasonable, we may replace the requirement $\lim_{S\to\infty} \prod_{S=p}^{\infty} F'(Y^{(s)}) = 0$ by $\rho(F'(Y^*)) < 1$. We should therefore define

Definition 3:

An exact convergence criterion is one which is derived from (1) directly or the infinite sum implied in (2).

Definition 4:

In approximate criterion is one which follows from the first order approximation in (2).

For some $Y^{(0)}$ definitions 2 and 4 are equivalent. Criteria which are global but approximate are possible, (e.g. sine functions of small amplitude) where it is necessary that the linear approximation is adequate

over the whole domain. Definitions 3 and 4 are introduced in contrast to any criterion obtained from an approximation to an iteration matrix itself obtained under one of the definitions 1 - 4. Such a criterion will be termed imprecise.

We assume the existence of Y* until theorem 25, and for reasons of simplicity alone we will assume until further notice that Y* is unique. Formally we assume the functional matrix $F(\cdot)$ in (1) in Lipschitzian[1] in each component. The assumption will be dropped later when multiple fixed points are considered.

Three consequences follow:

Corollary 1:

Theorem 1 is a global criterion and exact if (1) is linear. Then $J = B$ of (8.5). #

Many measures for rates of convergence have been proposed for non-linear iterations (see Ortega and Rheinboldt (1970))) but in view of theorem 1 following from a linear approximation it is justified to retain the use of those proposed for linear iterations (Chapter 8) within this neighbourhood of Y*.

Corollary 2:

(2) fits the linear approximation which introduced the analysis of linear iterations, so within the nighbourhood required by theorem 1 we can apply the results of Chapters 8 - 10. Regarding each column of $E^{(s)}$, $e_i^{(s)}$, as the error vector of step S, application of those results follows directly from $||e_i^{(s)}||_r < \tau_i \quad i = 1 \ldots M$; $||E^{(s)}||_r < \tau$, for tolerances τ_i, $\tau = \max_i(\tau_i)$ and a suitable norm being implied for convergence since

---

1. $F(\cdot)$ satisfies Lipschitz conditions at a point $\bar{X}$ if for some scalar p $|F(X) - F(\bar{X})| < p(|X - X|)$ for all X within a neighbourhood of $\bar{X}$. If all partial derivatives are continuous at each point of a closed region then it satisfies Lipschitz conditions for that region (Morishima (1964)). Evidently by (2) the iteration function is Lipschitzian locally and we only need the simplification of unique Y* for any global results.

$E^{(s+1)} = J|_{Y*}(E^{(s)})$ is a result of (2). #

Corollary 3:

$J^{(s)} = \{\partial F_i / \partial y_j^{(s)}\}$ where $\partial F_i / \partial y_j^{(s)} = \{\partial F_{\ell i} / \partial y_{tj}^{(s)}\}$ is a submatrix $i, j = 1 \ldots M$; $\ell, t = 1 \ldots T$. Since $y_j^{(s)}$ is a function of $\hat{\beta}_j^{(s)}$ and $\hat{\gamma}_j^{(s)}$, $J^{(s)}$ is both stochastic and dependent on S by virtue of the iteration. $\rho(J^{(s)})$ is thus both stochastic and dependent on S. #

12.2.2

So far we have ignored any possible effects instrument iteration may have on the convergence of an iterative estimator. If an iterative estimator converges for any instrument formation method, we may expect convergence conditions with respect to S and k to be related. But there is no necessity that an estimator will be convergent for all instrument formation methods in a given mode.

We can utilise Bodin's (1974) demonstration that the convergence criterion for an iterative estimator may be decomposed into two parts. Our purpose is more than just that demonstration. The problem is using (11.31) to obtain QRF or RF instruments or indeed direct instruments is that there will always be an error $\eta^{(s,k_s)} = Y^{(s,c_s)} - Y^{(s,k_s)} \neq 0$. We define the sequence c_s as the index for which (11.31) holds without error[1] (i.e. in (11.31) we have $c_s = k + 1$) and k_s as the index at which instrument iteration stops (i.e. $k_s = \max(k)$; $k_s = c_s$ for RF instruments; $k_s = 1$ for direct instruments) for each S. Applying theorem 2.12, full asymptotic efficiency and consistency of any iterative estimator at step S depends a valid instrumental variable interpretation and plim $T^{-1} N^{(s)'} X = X^{*'} Y$ with instrument $N^{(s)}$. In fact[2] we will have

---

1. Strictly $c_s = \infty$ $S = 1 \ldots$ but with a finite number of significant figures in the computer $c_s < \infty$ always holds. Then $Y^{(s,c_s)}$ is equivalent to the inversion solution (4.36).

2. for (5.4) we need plim $T^{-1} N^{(s)'} X = X^{*'} \Sigma^{-1} X$ instead.

$$\text{plim } T^{-1} N^{(s)'} X = \begin{bmatrix} Y^{*'}Y + \eta^{(s,k_s)'} Y & Y^{*'}Z \\ Z'Y & Z'Z \end{bmatrix} \neq X^{*'}X \qquad (3)$$

even if plim $T^{-1}\ \eta^{(s,k_s)'}Z = 0$ and plim $\hat{\Pi}^{(s)} = \Pi$ which remains to be proved.

Now $\eta^{(s,k_s)} = 0$ iff $Y^{(s)}$ is calculated by inversion (4.36) or if $k_s = \infty$. The second case is impractical; iteration is an infinite solution procedure but $\eta^{(s,k_s)}$ may be arbitrarily small and insignificant within rounding errors, measurement and misspecification biases, if k_s is large. The first case is unavailable for nonlinear models and in any case inevitably there will be rounding errors in the inversion when it is used. Either inversion or full iteration may in many cases be such an undesirable computational burden. $\eta^{(s,k_s)} \neq 0$ is likely in practice; its size will vary with k_s.

Theorem 2:

The convergence criteria of iterative estimators are invariant to the sequence k_s, if $\rho(I - \beta) < 1$ within the neighbourhood defined in theorem 1.

Proof:

From (11.31) we can write the general iteration

$$Y^{(s+1,c_{s+1})} - \eta^{(s+1,k_{s+1})} = Y^{(s+1,c_{s+1})}(I-\beta^{(s+1)}) - \eta^{(s+1,k_{s+1}-1)}(I-\beta^{(s+1)}) + Z\Gamma^{(s+1)} \qquad (4)$$

giving

$$Y^{(s+1,c_{s+1})} - \eta^{(s+1,k_{s+1})} = Y^{(s,c_s)}(I-\beta^{(s+1)})^{k_{s+1}} - \eta^{(s,k_s)}(I-\beta^{(s+1)})^{k_{s+1}} + Z\Gamma^{(s+1)} \sum_{i=0}^{k_{s+1}-1} (I-\beta^{(s+1)})^i \qquad (5)$$

where

$$Y^{(s+1,c_{s+1})} = Z\Gamma^{(s+1)} [\beta^{(s+1)}]^{-1}, \text{ by definition,} \qquad (6)$$

which are iterates in the Z-space. (5) may be regarded as two iterations

(a)

$$\eta^{(s,k_S)}=\eta^{(0,k_1)}\prod_{i=2}^{S}[I-\beta^{(i)}]^{k_i} = \eta^{(0,0)}\prod_{i=1}^{S}[I-\beta^{(i)}]^{k_i} \quad k_i \geq 1 \tag{7}$$

where $\eta^{(0,0)} = Y^{(0,c_0)} - Y^{(0)}$. (8)

(b)

$$Y^{(s+1,c_{s+1})} = Y^{(s,c_S)}[I-\beta^{(s+1)}]^{k_{s+1}} + Z\Gamma^{(s+1)}\sum_{i=0}^{k_{s+1}-1}(I-\beta^{(s+1)})^i \tag{9}$$

where $Y^{(s,c_S)}$ is generated by (6).

A sufficient condition for the convergence of (5) is $\rho(I - \beta^{(s+1)}) < 1$, all $S \geq 1$, since then (7) is convergent and also

$$||Y^{(s+1,c_{s+1})}|| \leq ||Y^{(s,c_S)}||\ ||I - \beta^{(s+1)}||^{k_{s+1}}$$

so

$$||Y^{(s,c_S)}|| \geq ||Y^{(s+1,c_{s+1})}|| \tag{10}$$

from (9). By theorem 1, the local sufficient condition becomes $\rho(I-\beta) < 1$. Then the iteration (i) will converge in all cases when the basic iterative estimator in $Y^{(s,c_S)}$ is convergent. A necessary and sufficient condition for the convergence of (5) will also include

$$\lim_{S\to\infty}[Y^{(s,c_S)}(I - \beta^{(s+1)})^{k_{s+1}} - \eta^{(0,0)}\prod_{i=1}^{S}(I - \beta^{(i)})^{k_i}] = 0 \tag{11}$$

as an alternative to (10).

[Finally if $c_s = k_s = \infty$ then $\eta^{(s,k_S)} = 0$ all $S \geq 1$ but the sufficient condition remains by (10)]. The theorem now follows. #

Convergence is unaffected by instrument formation methods under the conditions of theorem 2. Thus:

Corollary 4:

Rounding errors have no effect on convergence if $\rho(I - \beta) < 1$ (see also Ostrowski (1960)). #

Corollary 5:

If instrument iteration (11.31) is accelerated (at least with respect to k) by JOR or SOR methods Theorem 2 holds if $(I - \beta)$, or $F'(Y^*)$, has all roots with real part less than unity. #

We note that accelerating instrument formation may make $\eta^{(s,k_s)}$ very small for a moderate sized k_s. However, the important point from theorem 2 is that convergence in an estimator not using RF instruments always implies $\lim_{S\to\infty} \eta^{(s,k_s)} = 0$ and convergence to the same Y^*. We conclude therefore

Corollary 6:

If any estimator converges under these assumptions, consistency and asymptotic efficiency, appropriate to that information class, follow.

Proof:

Viewing the limit of (3) as $S \to \infty$. #

12.2.3

Further results follow for estimators without RF instruments at steps before convergence. The reader is referred to theorems 4.8 and 4.9. Here we treat inconsistent starts and also the efficiency of estimators without RF instruments to complete the theory of Chapter 4.7. RF instruments with numerical errors of course are included in this analysis.

Theorem 3:

The estimates at step S from such estimators are consistent but not asymptotically efficient within their information class iff Z is a basis for $Y^{(0)}$. Asymptotic efficiency is attained at convergence (i.e. $S = \infty$). Conversely if Z is not a basis for $Y^{(0)}$, consistency and asymptotic efficiency are obtained only at convergence; previous iterates have neither property. Iterates using RF instruments have both properties for

all $S \geq 2$ (or 3 depending on the information class).

Proof:

By (6) plim $T^{-1}Y^{(s,c_s)'}u = 0$ all $S \geq 1$. By (9) plim $T^{-1}Y^{(0,c_0)'}u = 0$. Consistency in the IIV family follows since

$$\text{plim } T^{-1}(Y^{(s,k_s)'}u) = \text{plim } T^{-1}Y^{(s,c_s)'}u + \text{plim } T^{-1}\eta^{(s,k_s)'}u = 0 \quad (12)$$

whenever

$$\text{plim } T^{-1}\left[\prod_{i=1}^{S+1}(I - \beta^{(i)})^i\right]' (Y^{(0,c_0)} - Y^{(0)})'u = 0 \quad (13)$$

which is satisfied if Z is a basis for $Y^{(0)}$.

In the FP family therefore we need in addition to the above

$$\text{plim } T^{-1}(X^{(s,k_s)'}X^{(s,k_s)}) = \text{plim } T^{-1}(X^{(s,k_s)'}X) \quad (14)$$

for consistency. This follows since

$$Y^{(s,k_s)'}Y^{(s,k_s)} = Y^{(s,k_s)'}Y + \eta^{(s,k_s)'}\hat{V}^{(s,k_s)} - Y^{(s,c_s)'}\hat{V}^{(s,k_s)} \quad (15)$$

and plim $T^{-1}\eta^{(s,k_s)'}V^{(s,k_s)} = \text{plim } T^{-1}\left[\prod_{i=1}^{S+1}(I-\beta^{(i)})^i\right]'$

$$.(Y^{(0,c_0)} - Y^{(0)})' M^{(s,k_s)}(I-\beta^{(s+1)})^{-1} = 0 \quad (16)$$

where

$$M^{(s,k_s)'} = I - X(X^{(s,k_s)'}X^{(s,k_s)})^{-1}X^{(s,k_s)'}.$$

Also

$$\text{plim } T^{-1}Z'Y^{(s,k_s)} = \text{plim } T^{-1}Z'Y \quad (17)$$

since plim $T^{-1}Z'V^{(s,k_s)} = 0$. But this equivalent IV estimator is consistent.

Assuming $Y^{(0)}$ has Z as a basis, asymptotic efficiency follows by theorem 2.12 iff plim $T^{-1}\eta^{(s,k_s)'}X = 0$. Then plim $T^{-1}Y^{(s,k_s)'}Y = \text{plim } T^{-1}Y^{*'}Y$ but also

$$\text{plim } T^{-1}Y(s,k_S)'Y=\text{plim } T^{-1}Y(s,c_S)'Y+\text{plim } T^{-1}\eta(s,k_S)'Y=\text{plim } T^{-1}Y^{*\prime}Y+\text{plim}T^{-1}\eta(s,k_S)'Y \tag{18}$$

$$\text{Hence plim } T^{-1}\eta(s,k_S)'Y = 0 = \text{plim } T^{-1}\eta(s,k_S)'Z \tag{19}$$

is required. Conversely if Z is not a basis for $Y^{(0)}$,

$$\text{plim } T^{-1}Y(s,k_S)'u =-\text{plim } T^{-1}\eta(s,k_S)'u=-\text{plim } T^{-1}\left[\sum_{i=1}^{S+1}(I-\beta^{(i)})^i\right]'Y^{(0)\prime}u \neq 0 \tag{20}$$

barring consistency in either family. Consequently, since, unless $S = \infty$, and $\rho(I - \beta) < 1$ and $Y^{(s,k_S)}$ is within a neighbourhood of Y^*, plim $T^{-1}Y(s,c_S)'Y$ = plim $T^{-1}Y^{*\prime}Y$, plim $T^{-1}\eta(s,k_S)'Y \neq 0$ barring asymptotic efficiency.

However we have by corollary 6 above $\lim_{S\to\infty} \eta^{(s,k_S)} = 0$ if an estimator converges and does not use RF instruments and $\eta^{(s,c_S)} = 0$, every S, by definition of c_S if it does use them; the properties at convergence and the last statement are proved. #

Comment:

Since convergence involves $S < \infty$, the properties mentioned for convergent estimators are only virtual properties.

Theorem 3 provides an important qualification, and some additions, to the results of Chapters 4 and 5. The problems caused by the nonzero errors $\eta^{(s,k_S)}$ are very different than those suggested by Brundy and Jorgenson (1974). Indeed the consequences may be summarised in

Corollary 7:

(a) the errors $\eta^{(s,k_S)}$ make no difference, in any circumstances where plim $T^{-1}Z'U = 0$, to the statistical properties of any estimator at convergence;

(b) the objective in convergence, where RF instruments are not used, is the attainment of these properties whatever the arbitrary start, as well as independence of start values and a maximum likelihood interpretation; more important is the attainment of asymptotic efficiency.

(c) there will be a difference in small sample distributions and hence statistical properties between estimators with RF and other instrument formation methods, even within the same family. This is not obvious since the same estimation phase is used for a given $Y^{(0)}$ within a family and estimators only vary with the iterative algorithm. However for a given $Y^{(0)}$ we have established the differences to be negligable at convergence because $\eta^{(s,k_s)}$ are then negligable within a tolerance. For previous iterates the difference will be more pronounced. Within a family, small sample distributions differ to the extent different algorithms introduce different rounding errors into the instrument process.

(d) There is a statistical reason for a differentiation of RF instrument estimators from the others - as well as the computational distinction already suggested. This is emphasised from now on.

(e) The selection of $Y^{(0)}$ has statistical consequences, where RF instruments are not used. A general rule is $Y^{(0)}$ be some (any) linear combination of Z or a subset of it. If this is not done an estimator cannot be stopped before convergence without the loss of the desired statistical properties.

The results of Theorem 3 are an important supplement to the general theory of Chapters 5 and 6. They appear to be new - not only has the structure of different estimators been unclear in previous studies but also there has been no consideration of the possible differences in statistical properties between iterative estimators. Further we have the first rule in selecting suitable starts (apart from guessing $Y^{(0)}$ close to Y^* implied by Theorem 1 and Corollary 4.4). We observe that errors $\eta^{(s,k_s)}$, at least in probability limits are in the orthogonal complement to both the Z and U spaces. It follows therefore that they may be interpreted as a sequence of disturbances in the iterative process either due to $Y^{(0)}$ being selected partly out of the Z space or due to rounding errors out of either space introduced during iterations. Indeed the above may

also be treated as a general analysis of the effects of rounding errors in the iterative process, which has not been carried out before. However these rounding errors have the property $\lim_{T\to\infty} \mathrm{Var}(\eta^{(s,k_S)}) = 0$ - not all rounding errors may have this property in which case standard measurement error results of regression theory apply.

12.2.4

To derive convergence criteria for estimation iteration we consider the fixed point in a neighbourhood given by

$$||Y^{(s)} - Y^*||_r \leq \Delta_1, \quad \Delta_1 > 0 \text{ and } ||\eta^{(s,k_S)}||_r \leq \Delta_2, \quad \Delta_2 > 0, \qquad (21)$$
$$1 \leq r \leq \infty$$

so

$$||Y^{(s,k_S)} - Y^*||_r \leq \Delta_1 + \Delta_2 \qquad (22)$$

for $Y^{(s)}$ any general iteration with "error" $\eta^{(s,k_S)}$.

Then by Theorem 1 convergence follows if the transformation from step S to S + 1 results in

$$\rho \begin{bmatrix} \dfrac{\partial Y^{(s+1)}}{\partial Y^{(s)}} & \vdots & \dfrac{\partial Y^{(s+1)}}{\partial \eta^{(s,k_S)}} \\ \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ \dfrac{\partial \eta^{(s+1,k_{S+1})}}{\partial Y^{(s)}} & \vdots & \dfrac{\partial \eta^{(s+1,k_{S+1})}}{\partial \eta^{(s,k_S)}} \end{bmatrix}_{Y^*} < 1. \qquad (23)$$

Obviously the instrument iteration need not affect convergence if $\rho(I-\beta)<1$ or if, say, semi-iterative iteration is used (theorem 10.8) (although it may affect the speed of convergence of the estimator) so we can sublimate it within the estimation iteration and separate out the "error" component when contrasted with a basic iteration, say (11.15a). In Theorems 2 and 3 we established that for consistent $Y^{(0)}$, $\partial Y^{(s+1)}/\partial \eta^{(s,k_S)} = 0 = \partial \eta^{(s+1,k_{S+1})}/\partial Y^{(s)}$ in probability limit all S (and for all T at convergence) when evaluated at Y^*.

We note from (7) $\partial\eta^{(s+1,k_{s+1})}/\partial\eta^{(s,k_s)} = I - \beta^{(s+1)}$; so by (23) $\rho(I - \beta) < 1$ is required since a block diagonal matrix has eigenvalues equal to those of the nonzero blocks. It remains to determine $[\partial Y^{(s+1)}/\partial Y^{(s)}]_{Y*}$ for various estimators so that conditions for (23) are explicit. We note where $k_s = c_s$, $\eta^{(s,c_s)} = 0$ so $\rho[\partial Y^{(s+1)}/\partial Y^{(s)}]$ is by itself decisive; the same is true for any error free estimator at Y^* whatever instruments.

Two results of Bodin will be employed.

Theorem 4:

The submatrix $\{\partial F_i/\partial y_j^{(s)}\} = \beta_{ij}^{(s)} \otimes I_T$ if $u_i = 0$ in relation i.

Proof:

Theorem 5.6 Bodin- replace Π_j by $y_j^{(s)}$ in my notation. #

Theorem 5:

The submatrix $\{\partial F_i/\partial y_j^{(s)}\} = 0$ if relation is is exactly identified.

Proof:

Theorem 5.7 Bodin; replace Π_j by $y_j^{(s)}$ again. #

Theorem 6:

The number of nonequivalent orderings in any recursive estimator is at most $M^*!/2$ where M^* is the number of overidentified relations.

Proof:

Construct a position matrix P such that $p_{ij} = \begin{cases} 1 \text{ if } \partial F_i/\partial y_j^{(s)} \neq 0 \\ 0 \text{ otherwise} \end{cases}$

The theorem follows from Theorem 5 using P = B in Theorem 8.8. Bodin also demonstrates part of this result for RPFP. Naturally Theorem 8.8 applies to a full B matrix and it will be confirmed below that $p_{ij} \neq 0$ iff $\beta_{ij} \neq 0$. Nevertheless it may be that, even so, more equivalent orderings are possible after the removal of $M - M^*$ zero rows. #

12.3 Convergence conditions

Using the analysis of the previous section we obtain the convergence conditions for various iterative estimators to the given fixed point. From Theorem 1, we consider an arbitrary start $Y^{(0)}$ within a neighbourhood of Y^*. We start from (23). In each case $\partial\eta^{(s+1,k_{s+1})}/\partial\eta^{(s,k_s)} = (I - \beta)$ at Y^* unless $\eta^{(s,k_s)} = 0$ applies. Error control therefore assumes that instrument iterations employ, where necessary, a convergent acceleration which overrides the additional requirement $\rho(I - \beta) < 1$ which would otherwise apply. Then $\rho(\partial Y^{(s+1)}/\partial Y^{(s)})$ itself is decisive. We consider only limited information methods at this stage.

12.3.1 Fixed point methods:

For any FP based method, under the current nonlinearity assumption, (11.18) holds. Thus for FP on the i^{th} equation

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} = M_i \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \beta_i^{(s+1)} + M_i Y_i^{(s)} \frac{\partial \beta_i^{(s+1)}}{\partial y_{tj}^{(s)}} \tag{24}$$

from (11.19), but from (11.18)

$$\frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}} = (Y_i^{(s)'} M_i Y_i^{(s)})^{-1} \frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}} M_i (y_i - Y_i^{(s)} \hat{\beta}_i^{(s+1)}) - (Y_i^{(s)'} M_i Y_i^{(s)})^{-1} Y_i^{(s)'} M_i \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} \tag{25}$$

New we note $y_i - Y_i^{(s)} \hat{\beta}_i^{(s+1)} = Z_i \hat{\gamma}_i^{(s+1)} + \hat{v}_i^{(s+1)}$ and $M_i Z_i = 0$. Since

$\hat{v}_i^{(s+1)} = p_i^{(s)} v_i$ where $p_i^{(s)} = I - X_i^{(s)} (X_i^{(s)'} X_i^{(s)})^{-1} X_i^{(s)'}$ by (4.27)

and

$$Z_i = X_i^{(s)} \begin{bmatrix} 0 \\ \vdots \\ I \end{bmatrix} \text{ then } M_i \hat{v}_i^{(s+1)} = \hat{v}_i^{(s+1)} \text{ because } M_i P_i^{(s)} = P_i^{(s)}, i = 1 \ldots M \tag{26}$$

$$\therefore \quad \frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} = Q_i^{(s)} \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} +$$

$$T_{(i)}^{(s)} \frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}} \hat{v}_i^{(s+1)} \qquad i,j = 1 \ldots M \qquad t = 1 \ldots T \qquad (27)$$

where $Q_i^{(s)} = M_i(I - Y_i^{(s)}(Y_i^{(s)'} M_i Y_i^{(s)})^{-1} Y_i^{(s)'} M_i)$

and $T_{(i)}^{(s)} = M_i Y_i^{(s)}(Y_i^{(s)'} M_i Y_i^{(s)})^{-1}$

But $\frac{\partial y^{(s+1)}}{\partial y^{(s)}} = \frac{\partial Y^{(s+1)}}{\partial Y^{(s)}}$ has $\left\{\frac{\partial y_i^{(s+1)}}{\partial y_j^{(s)}}\right\}$ as the i,j^{th} submatrix;

$\frac{\partial y_i^{(s+1)}}{\partial y_j^{(s)}}$ has $\left\{\frac{\partial y_{ti}^{(s+1)}}{\partial y_{kj}^{(s)}}\right\}$ as the t,k^{th} element for $i,j = 1 \ldots M$ and $t,k = 1 \ldots T$ (corollary 3 above).

$$\text{Also } \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} = \left.\begin{array}{ll} \beta_{ji}^{(s+1)} & \text{if } \beta_{ji} \neq 0 \\ 0 & \text{otherwise} \end{array}\right\} \quad \text{each } t = 1 \ldots T$$

so for the first term on the right of (27) corresponding to the submatrix

$$\frac{\partial y_i^{(s+1)}}{\partial y_j^{(s)}} \text{ we have } \begin{cases} Q_i^{(s)} [\hat{\beta}_{ji}^{(s+1)} \otimes I_T] & \text{if } \beta_{ji} \neq 0 \\ 0 & \text{otherwise,} \end{cases} \qquad (28)$$

$i,j = 1 \ldots M.$

Next consider the second right hand side term of (27). Evidently (26) implies $T_{(i)}^{(s)'} \hat{v}_i^{(s+1)} = 0$ (an r_i column vector) for $i = 1 \ldots M$ since $Y_i^{(s)'} \hat{v}_i^{(s+1)} = [I : 0] X_i^{(s)'} P_i^{(s)} v_i$. Now with each

t, $T_{(i)}^{(s)} \dfrac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}} \hat{v}_i^{(s+1)} = T_{(i)j}^{(s)} \hat{v}_{ti}^{(s+1)}$ (an T-volumn vector) for $j=1....r_i$.

Hence corresponding to the submatrix $\dfrac{\partial y_i^{(s+1)}}{\partial y_j^{(s)}}$ we get

$$A_{ji}^{(s)} = \begin{cases} T_{(i)j}^{(s)} \hat{v}_i^{(s+1)'} & \text{for } \beta_{ji} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

$i = 1 .... M$ and $j = 1 .... r_i$.

Lemma 1:

$A_{ji}^{(s)}$ is nilpotent (i.e. has no nonzero eigenvalues).

Proof:

(a) By Dhrymes (1970) A.6.34 all nonzero eigenvalues of $A_{ji}^{(s)}$ are identical to those of $\hat{v}_i^{(s+1)'} T_{(i)j}^{(s)}$ which is the scalar zero. Hence $A_{ji}^{(s)}$ has T zero eigenvalues for all $i = 1 .... M$, $j = 1 .... r_i$ if $\beta_{ji} \neq 0$, and the same obviously if $\beta_{ji} = 0$ since it is a null matrix. #

Let $A^{*(s)}$, of order MT, have submatrices $A_{ji}^{(s)}$ $i, j = 1.... M$.

Now from (28) and (28a)

$$J^{(s)} = \frac{\partial Y^{(s+1)}}{\partial Y^{(s)}} = Q^{*(s)}[(I - \hat{\beta}^{(s+1)}) \otimes I_T] + A^{*(s)} \tag{30}$$

where $Q^{*(s)} = \text{diag } Q_i^{(s)}$. Notice the FP iteration is nonlinear as is evident from (30) with also $J^{(s)}$ a function of S. Therefore a global convergence criterion is not available. A local criterion may be obtained from theorem 1.

Theorem 7:

FP has the local convergence criterion $\rho(I - \beta) < 1$.

Proof:

$A^{*(s)}$ is nilpotent by lemma A.5 for all $S \geq 0$. By Lemma A.6 $J^{(s)}$ is convergent iff $\rho(Q^{*(s)}[(I - \hat{\beta}^{(s+1)}) \otimes I]) < 1$ and therefore the estimator is locally convergent in the sense of theorem 1 if[1] $\rho(Q[(I - \beta) \otimes I]) < 1$ by theorem 1. But Q is idempotent and symmetric; also $(I - \beta) \otimes I$ has eigenvalues equal to those of $I - \beta$ each repeated with multiplicity T (Anderson (1958)). Applying corollary A.3 the result follows. #

For RFFP we use (11.33a); then

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} - \frac{\partial Y_i^{(s+1)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} = \left(Y_i^{(s+1)} - (I - M_i)Y_i^{(s)}\right) \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}}$$

$$- (I-M_i) \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} \qquad (33)$$

We substitute in the first term of (33) using (25) which is unchanged except that in view of (11.31)

$$y_i - Y_i^{(s)} \hat{\beta}_i^{(s+1)} = Y_i^{(s+1)} \hat{\beta}_i^{(s+1)} - Y_i^{(s)} \hat{\beta}_i^{(s+1)} + Z_i \hat{\gamma}_i^{(s+1)} + v_i^{(s+1)}$$

$$= (V_i^{(s)} - V_i^{(s+1)})\hat{\beta}_i^{(s+1)} + v_i^{(s+1)} + Z_i\hat{\gamma}_i^{(s+1)} \qquad (34)$$

where again (26) and $M_iZ_i = 0$ hold. Thus we have the right hand side of (33) as

---

1. Q^*, A^* and $Q^{*(s)}$, $A^{*(s)}$ evaluated at Y^*

$$[Y_i^{(s+1)}-(I-M_i)Y_i^{(s)}][Y_i^{(s)'}M_iY_i^{(s)}]^{-1}\left[\frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}}M_i\left((V_i^{(s)}-V_i^{(s+1)})\hat{\beta}_i^{(s+1)}+v_i^{(s+1)}\right)\right.$$

$$\left.-Y_i^{(s)'}M_i\frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}}\hat{\beta}_i^{(s+1)}\right] - (I-M_i)\frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}}\hat{\beta}_i^{(s+1)} \qquad (35)$$

We shall evaluate $J^{(s)}$ for RFFP at Y^* so that within some neighbourhood of Y^*, it is appropriate now to take the approximation[1] to (35) given by $Y^{(s+1)} \simeq Y^{(s)}$ and $V^{(s)} \simeq V^{(S+1)}$ which leaves (35) as follows

$$T_{(i)}^{(s)}\frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}}\hat{v}_i^{(s+1)} - (I_T-Q_i^{(s)})\frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}}\hat{\beta}_i^{(s+1)} \qquad (36)$$

where now

$Q_i^{(s)} = (M_iY_i^{(s)}(Y_i^{(s)'}M_iY_i^{(s)})^{-1}Y_i^{(s)'}-I)M_i$ and $T_{(i)}^{(s)}$ is as before, giving $A^{*(s)}$ as before. We can write, with $Q^{*(s)} = \text{diag}\,(Q_i^{(s)})$,

$$\hat{\beta}^{(s+1)}\otimes I_T[\partial Y^{(s+1)}/\partial Y^{(s)}] = -(I_{MT}-Q^{*(s)})((I-\hat{\beta}^{(s+1)})\otimes I_T) + A^{*(s)} \qquad (37)$$

Thus

$$J^{(s)} = \frac{\partial Y^{(s+1)}}{\partial Y^{(s)}} = (-[\hat{\beta}^{(s+1)}]^{-1}\otimes I_T)(I_{MT}-Q^{*(s)})((I-\hat{\beta}^{(s+1)})\otimes I_T)+A^{*(s)} \qquad (38)$$

Theorem 8:

RFFP has local convergence criterion $\rho(-(I-\beta)\beta^{-1}) < 1$ and thus is convergent if $(I - \beta)$ has all roots $< \frac{1}{2}$ in real part.

---

1. The approximation is arbitrarily good for large enough S or a small enough neighbourhood.

Proof:

Apply lemma A.5 to $A^{*(s)}$, lemma A.6 and theorem 1 to $J^{(s)}$. Note $(I - Q^{*(s)})$ is idempotent. The nonzero eigenvalues of $-\beta^{-1} \otimes I_T(I_{MT}-Q^*)((I-\beta) \otimes I_T)$ equal those of $-(I_{MT}-Q^*)((I-\beta)\beta^{-1}\otimes I_T)$ (by Dhrymes (1970) A.6.34), $J^{(s)}$ having evaluated at Y^*. Application of corollary A.3 implies the result from $\rho(J) = \rho(-(I-\beta)\beta^{-1})$. If $I - \beta$ has eigenvalues λ_i then $\rho(J) = \max_i \frac{\lambda_i}{1-\lambda_i}$ so $\rho(J) < 1$ if $Re(\lambda_i) < \frac{1}{2}$ $i=1....M$. #

Comment:

Theorem 8 has provided an imprecise convergence criterion because an approximation was taken over and above that indicated by ignoring terms higher than first order in (2). It is an open question which of the two approxmations involves greater inaccuracy in any given case. Clearly theorem 1 requires the evalution of $J^{(s)}$ at Y^* in its local nature and so both approximations follow - but the reader should be aware of two limit-ations on the neighbourhood.

We now investigate the effects of parametric iteration on convergence conditions. First with direct instrument formation; take the form (11.19c).

$$\frac{\partial \hat{\Pi}_i^{(s+1)}}{\partial \hat{\Pi}_{jk}^{(s)}} = I_i^* \left(\frac{\partial (\hat{\Pi}^{(s)})_i}{\partial \hat{\Pi}_{jk}^{(s)}} \hat{\beta}_i^{(s+1)} + (\hat{\Pi}^{(s)})_i \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial \hat{\Pi}_{jk}^{(s)}} \right) \tag{39}$$

$$\begin{aligned} \frac{\partial \hat{\beta}_i^{(S+1)}}{\partial \hat{\Pi}_{jk}^{(S)}} &= [(\hat{\Pi}^{(S)})_i' Z' M_i Z (\hat{\Pi}^{(S)})_i]^{-1} \frac{\partial (\hat{\Pi}^{(S)})_i'}{\partial \hat{\Pi}_{jk}^{(S)}} Z' M_i (y_i - Z(\hat{\Pi}^{(S)})_i \hat{\beta}_i^{(S+1)}) \\ &\quad - [(\hat{\Pi}^{(S)})_i' Z' M_i Z (\hat{\Pi}^{(S)})_i]^{-1} (\hat{\Pi}^{(S)})_i' Z' M_i Z \frac{\partial (\hat{\Pi}^{(S)})_i}{\partial \hat{\Pi}_{jk}^{(S)}} \hat{\beta}_i^{(S+1)} \end{aligned} \tag{40}$$

from (25) writing $Y_i^{(s)} = Z(\hat{\Pi}^{(s)})_i$. The first term of (40) vanishes since $y - Z(\hat{\Pi}^{(s)})_i \hat{\beta}_i^{(s+1)} = Z_i \hat{\gamma}_i^{(s+1)} + \hat{v}_i^{(s+1)}$ and $M_i Z_i = 0$ so (26) holds and

$$\partial(\hat{\Pi}^{(s)})_i'/\partial\hat{\Pi}_{jk}{}^{(s)}Z'\hat{v}_i^{(s+1)} = \begin{cases} z_k'\hat{v}_i^{(s+1)} = 0 \text{ if } z_k \varepsilon Z_i \text{ since } Z_i'\hat{v}_i^{(s+1)} = 0 \\ 0 \qquad\qquad\qquad \text{if } z_k \notin Z_i \end{cases}$$

$$\frac{\partial\hat{\Pi}_i^{(s+1)}}{\partial\hat{\Pi}_{jk}^{(s)}} = I_i^*(I_K - (\hat{\Pi}^{(s)})_i[(\hat{\Pi}^{(s)})_i'Z'M_iZ(\hat{\Pi}^{(s)})_i]^{-1}(\hat{\Pi}^{(s)})_i'Z'M_iZ)(\hat{\beta}_{ki}^{(s+1)} \otimes I_K)$$

$$= I_i^* Q_i^{(s)}(\hat{\beta}_{ki}^{(s+1)} \otimes I_K) \qquad (41)$$

Theorem 9:

PFP has local convergence criterion $\rho(I - \beta) < 1$.

Proof:

$Q_i^{(s)}$ is idempotent and hence has s_i repeated eigenvalues $\lambda_k = 1$ and $K - s_i$ of $\lambda_\ell = 0$. For eigenvectors x_k corresponding to λ_k and x_ℓ for λ_ℓ, we have

$$I_i^*Q_i^{(s)}x_k = I^*x_k \quad k = 1 \ldots s_i \text{ and } I_i^*Q_i^{(s)}x_\ell = 0 \quad \ell = s_i + 1 \ldots K \qquad (42)$$

Premultiplication by I^* sets row j to zero where $\gamma_{ij} \neq 0$ but elsewhere the matrix is unchanged. Write x_k^*, x_ℓ^* as x_k and x_ℓ but with $x^*_{kj} = 0$ and $x^*_{\ell j} = 0$ and we have $(I_i^*Q_i^{(s)})x_k^* = x_k^*$ and $(I_i^*Q_i^{(s)})x^*_\ell = 0$. Hence $I_i^*Q_i^{(s)}$ has all eigenvalues zero or unity. From (41) evaluated at Y^*

$$J = \left.\frac{\partial\Pi^{(s+1)}}{\partial\Pi^{(s)}}\right|_{Y^*} = Q^*((I - \beta) \otimes I_K) \qquad (43)$$

where $Q^* = \text{diag}\,(I_i^*Q_i)$. The theorem follows by application of corollary A.3 and theorem 1, when (43) is used. #

If RF instruments were used in (11.19c) we would have $\hat{\Pi}_i^{(s+1)} = (\hat{\Pi}^{(s+1)})_i\,\beta_i^{(s+1)} + C_i^*y_i - C_i^*\,Z(\hat{\Pi}^{(s)})_i\,\hat{\beta}_i^{(s+1)}$. However (40) holds unchanged. (39) becomes

$$\frac{\partial \hat{\Pi}_i^{(S+1)}}{\partial \hat{\Pi}_{jk}^{(S)}} - \frac{\partial (\hat{\Pi}^{(S+1)})_i}{\partial \hat{\Pi}_{jk}^{(S)}} \hat{\beta}_i^{(S+1)} = [(\hat{\Pi}^{(S+1)})_i - C_i^* Z(\hat{\Pi}^{(S)})_i] \frac{\partial \hat{\beta}_i^{(S+1)}}{\partial \hat{\Pi}_{jk}^{(S)}}$$

$$- C_i^* Z \frac{\partial (\hat{\Pi}^{(S)})_i}{\partial \hat{\Pi}_{jk}^{(S)}} \hat{\beta}_i^{(S+1)} \qquad (44)$$

But $C_i^* Z[(\hat{\Pi}^{(S+1)})_i - (\hat{\Pi}^{(S)})_i] = C_i^*(\hat{V}_i^{(S)} - \hat{V}_i^{(S+1)}) = 0$

approximately within the neighbourhood sufficiently close to Y*. Also the first term on the right of (40) vanishes approximately in this neighbourhood since $y_i - Z(\hat{\Pi}^{(s)})_i \hat{\beta}_i^{(s+1)} = Z_i \gamma_i^{(s+1)} + v_i^{(s+1)} + Z[(\hat{\Pi}^{(s+1)})_i - (\hat{\Pi}^{(s)})_i]$ and we proceed as before with $\hat{\Pi}^{(s+1)} \simeq \hat{\Pi}^{(s)}$. The right hand side of (44) therefore becomes

$$-\{I_i^*(\hat{\Pi}^{(s)})_i[(\hat{\Pi}^{(s)})_i' Z' M_i Z(\hat{\Pi}^{(s)})_i]^{-1}(\hat{\Pi}^{(s)})_i' Z' M_i Z$$

$$+ (I_K - I_i^*)\} \frac{\partial (\hat{\Pi}^{(s)})_i}{\partial \Pi_{jk}^{(s)}} \hat{\beta}_i^{(s+1)} \qquad (45)$$

Introducing the a priori zero restrictions in $\hat{\beta}^{(s+1)}$ on the left of (44)

$$(\hat{\beta}^{(s+1)} \otimes I)(\partial \Pi^{(s+1)} / \partial \Pi^{(s)}) = - Q^{*(s)}((I - \hat{\beta}^{(s+1)}) \otimes I) \qquad (46)$$

where $Q^{*(s)} = \text{diag}\,(Q^*_i{}^{(s)})$; and

$$Q^*_i{}^{(s)} = I_K + I^*((\hat{\Pi}^{(s)})_i[(\hat{\Pi}^{(s)})_i' Z' M_i Z(\hat{\Pi}^{(s)})_i]^{-1}(\hat{\Pi}^{(s)})_i' Z' M_i Z - I_K) \qquad (47)$$

$$= I_K - I_i^* Q_i^{(s)} \text{ where } Q_i^{(s)} \text{ is defined in (41).}$$

Theorem 10:

RFPFP has local convergence criterion $\rho(-(I - \beta)\beta^{-1}) < 1$ and is convergent if $(I - \beta)$ has all roots $< \frac{1}{2}$ in real part.

Proof:

As in Theorem 9, $Q^{*}_{i}{}^{(s)}$ has all eigenvalues either zero or unity since $I_i{}^{*}Q_i{}^{(s)}$ has that property for any value of S. Hence $Q^{*(s)}$ also has that property. Evaluating (46) at Y*, $J = \partial\Pi^{(s+1)}/\partial\Pi^{(s)} = -\beta^{-1} \otimes I_K(Q^{*})((I-\beta) \otimes I_K)$. We can obtain the nonzero eigenvalues of J from those of $-Q^{*}((I-\beta)\beta^{-1} \otimes I_K)$, since application of corollary A.3 Dhrymes (1970) A.6.34 and Theorem 1 completes the proof. #

Comment:

(a) The remarks following theorem 8 on the two limitations on the neighbourhood, for which the imprecise convergence criterion of theorem 10 is appropriate, apply again here.

(b) The comments of Chapter A.3 do not apply to parametric iterations. Thus while schemes for optimising PFP and FP or RFPFP and RFFP are identical on asymptotic convergence criteria, they are different on average convergence criteria. Also the number of iterations to convergence is calculated as conventionally done (Chapter 8). We expect the same number of iterations but there will be cases when they differ considerably. Finally we note

Theorem 11:

Theorem 7 - 10 hold unaltered for endogenous iteration.

Proof:

Immediate in each case since either mixed or endogenous iteration the proofs follow from (11.18) and (11.19) or a reformed (11.18) and (11.19c). #

The convergence conditions for accelerated fixed point estimators follow immediately from the results of Chapter 8 and those above.

Theorem 12:

(i) Fractional fixed point versions of FP, PFP, EFP or EPFP have the local convergence criterion for some $\alpha > 0$ the real part of all roots of $(I - \beta)$ be less than unity.

(ii) Fractional fixed point versions of RFFP, RFPFP, RFEFP, RFEPFP have the same local convergence criterion.

(iii) The optimisation and comparison theorems of Chapters 8,9 apply locally with J for B. Using Corollaries (i) and (ii) of Chapter A.3.

Proof:

(i) By Theorem 8.8 on $(I - \beta)$ following application of Theorem 1 on the criteria of Theorems 7, 9 and 11 each of which is evaluated at Y*.

(ii) As in (i) we apply Theorem 8.8 but to $-(I - \beta)\beta^{-1}$ from Theorems 8, 10, 11. Let $(I - \beta)$ have eigenvalues $a_j + ib_j$ $j = 1 \ldots M$ then

$$\frac{-\alpha(a_j + ib_j)}{(1-a_j)-ib_j} + 1 - \alpha < 1 \text{ iff } a_j < 1 \quad \text{for } j = 1 \ldots M.$$

(iii) Obvious in the context of theorem 1. #

Theorem 13:

(i) Recursive fixed point versions of FP, PFP, EFP, EPFP have local convergence criterion for some $\alpha > 0$, the real parts of all roots of $(I - \beta)$ be less than unity.

(ii) Recursive fixed point versions of RFFP, RFPFP, RFEFP, RFEPFP have the same local convergence criterion.

(iii) Any recursive fixed point estimator is convergent globally only if $0 < \alpha < 2$.

(iv) The optimisation and comparison theorems of Chapters 8, 9 apply locally with J for B. Using Corollaries (i) and (ii) of Chapter A.3.

Proof:

(i) Application of Theorem 1 followed by Theorem 8.10 on the criteria of Theorems 7, 9, 11 each of which is evaluated at Y*.

(ii) As in (i) to Theorems 8, 10, 11. The same steps as Theorem 12 (ii) are necessary to obtain the criterion in terms of the eigenvalues of $(I - \beta)$ rather than those of $(I - \beta)\beta^{-1}$.

(iii) Immediate from Theorem 8.5 - this is a global property.

(iv) Obvious in the context of theorem 1. #

Theorems 12 and 13 evidently may be rephrased: $\exists$ some $\alpha > 0$ for which the estimators mentioned are convergent iff no root of $(I - \beta)$ exceeds unity.

We note all Fractional and Recursive fixed point estimators have the same local convergence criteria. To the extent $Y^{(s)} \neq Y^*$ the criterion for any method is appropriate only within a neighbourhood. This is true for these methods in particular but for the range of α values for which a fractional or recursive fixed point estimator is convergent, given $Y^{(s)} \neq Y^*$ within the neighbourhood, is not clear. The criterion takes on an additionally local nature since for a given α value the neighbourhood may effectively be reduced.

Evidently from Chapter 10 accelerations based on GRF methods or any semi-iterative method can always be made convergent. If these accelerations are used as suggested in Chapter 14 then the estimator will be convergent at any S value, whatever the start. Thus for these estimators we can get global convergence (though possibly to a local solution). We treat other estimators which are globally convergent methods in the next section.

Meanwhile

Theorem 14:

Under a GEID specification the local convergence criteria in Theorems 7 - 13 are unchanged.

Proof:

In all of those theorems the proofs may be repeated. Nothing more restrictive than $X_i^{(s)'} v_i^{(s+1)} = 0$ has been noted; a property in both REID and GEID systems. #

12.4 Convergence Conditions - IIV

All IIV estimators differ from their FP counter parts only in the normal equation system. We have (11.15) and (11.17) in place of (11.18) and (11.19) for example. We therefore replace (25) by

$$\frac{\partial \hat{\beta}_i^{(S+1)}}{\partial y_{tj}^{(S)}} = (Y_i^{(S)'} M_i Y_i)^{-1} \frac{\partial Y_i^{(S)'}}{\partial y_{tj}^{(S)}} M_i (y_i - Y_i \hat{\beta}_i^{(S+1)})$$

$$= (Y_i^{(S)'} M_i Y_i)^{-1} \frac{\partial Y_i^{(S)'}}{\partial y_{tj}^{(S)}} \hat{u}_i^{(S+1)} \qquad (48)$$

since $y_i - Y_i \hat{\beta}_i^{(S+1)} = Z_i \hat{\gamma}_i^{(S+1)} + \hat{u}_i^{(S+1)}$ and $M_i Z_i = 0$.

Moreover

$$\hat{u}_i^{(s+1)} = [I - X_i (X_i^{(s)'} X_i)^{-1} X_i^{(s)'}] u_i \text{ so } (I - M_i)\hat{u}_i^{(s+1)} = 0 \qquad (48a)$$

we replace (40) by

$$\frac{\partial \hat{\beta}_i^{(S+1)}}{\partial \hat{\Pi}_{jk}^{(S)}} = [(\hat{\Pi}^{(S)})_i' Z' M_i Y_i]^{-1} \frac{\partial (\hat{\Pi}^{(S)})_i'}{\partial \hat{\Pi}_{jk}^{(S)}} Z' M_i (y_i - Y_i \hat{\beta}_i^{(S+1)})$$

$$= ((\hat{\Pi}^{(S)})_i' Z' M_i Y_i)^{-1} \frac{\partial (\hat{\Pi}^{(S)})_i'}{\partial \hat{\Pi}_{jk}^{(S)}} Z' \hat{u}_i^{(S+1)} \qquad (49)$$

for similar reasons.

Theorem 15:

The local convergence criterion of IIV is $\rho(I - \beta) < 1$.

Proof:

From (11.17) we have

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} = \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} - (I - M_i)y_i \frac{\hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}} + \frac{Y_i^{(s)} \partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}}$$

Combining this with (48) gives

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} = \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} + [Y_i^{(s)} - (I-M_i)Y_i](Y_i^{(s)'} M_i Y_i)^{-1} \frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}} \hat{u}_i^{(s+1)}$$

Define $T_{(i)}^{(s)} = [Y_i^{(s)} - (I-M_i)Y_i](Y_i^{(s)'} M_i Y_i)^{-1}$ and note $T_{(i)}^{(s)'} u_i^{(s+1)} = 0$ i=1....M,

from $Y_i^{(s)'} \hat{u}_i^{(s+1)} = (I:0)X_i^{(s)'}[I - X_i(X_i^{(s)'} X_i)^{-1} X_i^{(s)'}]u_i = 0$ and (48a).

Define $A_{ji}^{(s)} = \begin{cases} T_{(i)j}^{(s)} \hat{u}_i^{(s+1)'} & \text{for } \beta_{ji} \neq 0 \text{ and } A^{*(s)} \text{ as in (30) i.e. we have} \\ 0 & \text{otherwise} \end{cases}$

replaced $\hat{v}_i^{(s+1)}$ with $\hat{u}_i^{(s+1)}$. $A^{*(s)}$ is nilpotent by lemmas 1 and A.5.

Then

$$J^{(s)} = \frac{\partial Y^{(s+1)}}{\partial Y^{(s)}} = (I - \hat{\beta}^{(s+1)}) \otimes I_T + A^{*(s)} \qquad (50)$$

Apply Theorem 1 and lemma A.6 to (50) and the result follows. #

Definition:

A zero iteration matrix is an iteration matrix which is nilpotent.

Theorem 16:

RFIIV is always locally convergent. In fact it has zero iteration matrix.

Proof:

We rewrite (11.33) as

$$y_i^{(s+1)} = (I-M_i)y_i + Y_i^{(s+1)} \hat{\beta}_i^{(s+1)} - (I-M_i)Y_i \hat{\beta}_i^{(s+1)} \qquad (51)$$

and thus

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} - \frac{\partial Y_i^{(s+1)}}{\partial y_{tj}^{(s)}} \hat{\beta}_i^{(s+1)} = [Y_i^{(s+1)} - (I-M_i)Y_i] \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}} \qquad (52)$$

Combining with (48) which holds unchaged, the right hand side of (52) becomes

$$[Y_i^{(s+1)} - (I-M_i)Y_i](Y_i^{(s)'} M_i Y_i)^{-1} \frac{\partial Y_i^{(s)'}}{\partial y_{tj}^{(s)}} \hat{u}_i^{(s+1)} \qquad (53)$$

Define $T_{(i)}^{(s)} = [Y_i^{(s)} - (I-M_i)Y_i](Y_i^{(s)'} M_i Y_i)^{-1}$ so that by (48a) $T_{(i)}^{(s)'} \hat{u}_i^{(s+1)} = 0$ $i = 1 \ldots M$. Specify $A^{*(s)}$ of order MT having j,ith submatrix $A_{ji}^{(s)'} = \begin{cases} T_{(i)j}^{(s)} \hat{u}_i^{(s+1)'} & \text{for } \beta_{ji} \neq 0 \\ 0 & \text{otherwise} \end{cases}$. $A^{*(s)}$ is nilpotent by lemmas 1 and A.5. (53) contains $T_{(i)j}^{(s)} \hat{u}_i^{(s+1)}$ and $[V_i^{(s)} - V_i^{(s+1)}] (Y_i^{(s)'} M_i Y_i)^{-1} \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}}$ which vanishes within a neighbourhood defined by the approximation $V_i^{(s)} \simeq V_i^{(s+1)}$ also used in theorem 8. Thus

$$J^{(s)'} = \frac{\partial Y^{(s+1)}}{\partial Y^{(s)}} = (\hat{\beta}^{(s+1)^{-1}} \otimes I) A^{*(s)} \qquad (54)$$

Application of lemma A.6 implies $\rho(J^{(s)}) = 0$ all $S \geq 0$ including at convergence. #

Theorem 17:

(i) PIIV has local convergence criterion $\rho(I - \beta) < 1$.

(ii) RFPIIV is always locally convergent, with zero iteration matrix for Theorem 1.

Proof:

(i) From (49)

$$\frac{\partial(\hat{\Pi}^{(S)})_i'}{\partial \hat{\Pi}_{jk}^{(S)}} Z'\hat{u}_i^{(S+1)} = \begin{cases} z_k' \hat{u}_i^{(S+1)} = 0 \text{ if } z_k \in Z_i \text{ since } Z_i' \hat{u}_i^{(S+1)} = 0 \\ 0 \quad \text{if } z_k \notin Z_i \end{cases} \tag{55}$$

Hence $\partial\hat{\beta}_i^{(S+1)}/\partial\hat{\Pi}_{jk}^{(S)} = 0$ and from (39) $\partial\hat{\Pi}_i^{(S+1)}/\partial\hat{\Pi}_{jk}^{(S)} = I^*(\hat{\beta}_{ik}^{(S+1)} \otimes I_K)$.

I^* is idempotent. Evaluating $J^{(S)}$ at Y^* we have $J = (I^*(I-\beta) \otimes I_{KM})$ and the theorem follows by application of Theorem 1 and corollary A.3.

(ii) The basic iteration equivalent to (11.19c) is

$$\hat{\Pi}_i^{(S+1)} = (\hat{\Pi}^{(S+1)})_i \hat{\beta}_i^{(S+1)} + C^*_i y_i - C^*_i Y_i \hat{\beta}_i^{(S+1)} \tag{56}$$

and thus instead of (44)

$$\frac{\partial\hat{\Pi}_i^{(S+1)}}{\partial\hat{\Pi}_{jk}^{(S)}} - \frac{\partial(\hat{\Pi}^{(S+1)})_i}{\partial\hat{\Pi}_{jk}^{(S)}} \hat{\beta}_i^{(S+1)} = - C^*_i Y_i \frac{\partial\hat{\beta}_i^{(S+1)}}{\partial\hat{\Pi}_{jk}^{(S)}} = 0 \tag{57}$$

as in part (i). #

Comment:

Comment (b) following theorem 10 applies to IIV estimators using direct instruments.

Theorem 18:

Theorems 15 - 17 hold unaltered for endogenous iteration.

Proof:

As Theorem 11 for (15), (17) or reformed (15) and (19c). #

Theorem 19:

(i) Fractional versions of IIV, PIIV, EIIV, EPIIV have the local convergence criterion that the real parts of the roots of $I - \beta$ be less than unity.

(ii) Fractional versions of RFIIV, RFPIIV, RFEIIV, RFEIIV, RFEPIIV are always locally convergent if $0 < \alpha < 2$, where $\alpha = 1$ is optimal.

(iii) Comparison and optimisation theorems of Chapters 8, 9 apply locally to (i) with $I - \beta$ for B.

Proof:

Identical to Theorem 12 using Theorems 15 - 18; where in (ii) the values of α are immediate since the iteration matrix is $\alpha J + (1-\alpha)I$ where J is the iteration matrix of the underlying estimator - $\rho(J) = 0$ in each case here. #

Theorem 20:

(i) Recursive verions of IIV, PIIV, EIIV, EPIIV have the local convergence criterion that the real parts of the roots of $I - \beta$ be less than unity.

(ii) Recursive versions of RFIIV, RFPIIV, RFEIIV, RFEPIIV are always locally convergent for $0 < \alpha < 2$; $\alpha = 1$ is optimal for asymptotic convergence rates.

(iii) Any recursive IIV estimator is locally convergent only if $0 < \alpha < 2$ - a global property.

(iv) Comparison and optimisation theorems of Chapters 8,9 apply to (i) locally with $I - \beta$ for B.

Proof:

Identical to Theorem 13 using Theorems 15 - 18. The α values in (ii) are derived as in Theorem 19 since the iteration matrix is then for the J of theorem 19, $(I-\alpha L)^{-1}(\alpha U + (1-\alpha)I)$ with L,U strictly lower and upper triangular submatrices of J. For RFIIV or RFEIIV $J^S = 0 \quad S \geq MT - 1$

and for RFPIIV or RFPEIIV $J = 0$ where the asymptotic convergence rate is $\lim_{S\to\infty} \frac{1}{S} \log ||J^S||$. #

Corollary 8:

Recursively accelerated IIV estimators with RF instruments have convergence properties locally invariant to the ordering of the equations.

Proof:

All have zero iteration matrices for any ordering. #

Theorem 21:

Under a GEID specification Theorems 15 - 20 hold unaltered.

Proof:

Identical to Theorem 14. #

The comments following Theorem 31 apply equally to IIV estimators.

Remark:

In each of theorems 7,8,11-16, 18-21 a matrix $A^{*(s)}$ was defined with typical element $T^{(s)}_{(i)kj}\ \hat{u}^{(s+1)}_{ti}$ for $t,k = 1 \ldots T$; $i = 1 \ldots M$; $j = 1 \ldots r_i$. But in each case $\lim_{T\to\infty} E(T^{(s)}_{(i)}\ \hat{u}^{(s+1)}_{ti}) = 0$ so as T becomes large we expect $A^{*(s)} \to 0$. Because of the difficulties in assessing the number of iterations to convergence, mentioned in the Appendix, when $J^{(s)}$ is the sum of two matrices, one nilpotent, it is assumed from now on that T is sufficiently large for $A^{*(s)}$ to be neglected. Further justification for this is provided by

Lemma 2:

If $J^{(s)} = B^{(s)} + A^{*(s)}$ then $\rho(J^{(s)}) = \rho(B^{(s)})$ all $T \geq M$.

Proof:

Bergstrom (1974) for each $B^{(s)}$ in these theorems. #

This convention will be used throughout Chapter 13 in assessing and comparing computational burdens of various estimators. It must be remembered that this convention does not represent the best view of the

computation properties of these estimators - $B^{(s)}$ and $A^{*(s)}$ could be offsetting to $\rho(J^{(s)})$ if $T < M$. For all the models considered here $T \geq M$, but the point must be made to preserve the availability criterion. Finally we note that PIIV estimators with RF instruments are at least as good as IIV ones with RF instruments and generally better, having a null iteration matrix rather than having expected value zero. Lemma A.4 implies the zero iteration matrix vanishes for $S \geq MT - 1$. In any case on convergence and speed of convergence criteria IIV, PIIV estimators with RF instruments will be preferred to other methods.

12.5 The QRF instrument case:

To the results of Sections 4.1 and 4.2 we add convergence conditions for FP, or IIV estimators in endogenous or parametric iteration when QRF instruments are used. The conditions obtained are combinations of the corresponding conditions for direct and RF instruments - combinations which may be very complicated and in any even are not interesting in themselves. However it is necessary to derive the iteration matrices so that we may investigate if there is an optimal sequence of values k_s to maximise the speed of convergence and/or the total work done to achieve convergence. Only two examples of convergence conditions for QRF instrument estimators are sketched in here.

Convergence conditions for IIV and EIIV are obtained from the iteration

$$
\begin{aligned}
y_i^{(s+1,k_{s+1})} &= [Y_i^{(s+1,k_{s+1}-1)} - (I-M_i)Y_i]\hat{\beta}_i^{(s+1)} \;)\; (I-M_i)y_i \\
&= \{Y^{(s,k_s)}[I-\hat{\beta}^{(s+1)}]^{k_{s+1}-1} + \sum_{\ell=0}^{k_{s+1}-2} Z\hat{\Gamma}^{(s+1)}(I-\hat{\beta}^{(s+1)})^{\ell}\}_i \hat{\beta}_i^{(s+1)} \\
&\quad + (I-M_i)(y_i - Y_i\hat{\beta}_i^{(s+1)})
\end{aligned}
\tag{58}
$$

Writing $(I-M_i)(y_i - Y_i\hat{\beta}^{(s+1)}) = Z_i\hat{\gamma}^{(s+1)}$ (59)

by (48a);

$$\frac{\partial y_i^{(s+1,k_{s+1})}}{\partial y_{tj}^{(s,k_s)}} = \left\{ \frac{\partial Y^{(s,k_s)}}{\partial y_{tj}^{(s,k_s)}} [I-\hat{\beta}^{(s+1)}]^{k_{s+1}-1} - (k_{s+1}-1) Y^{(s,k_s)} [I-\hat{\beta}^{(s+1)}]^{k_{s+1}-2} \cdot \right.$$

$$\left. \frac{\partial \hat{\beta}^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} + \sum_{\ell} \left[Z \frac{\partial \Gamma^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} (I-\beta^{(s+1)})^{\ell} - \ell \hat{\Gamma}^{(s+1)} (I-\hat{\beta}^{(s+1)})^{\ell-1} \frac{\partial \beta^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} \right] \right\}_i \hat{\beta}_i^{(s+1)}$$

$$+ \left\{ Y^{(s,k_s)} (I-\hat{\beta}^{(s+1)})^{k_{s+1}-1} + Z\hat{\Gamma}^{(s+1)} \sum_{\ell} (I-\hat{\beta}^{(s+1)})^{\ell} \right\}_i \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}}$$

$$+ \frac{Z_i \partial \gamma_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} \qquad (60)$$

Now $\dfrac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} = 0$ if $y_j \notin Y_i$

$$= (Y_i^{(s,k_s)'} M_i Y_i)^{-1} \frac{\partial Y_i^{(s,k_s)'}}{\partial y_{tj}^{(s,k_s)}} \hat{u}_i^{(s+1)} \quad \text{otherwise} \qquad (61)$$

and from (11.13)

$$\frac{\partial \hat{\gamma}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} = -(Z_i'Z_i)^{-1} Z_i'Y_i \frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} \qquad (62)$$

$\dfrac{\partial \hat{\beta}^{(s+1)}}{\partial y_{tj}^{(s,k_s)}}$ and hence $\dfrac{\partial \hat{\Gamma}^{(s+1)}}{\partial y_{tj}^{(s,k_s)}}$ have nonzero elements given by (61) and

$-(Z_i'Z_i)^{-1} Z_i'Y_i (Y_i^{(s,k_s)'} M_i Y_i)^{-1} \dfrac{\partial Y_i^{(s,k_s)'}}{\partial y_{tj}^{(s,k_s)}} \hat{u}_i^{(s+1)}$ respectively $i = 1 \ldots M$

Define $A_{ji}^{(s)} = \begin{cases} T_{(i)j}^{(s)} \hat{u}_i^{(s+1)'} & \text{if } \beta_{ji} \neq 0 \\ 0 & \text{otherwise} \end{cases}$ and $A^{*(s)}$ of order MT having j,i^{th} submatrix $A_{ji}^{(s)}$; where $T_{(i)}^{(s)} = (I-M_i)Y_i(Y_i^{(s,k_s)'}M_iY_i)^{-1}$ so $T_{(i)}^{(s)'} \hat{u}_i^{(s+1)} = 0$. By lemma 1 $A_{ji}^{(s)}$ is nilpotent; by lemma A.5 $A^{*(s)}$ also is.

We can now write

$$\frac{\partial y_i^{(s+1,k_{s+1})}}{\partial y_j^{(s,k_s)}} = \left\{ \frac{\partial Y^{(s,k_s)}}{\partial y_j^{(s,k_s)}} [I-\hat{\beta}^{(s+1)}]^{k_{s+1}-1} \right\}_i \hat{\beta}_i^{(s+1)} + E_{ji}^{(s)}$$

$$= [(I-\hat{\beta}^{(s+1)})_{ji}^{k_{s+1}-1} \hat{\beta}_{ji}^{(s+1)} \otimes I_T] + E_{ji}^{(s)} \qquad (63)$$

where $E_{ji}^{(s)}$ is j,i^{th} submatrix from $B^{(s)} A^{*(s)} C^{(s)}$ with $B^{(s)}$ and $C^{(s)}$ defined in the obvious manner to collect all terms respectively pre and post multiplying $T_{(i)}^{(s)} \hat{u}_i^{(s+1)'}$ $i = 1 \ldots M$ in (60).

In the FP family (25) gives

$$\frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} = \begin{cases} 0 & \text{if } \beta_{ji} = 0 \\ [Y_i^{(s,k_s)'}M_iY_i^{(s,k_s)}]^{-1} \left[\frac{\partial Y_i^{(s,k_s)'}}{\partial y_{tj}^{(s,k_s)}} \hat{v}_i^{(s+1)} - Y_i^{(s,k_s)'} M_i \frac{\partial Y_i^{(s,k_s)}}{\partial y_{tj}^{(s,k_s)}} \hat{\beta}_i^{(s+1)} \right] \end{cases} \qquad (64)$$

The iteration is (58) with M_i premultiplying the first term on the right and $(I - M_i)y_i$ as the last. Thus (60) applies with all terms premultiplied by M_i and the last omitted.

Also

$$\frac{\partial \hat{\gamma}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}} = -(Z_i'Z_i)^{-1}Z_i'\left[\frac{\partial Y_i^{(s,k_s)}}{\partial y_{tj}^{(s,k_s)}}\hat{\beta}_i^{(s+1)} + Y_i^{(s,k_s)}\frac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s,k_s)}}\right]$$

$$= -(Z_i'Z_i)^{-1}Z_i'\left[\{I - Y_i^{(s,k_s)}(Y_i^{(s,k_s)'}M_iY_i^{(s,k_s)})^{-1}Y_i^{(s,k_s)}M_i\}\hat{\beta}_{ij}^{(s+1)}\right.$$

$$\left. + (Y_i^{(s,k_s)'}M_iY_i^{(s,k_s)})^{-1}\frac{\partial Y_i^{(s,k_s)'}}{\partial y_{tj}^{(s,k_s)}}\hat{v}_i^{(s+1)}\right] \qquad (65)$$

We substitute (64) and (65) into the new version of (60). By similar arguments to the IIV case all terms involving $v_i^{(s+1)}$ vanish into a matrix defined analogously to $E_{ji}^{(s)}$ where $A^{*(s)}$ is again nilpotent. The remaining terms indicate an extremely complicated iteration matrix. It is clear the iteration matrix involves some kind of rational function of elements of $\beta_{ij}^{(s+1)}$ (which probably yields overall a similar function of $(I - \beta)$) of which the coefficients themselves are nonidempotent matrices; so that with both additive and multiplicative terms in the iteration matrix there is no possibility of obtaining any analytic expressions for its spectral radius in FP estimators using QRF instruments.

A similar analysis may be undertaken with parametric iteration. We can conclude after such analysis

Theorem 22:

IIV estimators with QRF instruments have local convergence criterion $\rho((I - \hat{\beta}^{(s+1)})^{k_s}) < 1$, i.e. $\rho(I - \beta) < 1$.

Proof:

(63) implies $J^{(s)} = [I - \hat{\beta}^{(s+1)}]^{k_s} \otimes I + B^{(s)}A^{*(s)}C^{(s)}$. By lemma A.6 (ii) $B^{(s)}A^{*(s)}C$ is nilpotent hence applying lemma A.6 (i) and theorem 1 we have the result. The parametric version also results in (63), without $E_{ji}^{(s)}$. #

Substituting $k_s = 1$ into theorem 22 gives theorem 15. Note if $\rho(I - \beta) > 1$ then for any $k_s \gtrsim 1$ we have nonconvergence. However,

Corollary 9:[1]

RFIIV or RFEIIV as well as RFPIIV and RFEPIIV are always globally convergent.

---

1. The alterntive proof follows from theorem 16. Define $T_{(i)}^{(s)} = [Y_i^{(s+1)} - (I-M_i)Y_i](Y_i^{(s)'}M_iY_i)^{-1}$ then $T_{(i)}^{(s)'}\hat{u}_i^{(s+1)} = 0$ and $A^{*(s)}$ is defined with this $T_{(i)}^{(s)}$ in theorem 17. This follows because $X^{(s+1)'}X_i(X_i^{(s)'}X_i)^{-1}X_i^{(s)'} = \begin{bmatrix} Y^{(s+1)'} \\ Z' \end{bmatrix}[I - M_i + M_iB_iM_i]$

after much manipulation where $B_i = Y_i(Y_i^{(s)'}M_iY_i)^{-1}Y_i^{(s)'}$. But $Y_i^{(s+1)} = (Z \vdots Z)\begin{bmatrix} \Pi^{(s+1)}R_i & 0 \\ 0 & 0 \end{bmatrix}$ where $Y_i = YR_i$ defines R_i. Thus substituting $Z\Pi^{(s+1)}$ for $Y^{(s+1)}$ in the RF instrument estimator we have

$$Y_i^{(s+1)'}X_i(X_i^{(s)'}X_i)^{-1}X_i^{(s)'} = \begin{bmatrix} R_i'\Pi^{(s+1)'} & 0 \\ 0 & 0 \end{bmatrix} P_{1i}P_{1i}'\begin{bmatrix} Z' \\ Z' \end{bmatrix}X_i(X_i^{(s)'}X_i)^{-1}X_i^{(s)'}$$

$$= R_i'\begin{bmatrix} (\Pi)_i^{(s+1)} \\ \cdots\cdots \\ (\Pi)_i^{(s+1)} \end{bmatrix}'\begin{bmatrix} Z_i' \\ \hline Z_i'(I-M_iB_iM_i) \end{bmatrix}$$

where for notational convenience we define for each $i = 1 \ldots M$ the permutation matrix P_{1i} such that $ZP_{1i} = (Z_i \vdots \bar{Z}_i)$ where $Z_i \cup \bar{Z}_i = Z$ and $Z_i \cap \bar{Z}_i$ is the null set. $\Pi^{(s+1)}$ is partitioned conformably. Note $P_{1i}P_{1i}' = I$. Define also the permutation matrix P_{2i} such that $YP_{2i} = (Y_i \vdots \bar{Y}_i)$ where $Y_i \cup \bar{Y}_i = Y$ and $Y_i \cap \bar{Y}_i$ is null and note $P_{2i}'P_{2i} = I$. Since then $R_iP_{2i}'P_{2i}\begin{bmatrix} (\Pi)_i^{(s+1)'} \\ \cdots i \cdots \\ (\bar{\Pi})_i^{(s+1)'} \end{bmatrix}$ sends at least the lower $K - s_i$ rows to zero.

($M - r_i \gtrsim K - s_i$ by identification conditions) we have $Y_i^{(s+1)'}X_i(X_i^{(s)'}X_i)^{-1}X_i^{(s)'} = (\Pi)_i^{(s+1)'}Z_i'$ at most and hence $Y_i^{(s+1)'}\hat{u}_i^{(s+1)} = 0$. Corollary 9 now follows from the rest of theorem 16. #

Proof:

Substitute $k_s = \infty$ in theorem 22; then $J^{(s)} = B^{(s)}A^{*(s)}C^{(s)}$ which is a zero iteration matrix for each $S \geq 0$. #

Corollary 10:

(i) The approximate optimal sequence k_s for IIV estimators is $k_s = \infty$ on an asymptotic convergence rate criterion.

(ii) It is impractical to calculate the optimal sequence k_s for FP estimators.

Proof:

(i) This optimal sequence k_s minimises the number of steps to achieve convergence. Theoretically in PIIV $k_s = \infty$ implies a null iteration matrix and hence instantaneous convergence. Any other k_s sequence implies a nonzero spectral radius for the iteration matrix in IIV itself.

(ii) Follows immediately from the remarks preceding Theorem 22. #

Part (i) of corollary 10 represents a purely theoretical point. Before convergence we know that full asymptotic efficiency does not follow unless RF instruments are used and anyway the relation between spectral radius and speed of convergence is an approximation at best. For an average convergence rate since RF instruments require considerable computation we should calculate the minimum point in any trade off in work done between decreasing the estimation iterations and increasing those of instrument formation. There may therefore be some other optimal sequence k_s but it would have to be determined empirically in particular cases. Some attempt is made to assess the optimal k_s for average convergence rates in Chapter 13.

12.6 Global convergence conditions and nonlinearity

In view of the nonlinearity of the iterative process necessary to achieve a fixed point in the parameter space, any convergence analysis

was based on Theorem 1. However it is of interest to see if any further generalisations to global convergence results are possible.

12.6.1 Fixed point estimators:

No generalisations or extensions are possible to the convergence criteria for any of the fixed point estimators. Using the infinite sum in (2), the terms $J_m|_{Y^*} = F^{(m)}(Y^*)$ $m = 1 \ldots .\infty$ are available by repeated differentiation and evaluation from $J^{(s)}$ in Theorems 7, 8, 9 and 10 respectively. However it would not be possible to extract a convergence criterion from this extended expression. Similarly no criterion would be available if each term were not evaluated at Y^* - even if just the linear term were taken. Neither exact nor global criteria may be obtained.

12.6.2 IIV estimators:

The same remarks apply to IIV estimators with direct or QRF instruments, as applied to all FP estimators in 12.6.1. Neither exact nor global criteria of any kind are available. But

Theorem 23:

In theorems 16 - 21, for RFIIV, RFEIIV we have global, but for RFPIIV and RFEPIIV we have global and exact convergence criteria.

Proof:

Global conditions because $\rho(J^{(s)}) = 0$ all $S \geq 0$ and independent of S. RFPIIV and RFEPIIV have $J^{(s)} = 0$ all $S \geq 0$ so $J_m^{(s)} = 0$ all $m = 1 \ldots .$ and substitution in (2) completes the proof by (21). #

In the context of Theorem 23 we have assumed $F(\cdot)$ in (1) satisfies Lipschitz conditions over the whole real domain. If we replace that assumption with the existence of at least one Y^* they are still global conditions in the sense that convergence for arbitrary start $Y^{(0)}$ always occurs but not necessarily to the same fixed point each time. It is interesting that this theorem supports the conjecture of Maddala (1971) that the fixed point

theorem applies to these estimators (method 1 RRFIIV in his notation) rather than FP itself. Theorem 23 is extended in Chapter 14.

12.6.3 Nonlinearities

Lemma 3:

(a) Endogenous iteration is available iff the model specification is, or is transformed to be, linear in the parameters.

(b) Then the Theorems 7 - 23 hold unaltered.

Proof:

(a) Let the model be $\sum_{j=0}^{K+M} f_{ij}(X_{ij})\delta_{ij} = u_i \quad i = 1 \ldots M$ (68)

where X_{ij} is the subset of $X_i = [Y_i : Z_i]$ which occurs in the j^{th} nonlinear function $f_{ij}(\cdot)$. Then, with some normalisation $\delta_{ii} = 1$, endogenous iteration may be undertaken as in Chapter 11, such that for each $\delta_{ij} \neq 0$ where $Y_i^{(s)}$ is replaced by $f_{ij}^{(s)}(X_{ij})$ if $y_j \in X_{ij}$ all $j = 1 \ldots M$ and otherwise Z_i is replaced by $f_{ij}(X_{ij})$ all $j = 1 \ldots M$,
where $f_{ij}^{(s)}(X_{ij}) = -\sum_{j\neq i} f_{ij}^{(s-1)}(X_{ij})\hat{\delta}_{ij}^{(s)}$ or $-\sum_{j\neq i} f_{ij}^{(s)}(X_{ij})\hat{\delta}_{ij}^{(s)}$ as required, i.e. if $f_{ij}(x_{ij})$ contains at least one endogenous argument it is treated as an endogenous variable. Conversely if the model specification cannot be transformed to (68) neither IV or least squares estimation is available at the estimation phase so endogenous iterative estimators could not be used without alteration or approximation.

(b) If the model is as (68) and we use the rule of part (a) then plim $T^{-1} Z_i'u_i = 0$ implies plim $T^{-1}[f_{ij}^{(s)}(X_{ij})]'u_i = 0 \quad j = 1 \ldots K+M, i=1 \ldots M$.
In any of the Theorems 7 - 23 the only change is we write

$f_{ij}'(X_{ij}^{(s)}) \dfrac{\partial X_{ij}^{(s)}}{\partial y_{tk}^{(s)}}$, in place of $\dfrac{\partial Y_i^{(s)}}{\partial y_{tk}^{(s)}}$. The theorems therefore remain valid. #

So far the theorems are only valid if a model can be made linear in the parameters. We treat more general nonlinear models in Chapter 15.

12.7 Sphericality transforms

We now extend the convergence analysis to full information and autoregressive versions for each of the estimators considered. It is a simple matter to convert the previous convergence conditions to full information and/or autoregressive estimators by a sphericality transform in its disturbance terms, and reworking the foregoing theorems. Iterative estimators with sphericality transforms will be of the style (5.1) when ϕ is the appropriate variance - covariance matrix. For full information version $\hat{\phi}^{(s)} = \hat{\Sigma}^{(s)} \otimes I$ will be used and $\hat{\Sigma}^{(s)}$ is re-estimated at each iteration as in (5.17). In autoregressive verions we use $\hat{\phi}^{(s)} = \text{diag}\,(\hat{\Sigma}_i^{(s)})$ where $\Sigma_i^{(s)} = T^{-1}(y_i - Y_i\hat{\beta}^{(s)} - Z_i\hat{\gamma}_i^{(s)})(y_i - Y_i\hat{\beta}_i^{(s)} - Z_i\hat{\gamma}_i^{(s)})'$ is re-estimated at each iteration. A full information atuoregressive version would use (7.22), re-estimated at each iteration. Such estimators are those referred to in the discussion of information classes to achieve full asymptotic efficiency. The various intermediate states in the information class hierarchy can be reached by iterating the limited information version of an estimator inserting the implied $\hat{\phi}$. However these possibilities are not considered explicitly here.

We use the general iteration from Chapter 11.

$$\left.\begin{aligned} y^{(s+1)} &= (I-Q^{*(s)})y + Q^{*(s)}Y^{(s)}\hat{\beta}_o^{(s+1)} \qquad \text{for FP} \\ \text{or} \\ y^{(s+1)} &= (I-Q^{*(s)})y + [\hat{Y}^{(s)} - (I-Q^{*(s)})\hat{Y}]\hat{\beta}_o^{(s+1)} \qquad \text{for IIV} \end{aligned}\right\} \tag{69}$$

where $M^{*(s)} = \phi^{(s)^{-1}} Q^{*(s)}$ defines $Q^{*(s)}$, and $\hat{\beta}_o^{(s+1)} = (\hat{Y}^{(s)'} M^{*(s)} \hat{Y}^{(s)})^{-1} \hat{Y}^{(s)'} M^{*(s)} y$ for the IIV family, and $\hat{\beta}_o^{(s+1)} = (\hat{Y}^{(s)'} M^{*(s)} \hat{Y}^{(s)})^{-1} \hat{Y}^{(s)'} M^{*(s)} y$ for the FP family. $\phi^{(s)}$ is symmetric being the transform appropriate to the information class. Actually in the autoregressive case we could treat equations individually, but it is convenient to treat the general case. For FP

$$\frac{\partial y^{(s+1)}}{\partial y_{tj}^{(s)}} = \frac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}} [\hat{Y}^{(s)} \hat{\beta}_o^{(s+1)} - y] + Q^{*(s)} \frac{\partial Y^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)}$$

$$+ Q^{*(s)} \hat{Y}^{(s)} \frac{\partial \hat{\beta}_o^{(s+1)}}{\partial y_{tj}^{(s)}} \tag{70}$$

and for IIV

$$\frac{\partial y^{(s+1)}}{\partial y_{tj}^{(s)}} = \frac{\partial \hat{Y}^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} + [\hat{Y}^{(s)} - (I - Q^{*(s)}) \hat{Y}] \frac{\partial \hat{\beta}_o^{(s+1)}}{\partial y_{tj}^{(s)}} \tag{71}$$

$$+ \frac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}} [\hat{Y} \hat{\beta}_o^{(s+1)} - y]$$

Lemma 4:

For IIV family estimators

$$\frac{\partial y^{(s+1}}{\partial y_{tj}^{(s)}} = \frac{\partial \hat{Y}^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} + \hat{T}^{(s)} \psi^{*}{}_{tj}^{(s+1)} \quad t = 1 \ldots T \tag{72}$$

$$\text{and } j = 1 \ldots M$$

$$i = 1 \ldots r_j$$

where $\hat{T}^{(s)} = (\hat{Y}^{(s)} - (I-Q^{*(s)})\hat{Y})(\hat{Y}^{(s)\prime} M^{*(s)} \hat{Y})^{-1}$ and

$\hat{u}^{(s+1)} = \text{vec}\ \hat{u}_i^{(s+1)} = [I - \hat{X}(\hat{X}^{(s)\prime} \phi^{(s)-1} \hat{X})^{-1} \hat{X}^{(s)\prime} \phi^{(s)-1}]y$

with $\psi^{(s+1)} = \phi^{(s)-1} u^{(s+1)}$ and $(\psi^{*(s+1)}_{tj})_i = \psi^{(s+1)}_{tj}$ if $\beta_{ij} \neq 0$; 0 otherwise.

Proof:

$$\frac{\partial \hat{\beta}_o^{(s+1)}}{\partial y_{tj}^{(s)}} = (\hat{Y}^{(s)\prime} M^{*(s)} \hat{Y})^{-1} \left[\frac{\partial \hat{Y}^{(s)\prime}}{\partial y_{tj}^{(s)}} M^{*(s)} + \hat{Y}^{(s)\prime} \frac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}}\right] (y - \hat{Y}\hat{\beta}_o^{(s+1)})$$

so in (71) $\dfrac{\partial y^{(s+1)}}{\partial y_{tj}^{(s)}} = \dfrac{\partial \hat{Y}^{(s)}}{\partial Y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} + \left\{(\hat{Y}^{(s)} - (I-Q^{*(s)})\hat{Y})(\hat{Y}^{(s)\prime} M^{*(s)} \hat{Y})^{-1}\right.$

$$\left.\left[\frac{\partial \hat{Y}^{(s)\prime}}{\partial y_{tj}^{(s)}} M^{*(s)} + \hat{Y}^{(s)\prime} \frac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}}\right] - \frac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}}\right\}$$

$$(\hat{Z}\gamma_o^{(s+1)} + \hat{u}^{(s+1)}) \qquad (73)$$

It may be verified $M^{*(s)}\hat{Z} = \phi^{(s)-1} Q^{*(s)}\hat{Z} = 0$; also $Q^{*(s)}\ \hat{u}^{(s+1)} = \hat{u}^{(s+1)}$ using the notation $\hat{Z}' = (0 \vdots I)\hat{X}^{(s)\prime}$ so then $M^{*(s)}\ \hat{u}^{(s+1)} = \phi^{(s)-1}\ \hat{u}^{(s+1)}$. Using the notation $\hat{Y}^{(s)\prime} = (I \vdots 0)\hat{X}^{(s)\prime}$ we also have $\hat{Y}^{(s)\prime} \phi^{-1} \hat{u}^{(s+1)} = 0$ and $(I - Q^{*(s)})'\ \phi^{(s)-1}\ \hat{u}^{(s+1)} = 0$. By the identities above $\hat{T}^{(s)\prime}\ \hat{\psi}^{(s+1)} = 0$.

The term in $\dfrac{\partial \hat{Y}^{(s)\prime}}{\partial y_{tj}^{(s)}}\ M^{*(s)}$ gives $\hat{T}^{(s)}\ \hat{\psi}_{tj}^{(s+1)}$ in (73).

Next note $\hat{Y}^{(s)\prime} \dfrac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}} = \hat{Y}^{(s)\prime}\left[\dfrac{\partial \phi^{(s)-1}}{\partial y_{tj}^{(s)}} Q^{*(s)} + \phi^{(s)-1} \dfrac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}}\right]$

so $\hat{T}^{(s)}\hat{Y}^{(s)\prime} \dfrac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}} - \dfrac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}} = \hat{T}^{(s)}\hat{Y}^{(s)\prime} \dfrac{\partial \phi^{(s)-1}}{\partial y_{tj}^{(s)}} Q^{*(s)} + (\hat{T}^{(s)}\hat{Y}^{(s)\prime} \phi^{(s)-1} - I)\dfrac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}}$

(74)

But $\frac{\partial \phi^{(s)-1}}{\partial y_{tj}^{(s)}} = - \phi^{(s)-1} \frac{\partial \phi^{(s)}}{\partial y_{tj}^{(s)}} \phi^{(s)-1}$ so $\frac{\partial Q^{*(s)}}{\partial y_{tj}^{(s)}} = (I-Q^{*(s)}) \frac{\partial \phi^{(s)}}{\partial y_{tj}^{(s)}} M^{*(s)}$

follows directly. Substituting for both differentials on the right in (74) we get $- [\hat{T}^{(s)}\hat{Y}^{(s)'}M^{*(s)} + (I-Q^{*(s)})] \frac{\partial \phi^{(s)}}{\partial y_{tj}^{(s)}} M^{*(s)}$. The remainder of the second term on the right of (73) therefore vanishes because for any of the definitions of $\phi^{(s)}$ given, $\hat{u}^{(s)}$ depends only on $\hat{\delta}^{(s)}$, i.e. $y^{(s-1)}$ so

$$\frac{\partial \phi^{(s)}}{\partial y_{tj}^{(s)}} = 0 \quad t = 1 \ldots T, \; j = 1 \ldots M. \quad \#$$

Lemma 5:

For FP family estimators

$$\frac{\partial y^{(s+1)}}{\partial y_{tj}^{(s)}} = P^{(s)} \frac{\partial \hat{Y}^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} + \hat{R}^{(s)} \psi^{*}_{tj}{}^{(s+1)} \qquad \begin{matrix} t = 1 \ldots T, \\ j = 1 \ldots M \end{matrix} \tag{75}$$

where $P^{(s)} = Q^{*(s)}[I-\hat{Y}^{(s)}(\hat{Y}^{(s)'}M^{*(s)}\hat{Y}^{(s)})^{-1}\hat{Y}^{(s)'}M^{*(s)}]$ is idempotent and $\hat{R}^{(s)} = Q^{*(s)}\hat{Y}^{(s)}(\hat{Y}^{(s)'}M\hat{Y}^{(s)})^{-1}$ and $\hat{v}^{(s+1)} = \text{vec } \hat{v}_i^{(s+1)} = [I-\hat{X}^{(s)}(\hat{X}^{(s)'}\phi^{(s)-1}\hat{X}^{(s)})^{-1}\hat{X}^{(s)'}\phi^{(s)-1}]y$ and here $\psi^{(s+1)} = \phi^{(s)-1}\hat{v}^{(s+1)}$ and $\psi^{*}_{tj}{}^{(s+1)}$ as before.

Proof:

$$\frac{\partial \hat{\beta}_o^{(s+1)}}{\partial y_{tj}^{(s)}} = (\hat{Y}^{(s)'}M^{*(s)}\hat{Y}^{(s)})^{-1} \left\{ \left[\frac{\partial \hat{Y}^{(s)'}}{\partial y_{tj}^{(s)}} M^{*(s)} + \hat{Y}^{(s)'} \frac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}} \right] \right.$$

$$\left. (y-\hat{Y}^{(s)}\hat{\beta}_o^{(s+1)}) - \hat{Y}^{(s)'}M^{*(s)} \frac{\partial \hat{Y}^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} \right\}$$

and inserting this into (70) we get

$$\frac{\partial y^{(s+1)}}{\partial y_{tj}^{(s)}} = \left\{ Q^{*(s)}\hat{Y}^{(s)}(\hat{Y}^{(s)'}M^{*(s)}\hat{Y}^{(s)})^{-1} \left[\frac{\partial \hat{Y}^{(s)'}}{\partial y_{tj}^{(s)}} M^{*(s)} + \hat{Y}^{(s)'} \frac{\partial M^{*(s)}}{\partial y_{tj}^{(s)}} \right] - \frac{\partial Q^{*(s)}}{\partial y_{ti}^{(s)}} \right\}$$

$$(y-\hat{Y}^{(s)}\hat{\beta}_o^{(s+1)}) + Q^{*(s)}[I-\hat{Y}^{(s)}(\hat{Y}^{(s)'}M^{*(s)}\hat{Y}^{(s)})^{-1}\hat{Y}^{(s)'}M^{*(s)}] \cdot$$

$$\frac{\partial \hat{Y}^{(s)}}{\partial y_{tj}^{(s)}} \hat{\beta}_o^{(s+1)} \tag{76}$$

We repeat the steps of Lemma 2 identically with $\hat{v}^{(s+1)}$ and $\hat{R}^{(s)}$ replacing $\hat{u}^{(s+1)}$ and $\hat{T}^{(s)}$ respectively, because $y - \hat{Y}^{(s)} \hat{\beta}_o^{(s+1)} = \hat{Z} \hat{\gamma}_o^{(s+1)} + \hat{v}^{(s+1)}$ but $\frac{\partial \phi}{\partial y_{tj}^{(s)}} = 0$ as before. The result now obviously follows. #

Theorem 24:

Any iterative estimator employing any sphericality transform appropriate to some information class has an identical convergence criterion to its limited information version, i.e. theorems 7 - 21 above are valid exactly in each informtion class.

Proof:

(a) The IIV family: From (72) for each $t = 1 \ldots T$, $j = 1 \ldots M$ we fill in the t^{th} column of each submatrix of the j^{th} column block,

so $$\frac{\partial y^{(s+1)}}{\partial y^{(s)}} = (I - \beta) \otimes I_T + A^{*(s)} \qquad (77)$$

where $A^{*(s)} = \hat{T}^{(s)} \psi^{*(s+1)'}$ from (72) which by Lemmas 1,2 is nilpotent. Apply lemma A.6 to (77) and theorem 15 follows. Theorem 16 is valid by the same argument substituting $\hat{Y}^{(s+1)}$ for $\hat{Y}^{(s)}$ in the first term on the right of (73) and $\hat{T}^{(s)}$. Theorem 17 by substituting $\hat{Y}^{(s)} = \hat{Z}$ vec $\hat{\Pi}^{(s)}$ and $\hat{Y}^{(s+1)} = \hat{Z}$ vec $\hat{\Pi}^{(s+1)}$ complete the picture as section 4.

(b) The FP family: from (75) as in part (a)

$$\frac{\partial y^{(s+1)}}{\partial y^{(s)}} = P^{(s)} [(I-\beta) \otimes I_T] + A^{*(s)} \qquad (78)$$

where $A^{*(s)} = \hat{R}^{(s)} \psi^{*(s+1)}$ from (75), again nilpotent by lemmas 1 and 3. Apply lemma A.6 to (78) and theorem 7 follows using corollary A.3. Hence theorems 8,9,10 and the rest of section 3 follow in the obvious manner. #

Corollary 11;

The number of estimation iterations remains unaltered in full information and/or autoregressive versions from that of the limited information couterpart. Thus the estimators remain in the same relative positions for speed of convergence. #

Remember for autoregressive estimators where $y_{t-1,j} \in Z_i$ and $\Sigma_i \neq \sigma_i^2 I$, any one set of i,j, although the convergence criteria are valid, the statistical properties for forfeit at each estimation phase.

Theorem 24 and its corollary together with Chapter 7 answer the comment of Brundy and Jorgenson (1974) that the properties of autoregressive iterative estimators are unknown. In view of these results it will not be necessary to handle sphericality transforms in the estimators individually.

Theorem 24 is not unexpected since we achieve the same result considering another algorithm form to reach convergence with an iterative estimator under sphericality transform. We could iterate in two parts; treating $\phi^{(s)}$ as fixed, iterate to convergence in $\delta^{(s)}$ and $Y^{(s)}$ in the usual way; then treat $Y^{(s+1)}$ as fixed iterate to convergence in $\phi^{(s)}$ $\delta^{(s+\frac{1}{2})}$; repeat the cycles starting with $\phi^{(s+1)}$, $Y^{(s+1)}$, $\delta^{(s+1)}$ now available. Theorem 24 then follows because Theorem 1 of Oberhofer and Kmenta (1974) guarantees convergence in the second part of the cycle in all cases (see also Theorem 2.18) and convergence of the first part, and hence of the estimator as a whole, follows from the appropriate theorem of sections 3.1 or 4 above. Such algorithms are perhaps appropriate for RF instruments being fully iterated at each S. However there is no theoretical difference since S of convergence is the same in either type of algorithm. But the work done within the step S to S + 1 is obviously higher in these transforms and it is not recommended.

12.8 Existence theorems

We have assumed the nonlinear function in the iteration (1) implied by (4.27-28) to be Lipschitzian. If we can always justify the existence of at

least one fixed point Y*, we need then only investigate the problems implied by multiple solutions when this assumption is removed. We assume the existence of at least one Y* the moment estimation is undertaken - but nonconvergence could imply nonexistence, or a nonconvergent estimator (the latter possibility being removed by Theorem 14.1 below). A formal statement of the justification is

Theorem 25:[1]

Under the assumptions of at least one fixed point Y* exists in the mapping $F(\cdot)$ of (1).

Proof:

By the Oberhofer and Kenta (1974) Lemma. In that notation let a_1 be δ and a_2 be $X^{(s)}$ so U_1 and U_2 are the parameter and instrument spaces respectively. Then from Chapter 5, in IIV estimators f(a) is $L(\beta,\Gamma,\Sigma)$ which is continuous in T since U follows a multivariate normal distribution $N(0,\Sigma)$ and so $U_1 \times U_2$ is closed (on appropriate definiton) and bounded. Alternatively in FP estimators f(a) is the negative of the distance function (2.17) using (2.30). Thus an iterative maximisation procedure always yields a fixed point a* which therefore exists. #

However the fixed point need not be unique. The implication of statement (ii) of the Lemma is that by performing iterative maximisation over the whole space $U_1 \times U_2$ we can find the global maximum. With iterative estimation we should follow the same procedure and search the whole space by experimenting with different starts. Theorem 25 quotes assumptions which may be relaxed. For example U may follow any continuous multivariate probability distribution with finite variance. Or the model specification may be nonlinear if it is continuous in U and preserves these modal and continuity properties in the likelihood distribution. Thus the regularity

---

1. By Chapter 5, this holds in IIV estimators as a direct consequence of Lemma 2.1.

conditions for Theorem 25 are very general.

Theorem 26:

Theorem 25 is valid for uu' positive definite.

Proof:

The Lemma of Oberhofer and Kmenta holds in this case. Consider f(a) from Theorem 25. Partition $a = (a_1\ a_2\ a_3)$; $a_1 \quad U_1 \in R^{\Sigma(r_i+s_i)}$, $a_2 \in U_2 \leq R^{MxT}$, $a_3 \in U_3 \leq R^M$ i.e. a_3 is $\hat{\phi}$ and U_3 the residual space. Then $\exists$ t such that $T = \{a \mid a \in U_1 \times U_2 \times U_3,\ f(a) \geq t\}$ is nonempty and bounded and $U_1 \times U_2 \times U_3$ is closed. Then the iterative maximisation is defined as

(i) Let $a_2^{(0)} \quad a_3^{(0)}$ be an arbitrary start satisfying the given conditions and a_1 be such that $f(a_1 \quad a_2^{(0)} \quad a_3^{(0)}) \geq t$.

(ii) maximise $f(a_1\ a_2^{(0)})$ in U_1 getting $a_1^{(1)}$

(iii) maximise $f(a_1^{(1)} \quad a_2 \quad a_3^{(0)})$ in U_2 getting $a_2^{(1)}$ and soon in sequence

(iv) having obtained $a_1^{(j)} \quad a_2^{(j)}\ a_3^{(j)}$ all $0 \leq j \leq k$ we get $a_1^{(k+1)}$

from $\max_{U_1} f(a_1\ a_2^{(k)}\ a_3^{(k)})$, $a_2^{(k+1)}$ from $\max_{U_2} f(a_1^{(k+1)} \quad a_2 a_3^{(k)})$ and

$a_3^{(k+1)}$ from $\max_{U_3} f(a_1^{(k+1)} \quad a_2^{(k+1)} \quad a_3)$.

The Lemma follows the proof of Oberhofer and Kmenta applied to $\hat{a} = (a_1\ a)$, $\hat{a} = (a_2\ a_3)$, after it is applied to $f(\hat{a})$ alone at each step.

Theorem 26 then follows by identical reasoning to Theorem 25. #

CHAPTER 13: THE COMPUTATION OF ITERATIVE ESTIMATORS

13.1 Introduction

In this chapter we provide a computational analysis of the many iterative estimators, that have been mentioned in earlier chapters, to follow their statistical and convergence properties. This computational analysis is intended to illustrate the ease of use of iterative estimators and is undertaken in the conventional manner of numerical analysis; that is, analysing the algorithms for speed of convergence, and work done per iteration by operation counts. The aim is to provide a preference ordering over iterative estimators on computational grounds.

The preference ordering must of course be used in conjunction with the results on convergence criteria obtained earlier. However it is of particular importance since under classical assumptions we proved that members of the IIV family all give maximum likelihood estimates, so there is no discrimination possible between these estimators on either large or small sample properties. Given suitable starts, earlier iterates are asymptotically equivalent to maximum likelihood estimates and hence 2SLS. All members of the FP family give estimates asymptotically equivalent to maximum likelihood estimates given suitable starts. There is no discrimination here between iterative estimators and conventional ones. Chapter 16 argues that small sample properties are not generally available and even those results which are available do not permit effective discrimination between estimators.

Were we to discriminate by small sample properties we can still make no selection on statistical grounds within members of the FP family, or within the IIV family or between any IIV members and maximum likelihood estimates. We analyse the computational efficiency of the various algorithms to achieve a sharper estimator selection procedure. The other possible criterion - robustness - is inoperative since we are, with each iterative estimator aiming to

solve the nonlinear equation (5.10) so, within a group of estimators having identical small sample properties, robustness properties will also be identical.

The selection of the computationally efficient estimator is of particular importance because the computing burden in iterative estimation is very likely extremely heavy (e.g. Salmon and Eaton (1974)) if no steps are taken to reduce it. In many circumstances the burden of different estimators varies enormously. The problem is highlighted when it is required consistently to test several competing potential specifications, or update and reestimate the model.

In Section 2 we investigate how identities are best handled. In Section 3 we provide general expressions for the total operation count of the various estimators and then demonstrate weak separability to justify the estimator structure of chapter 11. Sections 4 and 5 provide a computational analysis within the class of iterative estimators. Section 6 extends this to comparisons with conventional estimators. The possibility that computations can be improved by least squares or IV estimates obtained from QR transforms is treated in Section 7. Finally we discuss any possible improvements by preselecting an optimal QRF scheme. Throughout illustrations are provided from the nine econometric models to demonstrate the significance of the results in practice.

Other aspects of numerical analysis, in particular numerical stability and accuracy, are of course relevant and are considered. However few general results are available and the importance of this aspect is even more dependent on a particular computing system than on the model itself, so it is not extensively analysed.

13.2 Identities in the Specification

13.2.1 In this chapter we relax the assumption of no identities in the

model (2.5). Using the same notation we write

$$I - \hat{\beta}^{(S)} = \begin{bmatrix} \hat{\beta}_{1g}^{(S)} & \beta_{1I} \\ \hat{\beta}_{2g}^{(S)} & \beta_{2I} \end{bmatrix}, \quad \hat{\Gamma}^{(S)} = [\hat{\Gamma}_g^{(S)} : \Gamma_I] \text{ and } U = [U_g : 0] \text{ with } Y = (Y_g : Y_I)$$

Y_g being the first g columns and Y_I the following M-g columns of Y; separating the first g equations being behavioural relations from the reamining M-g identities. The simplest treatment is to use the identities to determine the instruments for the next iterate given the known parameters, but ignore them at the estimation phase. The specification then implies, for $g < i \leq M$.

$$Y_i^{(S+1)} = Y_i^{(S)} \beta_i + Z_i \gamma_i \qquad (1)$$

but for $1 \leq i \leq g$ iteration as before. Then we have

Lemma 1:

The local convergence criteria of chapter 12 are all valid when identities are present: excepting the FP family with RF instruments when $(I-\beta^*)\beta^{-1}$ replaces $(I-\beta)\beta^{-1}$ where $\beta^* = \begin{bmatrix} \beta_{1g} & 0 \\ \beta_{2g} & 0 \end{bmatrix}$.

Proof:

(i) Theorem 12.4 directly in any of Theorems 12.7, 12.9 - 12.23.

(ii) In any event we note differentiating as in Theorems 12.7 - 12.23 the iteration (11.16) or (11.19) for $1 \leq i \leq g$ the elements of $J^{(s)}$ are as before but for $g < i \leq M$ from (1),

$$\frac{\partial y_i^{(S+1)}}{\partial y_{tj}^{(S)}} = \frac{\partial Y_i^{(S)}}{\partial y_{tj}^{(S)}} \beta_i \quad \text{or} \quad \frac{\partial Y_i^{(S+1)}}{\partial y_{tj}^{(S)}} \beta_i \qquad (2)$$

There are no terms in $\partial\hat{\beta}_i^{(s+1)}/\partial y_{tj}^{(s)}$. Hence the result follows for RF instruments in FP estimators (for $g < i \leq M$ all terms are on the LHS after differentiation). Any of the matrices whose eigenvalues were 0,1 in Theorems 12.7 - 12.23 were block diagonal matrices. For $g < i \leq M$ I_T will be the nonzero block and hence these matrices still have 0,1 eigenvalues. (Lemma 1 therefore follows immediately for all methods). Substituting $Z(\hat{\Pi}^{(s)})_i = Y_i^{(s)}$ in (2), as in (12.40), and proceeding as Theorems 12.9 and 12.17 covers parametric iteration. #

Comment:

The final M-g columns of the iteration matrix $I - \hat{\beta}^{(s)}$ are independent of S. For linear models these elements correspond to global conditions, although overall the criteria remain local. In nonlinear models $\partial F_i/\partial y_{tj}^{(s)}\big|_{Y(s)}$ depends on S even if F_i does not for $g < i \leq M$.

Some special cases of Lemma 1 have been noted in other work (Bergstrom (1974)., Bodin (1974)). We could now perform the computational analysis, treating identities in the simple fashion metnioned above. This is called Approach I. However there is an obvious possibility that instrument formation can be significantly accelerated using the a priori known coefficients where $Y_I^{(s)}$ is obtained recursively from $Y_g^{(s)}$ in a scheme involving only those coefficients. That scheme may be set up before iteration and need not be recalculated at each step. The possibility may be extended to eliminate the identities. (These possibilities are referred to as Approaches II and III respectively.) These issues have been investiated by Mitchell (1970), Agren (1972) and Lyttkens (1973) and we have to specify carefully which approach is being used in the computational analysis.

13.2.2 Approach I:

It is simplest for large systems. In direct instrument calculations are used we iterate

$$(Y_g^{(S+1)} : Y_I^{(S+1)}) = (Y_g^{(S)} \hat{\beta}_{1g}^{(S+1)} + Y_I^{(S)} \hat{\beta}_{2g}^{(S+1)} : Y_g^{(S)} \beta_{1I} + Y_I^{(S)} \beta_{2I})$$

$$+ Z(\hat{\Gamma}_g^{(S+1)} : \Gamma_I) \qquad (3a,3b)$$

having just obtained $\hat{\beta}_{1g}^{(S+1)}$, $\hat{\beta}_{2g}^{(S+1)}$ and $\hat{\Gamma}_g^{(S+1)}$ by OLS or IV in the estimation phase, as usual, for equations 1 .... g only. Clearly $Y_g^{(S+1)}$ and $Y_I^{(S+1)}$ may be calculated individually or together with impunity in this analysis. If, however, RF instruments are used we have

$$(Y_g^{(S+1)} : Y_I^{(S+1)}) = (Y_g^{(S+1)} \hat{\beta}_{1g}^{(S+1)} + Y_I^{(S+1)} \hat{\beta}_{2g}^{(S+1)} : Y_g^{(S+1)} \beta_{1I} + Y_I^{(S+1)} \beta_{2I})$$

$$+ Z(\hat{\Gamma}_g^{(S+1)} : \Gamma_I) \qquad (4a,4b)$$

where $\hat{\beta}_{1g}^{(S+1)}$, $\hat{\beta}_{2g}^{(S+1)}$ and $\hat{\Gamma}_g^{(S+1)}$ are obtained in the same way. $Y_g^{(S+1)}$, $Y_I^{(S+1)}$ may be determined either from the whole system written

$$Y^{(S+1)} = Z\Gamma^{(S+1)} \begin{pmatrix} I - \hat{\beta}_{1g}^{(S+1)} & - \beta_{1I} \\ - \hat{\beta}_{2g}^{(S+1)} & I - \beta_{2I} \end{pmatrix}^{-1} \qquad (5)$$

or

$$Y^{(S+1)} = Z \begin{pmatrix} \hat{\Gamma}_g^{(S+1)}(I - \hat{\beta}_{1g}^{(S+1)})^{-1} & \\ \Gamma_I & (I - \beta_{2I})^{-1} \end{pmatrix} \begin{pmatrix} I & - \beta_{1I}(I-\beta_{2I})^{-1} \\ -\hat{\beta}_{2g}^{(S+1)}(I - \hat{\beta}_{1g}^{(S+1)})^{-1} & I \end{pmatrix} \qquad (6)$$

(5) is the conventional RF instrument formation but (6) is (4a) and (4b) solved individually and then block simultaneously. (6) is computationally superior in a large system since inversion by partitioned matrices is faster. Although the dimension of submatrices to be inverted is unaltered there is substantial reductions in matrix multiplication due to the identity matrices, and one constant submatrix.

13.2.3 Approach II

The first extension of Approach I is to use group iterative methods in (3) or (4). Suppose we use (3),we could attempt to speed iterations by replacing $Y_g^{(S)}$ in (3b) by $Y_g^{(S+1)}$, given by (3a); that is impose a block Gauss-Seidel acceleration (or SOR if $\alpha \neq 1$ is used). Not only may such a device speed convergence but in some cases it may make it possible when it was not under (3). It may also allow some computational advantage to be taken of the a priori known coefficients. We get (writing E for exogeneous parts)

$$Y^{(S+1)}=Y^{(S)}\begin{pmatrix}\hat{\beta}_{1g}^{(S+1)} & 0\\ \hat{\beta}_{2g}^{(S+1)} & \beta_{2I}\end{pmatrix}\left[I-\begin{pmatrix}0 & \beta_{1I}\\ 0 & 0\end{pmatrix}\right]^{-1}+E \qquad (7)$$

whereas if we replace $Y_I^{(S)}$ in (3a) by $Y_I^{(S+1)}$ (i.e. do (7) in reverse order of blocks) we get

$$Y^{(S+1)}=Y^{(S)}\begin{pmatrix}\hat{\beta}_{1g}^{(S+1)} & \beta_{1I}\\ 0 & \beta_{2I}\end{pmatrix}\left[I-\begin{pmatrix}0 & 0\\ \beta_{2g}^{(S+1)} & 0\end{pmatrix}\right]^{-1}+E^{(S+1)} \qquad (8)$$

which is less convenient as both the inverse and $E^{(s+1)}$ must be recalculated at each iteration.

If behavioural equations are solved after the identity block, they should have coefficient estimates from using $Y^{(s+1)}$ to correspond with RFP - otherwise it is R_1FP. Conventionally behavioural equations are solved first, when the distinction disappears. Hereafter we take that to be the case. Then there is no change of character in the instruments but only an acceleration. This is also clear since either (7) or (8) may be used directly in the new form of (3a,3b), which in a large system with sparse β would be computationally preferable. For (7)

$$(Y_g^{(S+1)}:Y_I^{(S+1)})=(Y_g^{(S)}\hat{\beta}_{1g}^{(S+1)}+Y_I^{(S)}\hat{\beta}_{2g}^{(S+1)}:Y_g^{(S+1)}\beta_{1I}+Y_I^{(S)}\beta_{2I}) + E \quad (7a,7b)$$

The alternative is to replace $Y_g^{(S)}$ by $Y_g^{(S+1)}$ in (3a) or $Y_I^{(S)}$ by $Y_I^{(S+1)}$ in (3b), giving

$$Y^{(S+1)} = Y^{(S)} \begin{pmatrix} 0 & \beta_{1I} \\ \hat{\beta}_{2g}^{(S+1)} & \beta_{2I} \end{pmatrix} \left(I - \begin{pmatrix} \hat{\beta}_{1g}^{(S+1)} & 0 \\ 0 & 0 \end{pmatrix} \right)^{-1} + E^{(S+1)} \quad (9)$$

or

$$Y^{(S+1)} = Y^{(S)} \begin{pmatrix} \hat{\beta}_{1g}^{(S+1)} & \beta_{1I} \\ \hat{\beta}_{2g}^{(S+1)} & 0 \end{pmatrix} \left(I - \begin{pmatrix} 0 & 0 \\ 0 & \beta_{2I} \end{pmatrix} \right)^{-1} + E \quad (10)$$

respectively. Clearly(10) has the advantage. However it is not possible to eliminate $Y_g^{(S+1)}$ or $Y_I^{(S+1)}$ between (3a) and (3b), nor is it possible to perform (9) or (10) directly in the new form of (3a,3b). The instruments have evidently become partially RF in that one subvector in either case is dependent on the previous iterate of its complement and the current iterate of itself, (i.e.(3a) in (10).) However the computational advantage of (10) may be retained (especially for cases of sparse β where (10) is

unnecessarily burdensome) in

$$Y_g^{(S+1)} = Y_g^{(S)}\hat{\beta}_{1g}^{(S+1)} + Y_I^{(S)}\hat{\beta}_{2g}^{(S+1)} + E_g \qquad (10a)$$

$$Y_I^{(S+1)} = Y_g^{(S)}\beta_{1I}(I - \beta_{2I})^{-1} + E_I \qquad (10b)$$

This procedure is not a group acceleration (it is not possible to let the block size tend to one equation) but has, in (10a), merely introduced some RF instruments. This may be an advantage in itself in terms of speed of convergence, and it also exploits known coefficients to cut down the work per iteration. It is possible using direct instruments in the identity block to get repeated iterations for some model specifications (see Agren (1972), Lyttkens (1973)) where some equations have no exogenous part. The RF elements overcome this drawback.

2.4 Approach III

(10a,b) may be further speeded if $Y_g^{(S+1)}$ is used in (10b), which gives a group acceleration similar to (7a,b). On the other hand (7a,b) may have the advantages of partial RF instruments mentioned if $Y_I^{(S+1)}$ is used in (7b). In both, $Y^{(S+1)}$ is used on RHS of (3). And in either case, working from (3) we get on elimination of $Y_I^{(s+1)}$

$$Y_g^{(S+1)} = Y_g^{(S)}(\hat{\beta}_{1g}^{(S+1)} + \beta_{1I}(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)}) + Z(\hat{\Gamma}_g^{(S+1)} + \Gamma_I(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)}) \qquad (11)$$

We can therefore combine the acceleration of group Gauss-Seidel (or SOR) iteration with the advantages of partial RF instruments; which is obtained using $Y_g^{(S+1)}$ and $Y_I^{(S+1)}$ on RHS of (3b); and also comes from putting the computationally efficient (7) and (10) together.

We also obtain (11) if a direct instrument iteration is designed for the model with identities substituted out (i.e. a partial RF method). From (2.5)

eliminating Y_I we get

$$Y_g = Y_g(\beta_{1g}+\beta_{1I}(I-\beta_{2I})^{-1}\beta_{2g}) + Z(\Gamma_I(I-\beta_{2I})^{-1}\beta_{2g}+\Gamma_g)+U_g \qquad (12)$$

implying (11) for direct instruments.

(11) may be rewritten to give RF instruments (i.e. a purely RF instrument method)

$$Y_g^{(S+1)} = Z(\Gamma_I(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)}+\hat{\Gamma}_g^{(S+1)})(I-\hat{\beta}_{1g}^{(S+1)}-\beta_{1I}(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)})^{-1} \qquad (13)$$

which is exactly what results from substituting $Y_I^{(S+1)}$ from (4). Naturally (11) and (13) could be recast to give iterations in Y_I by obvious modifications to any of the above methods.

Group iterative methods may be used on (4). Corresponding to (7), replace $Y_g^{(S+1)}$ in (4b) by $Y_g^{(S)}$ yielding

$$Y^{(S+1)} = Y^{(S)} \begin{bmatrix} 0 & 0 \\ 0 & \beta_{2I} \end{bmatrix} \left(I - \begin{bmatrix} \hat{\beta}_{1g}^{(S+1)} & \beta_{1I} \\ \hat{\beta}_{2g}^{(S+1)} & 0 \end{bmatrix} \right)^{-1} + E^{(S+1)} \qquad (14)$$

This form is inconvenient since the inversion and $E^{(S+1)}$ must be recalculated at every iteration. This is always the case if only one endogeneous element of (4) is from $Y^{(S)}$. (14) may also be written

$$\left.\begin{aligned} Y_g^{(S+1)} &= [Y_I^{(S+1)}\hat{\beta}_{2g}^{(S+1)} + Z\hat{\Gamma}_g^{(S+1)}] (I - \hat{\beta}_{1g}^{(S+1)})^{-1} \qquad (15a) \\ Y_I^{(S+1)} &= [Y_g^{(S)}\beta_{1I} + Z\Gamma_I] (I - \beta_{2I})^{-1} \qquad (15b) \end{aligned}\right\}$$

and performed (15b) before (15a), which represents a substantial reduction in computation, the more so the more M > g. It is suggested (15) is a much faster method than (5) or (6), for introducing RF instruments into an iterative estimator, and can be further improved as

$$Y_g^{(S+1)}=\{Y_g^{(S)}\beta_{1I}(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)}+Z[\Gamma_I(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)}+\hat{\Gamma}_g^{(S+1)}]\}$$

$$(I-\hat{\beta}_{1g}^{(S+1)})^{-1} \qquad (16)$$

we can obtain expressions corresponding to (14) and (15) if $Y_I^{(S+1)}$ in (4a) is replaced by $Y_I^{(S)}$ to obtain a simplified way of calculating RF instruments. By eliminating $Y_g^{(S+1)}$ we obtain an iteration in Y_I similar to (16).

Alternatively we can create methods corresponding to (10). For (10) and (10a,b)

$$\left.\begin{aligned} Y_g^{(S+1)} &= Y_g^{(S)}\hat{\beta}_{1g}^{(S+1)} + Y_I^{(S+1)}\hat{\beta}_{2g}^{(S+1)} + E_g && (17a)\\ Y_I^{(S+1)} &= Y_g^{(S+1)}\beta_{1I}(I-\beta_{2I})^{-1} + E_I && (17b)\end{aligned}\right\}$$

and eliminating $Y_I^{(S+1)}$

$$Y_g^{(S+1)} = (Y_g^{(S)}\hat{\beta}_{1g}^{(S+1)}+E_g)(I-\beta_{1I}(I-\beta_{2I})^{-1}\hat{\beta}_{2g}^{(S+1)})^{-1} + E_I \qquad (18).$$

Again these iterations are a considerable computational improvement over the full RF procedure in a given iteration, both in having one inversion invariant to S and in eliminating one subset of the variables. Note iteration (17) itself, rather than (18), is impossible. However group iterative methods on (4) cannot preserve the full RF nature of the instruments since any attempt to obtain a subset of $Y^{(s+1)}$ recursively from one earlier

subset removes part of the simultaneity[1] in $Y^{(s+1)}$. (15) is simulaneous within blocks but not between, and (17) visa versa, so it makes no sense to argue one method is "closer" to providing RF instruments. Bother are partial RF.

We note no elimination of either Y_g or Y_I is possible in group iterations on (3).

13.2.5

Obviously there are many more ways of treating identities since there are many ways of substituting out the identities and many methods of grou iteration. Examples of Approach II are (7) - (10), (10a,b), (15a,b) (17a,b). Many Approach II methods can be converted to Approach III as (11), (13), (16) and (18). Approach II in the form (10) corresponds to Mitchell's scheme. Lyttkens suggests Approach III for the case of (11) obtained from (10). Agren's extension to cases IIa, IIb (his nomenclature) may be regarded as a special case of (11) or as alternatives in group iteration where some or all of the identities are substituted out. The rest of the explanation is new.

The estimation phase shows if any of these are practical. We can obtain $\hat{\beta}_{1g}^{(S+1)}$ $\hat{\beta}_{2g}^{(S+1)}$ and $\hat{\Gamma}_g^{(S+1)}$ from the estimating equation

$$Y_g = Y_g^{(S)} \hat{\beta}_{1g}^{(S+1)} + X^{*(S)} \hat{\beta}_{2g}^{(S+1)} + Z\hat{\Gamma}_g^{(S+1)} + \hat{U}_g^{(S)} \qquad (19)$$

where $X^{*(S)} = (Y_g^{(S)} \beta_{1I} + Z\Gamma_I)(I - \beta_{2I})^{-1}$, for use in any Approach III method. However we run the risk of substantially increased multicollinearity. In

---

1. Unless $Y_I^{(S+1)}$, say, only was calculated from (6) and used (15a) or(17a) That requires the complete inversion in (6) which, if done, it is inefficient not to complete the job in (6).

(11) and (13) $X^{*(S)}$ can be used to obtain $Y_g^{(S+1)}$ directly but in (16) and (18) a considerable amount of extra computation is required including an inversion of order g. Thus, although all iterations can be described without Y_I, (16) and (18) will be discarded as computationally inefficient in favour of (11) or (13). It seems likely that the same conclusion will follow from consideration of other alternatives to (11) and (13).

13.2.5

The generalisation of any of the methods in Approaches I, II or III to fractional or recursive estimators is immediate and obvious. The generalisation to endogeneous iteration is similarly straightforward. In Approaches I and II we take the i^{th} equations $1 \leq i \leq g$ and proceed as (11.16) or (11.19) and complete the iteration with (3b), (7b) or (10b). In Approach III we proceed as (11.16) or (11.19) for all $1 \leq i \leq g$ on the system

$$Y_g = [Y_g^{(S)} : X^{*(S)}] \begin{bmatrix} \hat{\beta}_{1g}^{(S+1)} \\ \hat{\beta}_{2g}^{(S+1)} \end{bmatrix} + Z\hat{\Gamma}_g^{(S+1)} + \hat{U}_g^{(S)} \qquad (20)$$

involving therefore only a redefinition of the variable $Y_i^{(S)}$ of (11.16) or (11.19). For RF instrument versions we proceed as (20) with redefined $Y_i^{(S+1)}$ as in (11.33), (11.33a) and (11.33b) etc.

The extension to parametric iteration is again immediate when $Z\hat{\Pi}^{(S)}$ replaces $Y^{(S)}$ at all points including the identities in equations (3) to (10a,b). In Approach III we again use redefined variables as in (20) but obtain parameter estimates from (19). RF instrument versions may be then used as usual if required.

13.2.7 Convergence:

Naturally different methods of handling the identities affect the iteration matrix and convergence speed. For Approach I we use Lemma 1. Even if the character of the instruments is preserved in Approach II the iteration matrix is not. Viewing the examples given

Theorem 1:

The local convergence criteria of (7), (10), (16) and (18) are that the spectral radii of the following matrices should be less than unity for FP or IIV,

(a) $\begin{pmatrix} \beta_{1g} & 0 \\ \beta_{2g} & \beta_{2I} \end{pmatrix} \left[I - \begin{pmatrix} 0 & \beta_{1I} \\ 0 & 0 \end{pmatrix} \right]^{-1}$;

(b) $\begin{pmatrix} \beta_{1g} & \beta_{1I}(I - \beta_{2I})^{-1} \\ \beta_{2g} & 0 \end{pmatrix}$;

(c) $\beta_{1I}(I - \beta_{2I})^{-1} \beta_{2g}(I - \beta_{1g})^{-1}$;

and

(d) $\beta_{1g}(I - \beta_{1I}(I - \beta_{2I})^{-1}\beta_{2g})^{-1}$ respectively

Proof:

In each case we determine $J^{(S)}$ and then take the local convergence criterion evaluated at Y^* as indicated by Theorem 12.1.

(a) Application of the iteration matrix of a group recursive acceleration in Chapter 10 to the system covered by Theorems 12.7 or 12.15; or direct application of Theorems 12.7 or 12.15 to (7) itself as a respecified FP or IIV iterative estimator.

(b) (10a,b) may be written as a complete FP, IIV estimator with iteration

$$Y^{(S+1)} = Y^{(S)} \begin{pmatrix} \hat{\beta}_{1g}^{(S+1)} & \beta_{1I}(I-\beta_{2I})^{-1} \\ \hat{\beta}_{2g}^{(S+1)} & 0 \end{pmatrix}$$

and then we apply (2) and Theorems 12.7 or 12.15 to this system.

(c) (15) may be rewritten as (16), and since the a priori restrictions will certainly be honoured we can treat (16) as an FP or IIV estimator. Hence $J^{(S)} = \partial Y_g^{(S+1)}/\partial Y_g^{(S)} = \beta_{1I}(I-\beta_{2I})^{-1}\,\hat{\beta}_{2g}^{(S+1)}(I - \hat{\beta}_{1g}^{(S+1)})^{-1}$ - the details may be checked by Theorems 12.7 or 12.15 and 12.9. Note Theorems 12.16 or 14.18 do not apply here since $J^{(S)}$ follows from

$$\frac{\partial Y_g^{(S+1)}}{\partial y_{gtj}^{(S)}}\,(I - \hat{\beta}_{1g}^{(S+1)}) - Y_g^{(S+1)}\,\frac{\partial \hat{\beta}_{1g}^{(S+1)}}{\partial y_{gtj}^{(S)}}$$

$$= \frac{\partial Y_g^{(S)}}{\partial y_{gtj}^{(S)}}\,\beta_{1I}(I - \beta_{2I})^{-1}\,\hat{\beta}_{2g}^{(S+1)} + Y_g^{(S)}\,\beta_{1I}(I - \beta_{2I})^{-1}\,\frac{\partial \hat{\beta}_{2g}^{(S+1)}}{\partial y_{gtj}^{(S)}} \qquad (21)$$

and we use (12.28) or (12.48) appropriately (i.e. RHS $\neq$ 0 even for IIV).

(d) Obvious from and very similar to (c). #

Corollary 1:

Fractional and recursive estimators have local convergence conditions that the real part of any root of (a) - (d) respectively be less than unity. Parametric iterative estimators have the same conditions (check: Theorems 12.9 and 12.17 in place of those others mentioned). #

For Approach III methods

Theorem 2

(a) $\rho(\beta_{1g} + \beta_{1I}(I - \beta_{2I})^{-1}\,\beta_{2g}) < 1$ is the convergence criterion for (11)

(b) for (13) it is the matrix in part (a) substituted for $I - \beta$ in Theorem 12.8 for FP, but for IIV estimators a zero iteration matrix (always convergent).

Proof:

Proceding as in Theorem 1

(a) Obvious using the same methods as Theorems 12.7 or 12.15 on (11).

(b) Immediate since (13) is shown as the RF instrument version of (11) so we apply Theorem 1^2.8 if FP, or Theorems 1^2.16, 1^2.18 if IIV. #

13.2.8

Finally we have to decide which of the these approaches are to be used. We may use (11), (13), (7), (18) or (10) or Approach I. For comparisons among iterative estimators, it is irrelevant which approach is used provided the same is used for each to preserve strict compatability. Since the computational burden of an iterative estimator is influenced by the speed of convergence it is desirable to preserve the character of the instruments and use the theorems of the previous chapter. We will use Approaches I or III - the precise form depending on the desired instrument formation.

Clearly comparing the iteration matrices of Chapter 12 and those of theorem 1, we can make no general statement on which approach is preferable. There is an expectation approach III will often be better but this is not always the case for all models and methods (compare tables A3 and A24 below). Therefore, apart from pointing out that eliminating the identities can make convergence possible for models otherwise nonconvergent, no further recommendation can be made on theoretical grounds (except IIV estimators with RF instruments have convergence speed unchanged).

13.3 The Computational Burden of the General Iterative Estimator:

13.3.1

We will assess and compare the computational burdens of the various iterative estimators by means of operation counts. The use of operation counts in numerical analysis is standard and given the algorithms in use, the operation counts may be easily built up from the following evaluations:

1. Inversion of n x n matrix $\frac{1}{2}(n^3 + n^2)$ multiplications; for a symmetric matrix $\frac{1}{4}(n^3 + n^2)$ multiplications.
2. Inner product of n x p and p x q matrices; npq multiplications and additions.
3. QR decomposition of n x p matrix; $\frac{n}{2}((n+1)(p+1)+20)$ multiplications and $n((n+1)(p+1)+3)$ additions.
4. The number of iterations in n x n matrix iterative scheme on matrix B is approximately $\log_e \tau / \log_e \rho(B)$ for tolerance τ.
5. Operation count of square root operation on scalar = 7 multiplications.

The inversion routine referred to is the product form of the inverse which is given by Westlake(1968)as computationally the most economical.It is quite possible that in some cases that this routine is not the most satisfactory in terms of numerical stability, rounding errors and accuracy. However there seems little consensus of opinion on which routine is most satisfactory, indeed it is almost certainly more influenced by the available computing power, and the accuracy of its arithmetic, than anything else. It is not within the scope of this analysis to answer that debate but of course the computational analysis could be reworked with whatever inversion routine is desired (and most are detailed in Westlake). The QR decomposition refers to Householder transformations which produce the QR form,as given by Businger and Golub(1965).The Householder transformations may not be ideal in all respects but are speedy, stable, readily available and have appeared before in econometric literature (Belsley(1974)).The operation count is taken from the routine inBusinger and Golub with a comparison taken as equivalent to an addition. We note the evaluation of the number of iterations is only approximate - especially where $\rho(B)$ is small. Finally the expression for a square root is very approximate and is an average figure for the PDP 8 software. It is not very significant here.

Estimated computational burdens are therefore rather approximate. Even if the formulae were precise the comparisons would be rather crude. No allowances are made for speed of accessing information, nor for the time in transferring information in and out of memories, for storage requirements, problems of disc usage, for recording intermediate results, etc. We include the additions in operation counts by defining q as the ratio of time taken for one multiplication to that for one addition. Values of q for various machines in current use as follows: for single precision calculations done individually with information already in CPU (for double precision all figures rise).

Table 1

q values

IBM	1800 series	1.27	PDP 8	1.0
IBM	360 series	4.27	PDP 11/45	2.0
IBM	370 series	2.28	CDC 7600	1.25
ICL	4/75	2.13	UNIVAC 1108	4.2

(Figures supplied by the manufacturers)

Hence $1 \leq q \leq 5$ was used, plus $q = 10$ to give some indication of results under double precision. It appears that additions are not often included in numerical analysis (e.g.Belsley(1974),Businger and Golub(1965),etc.) but it is important here in view of the substantial number of additions and the reduction of q towards unity in modern computers. The other shortcomings and inaccuracies are insignificant compared against the operation count, which includes both multiplications and additions, and are impractical for any general measures.

13.3.2

The discussion of Chapter 11 demonstrated, intuitively, that the structure of the class of iterative estimators may be analysed from seven independent decisions. It is obvious that this would be the case since each decision concerns one step in building up the estimator - each has an estimation phase, an iteration variable(s), the same option to linearity, to RF instruments, to various accelerations and so on. Clearly at each step one option can, if desired, be taken without affecting the list of options at any other step. That discussion was provided for explanatory purposes to highlight the family structure of the general estimator.

However we formally establish weak separability of the decisions to provide a simple method of selecting a computationally efficient estimator. It is apparent that a very large number of iterative estimators are available. If we consider only the options of chapter 11 we have 384; allowing five kinds of starts, $Y^{(0)} = 0$, $Y^{(0)} = Y$, PC, (4.26) or dummies, all of which are generally available, and considering FI, AR and FIAR versions, we have over 5,000; and that by no means exhausts the possibilities. In section 1 we argued the necessity of selecting the computationally efficient estimator. We note that the number of operations incurred by any decision is a discreet variable (a positive real number, since q is real). Therefore if we express the computational burden of the general estimator in terms of the operations incurred at each decision (each expressed in terms of the characteristics of the model, M, g, T, K, r_i, s_i, i = 1 .... M) the minimum must be selected by direct comparison, since the lack of continuous first order derivatives precludes an analytic decision rule.

Clearly 384 comparisons is a lot of work. More effective would be a branch and bound algorithm applied to the decision tree. In either case substantial extra calculation is required, which may or may not be worthwhile, but detracts from the attractiveness and competitiveness of iterative estimators. Further computer programmes would be required. Moreover a number of possible combinations of options are clearly inefficient (e.g. OLS or IV estimation with endogeneous iteration) in all or certain, a priori testable circumstances, and any estimator embodying such combinations should be passed over. In this context it is very easy to derive simple selection rules for certain options within a decision, a fact which should be exploited. Thus a number of branches may be pruned at the start, or enroute, and it looks attractive to demonstrate complete separability of decisions so that the computations may be minimised at each decision and thus overall - as for example in the well known multistage optimisation of consumer budgeting popularly associated with separable functions. However some, not necessarily all, options are independent at each decision - in particular the computation minimising ones may not be. We may therefore try instead to combine the pruning of branches and the simple written decision rules with the degree of separability present, to circumvent the necessity of a branch and bound algorithm. We note that the convergence criteria of Chapter 12 fall into the three groups defined by $I - \beta$, $-(I - \beta)\beta^{-1}$ and zero iteration matrices. This enables the 384 comparisons to be reduced to three and avoids further computations.

Consider any function $F(X)$, $X = (x_1 \ldots x_m)$ independent continuous variables ϵR^m with mutually exclusive subsets $\{X_i | X_i \subset X, i = 1 \ldots n\}$ and $\bar{X}$ the complement of $X_1 \ldots X_n$ in X. If $F^*(\cdot)$, $f_i(\cdot)$ $i = 1 \ldots n$ exist with the partial derivatives, $F'(X)$, $F^{*\prime}(f_1(a_1)\ldots f_n(a_n), a_m)$; $f_i'(a_i)$ continous and nonvanishing, in some neighbourhood of arbitrary points $f_i(a_i)$, a_m and $a_i \epsilon R^m$, $i = 1 \ldots n$. respectively. Then $F(X)$ is locally separable with respect to X_i if $F(X) = F^*(f_1(X_1) \ldots f_1(X_n), \bar{X})$. (Leontief (1947a)). We use the following : (Theorem I of Leontief (1947b) for proof).

Theorem 3:

F(X) is locally separable with respect to X_i under the conditions given here iff

$$\frac{\partial\left(\frac{\partial F}{\partial x_{ri}} \Big/ \frac{\partial F}{\partial x_{rj}}\right)}{\partial x_{pk}} = 0 \quad \text{all } p\ r\ i\ j\ k \text{ such that } p \neq r,\ \ p,r = 1 \ldots n;$$

$k = 1 \ldots m_p$; $i,j = 1 \ldots m_r$; X_p and X_r of length m_p and m_r respectively. #

To apply Theorem 3 to the case where x_i $i = 1 \ldots m$ are discreet and hence the partial derivatives locally discountinuous we need

Lemma 2:

Let x_i $i = 1 \ldots m$ of Theorem 3 be discreet. Consider one set of such values $x_i = a_i$. Define $Y = (y_1 \ldots y_m)$ independent continuous variables εR^m where y_i corresponds to x_i and further define Y_i corresponding to X_i and $\bar{Y}$ their complement in Y. Then F(X) is locally separable with respect to X_i iff F(Y) is to Y_i (in which case $F(X) = F^*(f_1(x_1) \ldots f_n(x_n), \bar{X})$) under the same regularity conditions on $F^*(\cdot)$ and $f_i(\cdot)$ and $F(\cdot)$.

Proof:

By Theorem 3 on F(Y) for its local separability. Then obtain the precise form of $F^*(\cdot)$ and $f_i(\cdot)$, $i = 1 \ldots n$. by proposition V Leontief (1947a).

Since each y_i is continuous over at least one possible discreet value of x_i, $F(X) = F^*(f_1(X_1), \ldots, f_n(X_n), \bar{X})$ holds as a particular value of $F(Y) = F^*(f_1(Y_1), \ldots, f_n(Y_n), \bar{Y})$ just obtained. Considering the sequence $x_i = a_{ik}$, k indexing all values of x_i, F(X) is separable at all points. #

For any estimator in the decision tree the total computational burden as given by the total operation count

$$C = (f_1 + f_2)\, S_1\ f_6\, f_7 + f_4 \left(\sum_{j=0}^{S_1 f_6 f_7} S_{2,j}\, f_{5,j}^{-1} \right) + Q \qquad (25)$$

Q = operation count of initial calculations, including starts.

f_i = operations incurred at decision i, i = 1 .... 7, for any estimator. Therefore in a given model f_i may take on C_i distinct discreet values, C_i being the number of options open at decision i. f_1 has $C_1 = 5$; IV or OLS at each step S, OLS with $Z\hat{\Pi}^{(S)}$ in place of $Y^{(S)}$ (parametric), or IV (parametric), or zero for endogeneous iteration. f_2 comprises decisions 2 and 3 together and hence $C_2 = 4$; mixed or endogeneous iteration, parametric or in $Y^{(S)}$ [1]. f_4 has $C_4 = 5$; RF or QRF instruments (any option under f_2) or direct ($f_4 = 0$). Note a wider range of estimators are implied here both at decision 1 and 4, (about four times as many). $f_{5,j}$, f_6 and f_7 are factor reductions in the number of iterations from accelerations in instrument formation, estimation iteration and the superimposition of SI acceleration respectively. Each is therefore defined as $\frac{\log \rho_1<1}{\log \rho_2}$, ρ_1 and ρ_2 being spectral radii of iteration matrices before and after the acceleration selected. $S_1 = \frac{\log \tau_1}{\log \rho(B)}$, B = iteration matrix of estimator chosen and

$$S_{2,j} = \frac{\log \tau_2}{\log \rho(I-\hat{\beta}^{(j)})}$$ for estimation and instrument formation tolerances τ_1 and τ_2.

We note instrument formation has always the same iteration matrix $(I - \beta^{(j)})$ at iteration j. If RF instruments are obtained by inversion (10.22) the right hand term of (25) must be replaced appropriately, but that is less general. Note S_1 depends on f_1 and f_4; f_4 on f_2.

Separability as defined in Theorem 3 is only a local phenomenon. Our definition of S_1 and $S_{2,j}$ are only approximate as usual and only locally valid by Theorem 12.1. The sequence $S_{2.j}$ will depend on the start

---

1. The option to nonlinear iteration is ignored as it has in effect already been rejected on computational grounds.

and speed of convergence, i.e. S_1, f_6, f_7. Thus it is sufficient for local separability to consider (25) within the same neighbourhood of Y*, as we used for convergence analysis, i.e. replace $\beta^{(j)}$ with β in S_2 ($f_{5,j}$ then becomes independent of j), and we have S_2 a constant in

$$C \simeq [(f_1 + f_2) + f_4(S_2 f_5 - 1)] S_1 f_6 f_7 + Q \qquad (26)$$

Theorem 4:

(a) C in (26) is locally separable with respect to f_6 and f_7.

(b) If S_1 is held constant, C in (26) is locally separable with respect to the subgroups, $(f_1 f_2)(f_4)(f_5)(f_6)(f_7)$.

Proof:

The evaluation of operation counts implies each option makes f_i a polynomial, $f_i = p_{ik}(M, g, T, K, r_j, s_j)$ $j = 1 \ldots M$, $k = 1 \ldots C_i$, $i = 1 \ldots 7$ of order 3 or less. Define new variables y_i, as in Lemma 2, corresponding to f_i $i = 1 \ldots 7$ and y_8 for S_1. We have $C'(y_1 \ldots y_8)$ of (26) satisfies the regularity conditions on $F(\cdot)$ in Theorem 3. Proposition V Leontief (1947a) ensures if we can write (26) as $C = C^*(C_1^*(Y_1) \ldots C^*_n(Y_n), \bar{Y})$, the regularity conditions on $C^{*\prime}(\cdot)$ and $C^*_i{}'(\cdot)$ must be satisfied for neighbourhoods $C^*_i(a_{ik})$, a_{mk} and a_{ik} respectively $k = 1 \ldots C_i$. Let the second order partial differential of Theorem 3 be $R_{ij \cdot k}$ $\{y_i, y_j\} \subset Y_r$ and $y_k \in Y_p$ and $Y_r \cap Y_p = 0$.

(a) $R_{ij \cdot 6}$ and $R_{ij \cdot 7}$ vanish for i,j = 1 .... 5. $R_{ij \cdot 8} \neq 0$ all i, j = 1 7

since $\partial y_8 / \partial y_1$ and $\partial y_8 / \partial y_4$ do not vanish. Apply Lemma 2 and the result follows.

(b) Make S_1 = constant. Now $R_{ij \cdot k} = 0$ i,j = 1,2 and k = 4,5. Again the result follows from Lemma 2 and part (a); decision 5 affects C through S_2 a constant, whatever f_4, implying the separability of $(f_4)(f_5)$. f_1 and f_2 are not fully independent and hereafter are treated as one decision. #

Let $\lambda = \rho(I - \beta)$ and $\lambda^* = \rho(-(I - \beta^*)\beta^{-1}$ of Lemma 1. Then $S_1 = \frac{\log \tau_1}{\log \lambda}$ if direct instruments are used, $S_1 = \frac{\log \tau_1}{\log \lambda^*}$ if RF instruments are used in any FP estimator and $S_1 = 0$ if they are used in an IIV one. It is simple to compare directly the following three expressions for each estimator in those classes,

$$C_D = (f_1 + f_2)\, f_6\, f_7 \frac{\log \tau_1}{\log \lambda} + Q_D \tag{27}$$

$$C_{RF/OLS} = [(f_1 + f_2) + f_4 \frac{\log \tau_2}{\log \lambda} f_5]\, f_6\, f_7 \frac{\log \tau_1}{\log \lambda^*} + Q_{RF/OLS} \tag{28}$$

$$C_{RF/IV} = [(f_1 + f_2) + f_4 \frac{\log \tau_2}{\log \lambda} f_5]\, f_6\, f_7\, a + Q_{RF/IV} \tag{29}$$

Note that if M = g, or if Approach III is used ((11) or (13)), $\log \lambda^* = \log (1 - \lambda) - \log \lambda$. In the case of Approach III λ refers to the matrices given in Theorem 2, not $I - \beta$ itself.

In (27) - (29) S_1 is held constant so Theorem 4 (b) may be applied. Hence we have

Corollary 2:

Minimum $C = \min \{C^*_D, C^*_{RF/OLS}, C^*_{RF/IV}\}$, where C^*_D, $C^*_{RF/OLS}$, $C^*_{RF/IV}$ are the minimum of (27) - (29) obtained by minimising independently, f_1 and f_2 together, f_4, f_5, f_6, f_7 (note, f_4 in (28) and (29) is never zero). #

An interesting divergence between theory and practice comes to light at (29). The constant a stands for the number of estimation iterations for that class of estimators. Naturally from the theory of Chapter 12, a = 0. Thus in theory IIV estimators with RF instruments are all automatically equally the best estimators! The theoretical analysis therefore ends at this point. It remains only to discover the most efficient IIV estimator with RF instruments and the recommendation is to use that, if it is superior to conventional methods.

Not suprisingly IIV estimators with RF instruments do not converge instantly - RFIIV generally converges in about ten iterations, apparently independently of the mode (see e.g. Bergstrom (1974), Hausman (1975)/ Brundy and Jorgenson (1974)/Maddala (1971)). For practical results and recommendations we have therefore to continue with the corollary above. a stands for that, approximately constant, small number of iterations that will be met in practice. At least five reasons can be advanced why $a \neq 0$ is found:

(i) the relation $a \simeq \frac{\log \tau_1}{\log \lambda}$ is only approximate to within a few numbers of iterations independently of a, and hence especially so as $\lambda \to 0$ at a rate $\to \infty$;

(ii) even if convergence were instantaneous[1] two or three (depending on the information class) iterations are necessary for consistency and asymptotic efficiency - and more if $\phi^{(S)}$ is to converge. Thus $a > 2,3$ in any event;

(iii) if RF instruments are obtained by iteration there is always the error $\eta^{(S+1,k_{S+1})} \neq 0$. If the instruments were obtained by inversion, in any inversion routine there will be an accumulation of rounding errors and thus in practice a nonzero sequence of errors in estimation iteration, defined by $Y^{(S+1)} = Z\hat{\Gamma}^{(S+1)}((I - \hat{\beta}^{(S-1)})^{-1}_{*} + \varepsilon^{(S+1)})$ where * indicates inversion with no rounding errors. In either case $\eta^{(S+1,k_{S+1})} = \varepsilon^{(S+1)} = 0$ all S if the zero iteration matrix is to apply (by (12.82) and Theorem 12.16 etc.). Inevitably we have an element of direct instrument formation in the iteration matrix equivalent to having used QRF instruments (i.e. Theorem 12.22 applies). Now the theorem is only locally correct so the start may influence it.

(iv) In FI and AR versions the RF instrument set is correctly $\hat{N}^{(S+1)} = (\Sigma^{(S+1)} \otimes I)^{-1}\hat{X}^{(S+1)}$ not $(\Sigma^{(S)} \otimes I)^{-1}\hat{X}^{(S+1)}$. In other words $\phi^{(S)}$ must be fully iterated through to $\phi^{(S+1)} = \beta^{-1\,(S+1)'}V^{(S+1)'}V^{(S+1)}\hat{\beta}^{-1(s+1)}$ (or AR equivalently). This may not always be done.

---

1. Recall we regard 2SLS and 3SLS as instantaneous iterative estimators since $Y^{(S)} = X\Pi^{+}$ is used for all S.

(v) RFIIV has a midpotent iteration matrix with only zero expected value.

(vi) Possibly identities have been treated to give other than purely RF instruments as (5) or (13).

The fact remains, a being very small tends to dominate even the practical minimisation of (26). There appears to be no method of predicting a, except by guessing a small number, and as also λ is unknown a priori, we concur with the theoretical analysis in that we expect (non mathematically) (29) to be the optimum, and it is therefore likely the best of (27) - (29). It will be recommended formally at the end of the chapter in that both computationally and statistically that IIV estimators with RF instruments are the best iterative estimators. But we need to check the few RF instrument formations are not likely to outweigh a large number of estimation iterations, and also to discover the most efficient such IIV estimator.

13.4 Comparisons of computation of iterative estimators:

Lemma 3:

Among FP estimators where M = g, or under Approach III, estimation iterations are minimised by (27) if $\lambda \in [0,1)$ but by (28) if $\lambda \in [-\infty,0]$.

Proof:

Obvious; also for accelerations with similar parameters. #

We now derive simple rules for selecting the computationally efficient estimator. Within each group (27)-(29) we seek only the best, but no preference ordering. The results of Section 4 and 5 of this chapter depend only on the total operation count implied by each estimator or each decision and comparing that with another comparable operation count. The decision rule is presented in terms of the characteristics of the model (M, g, T, K, r_i, S_i all non-negative, i = 1 .... M) and is illustrated with reference to the nine models. In either case the derivation of the results is quite straightforward but extremely tedious. Therefore no derivation is presented, but with the aid of the operation evaluations in Section 3 the enthusiastic reader may easily reconstruct each case. The decision rules frequently turn out to be

very complicated due to the burden of initial calculations and therefore unilluminating. Therefore simplified approximate rules are presented here and the precise rule in each case is given in Appendix 2 of this Chapter (assuming the same $Y^{(0)}$ is used for each estimator compared). The numerical illustrations however are worked out including initial calculations where relevant.

13.4.1 Direct instrument estimators (27):

Using the separability of f_6 and f_7, we have only to consider $\min(f_1 + f_2)$ in (27) and minimise f_6 and f_7 separately. f_1 and f_2 both have four options. However mixed iterations always require $f_1 \neq 0$. Thus some two options are removed.

We note first that $f_1 \neq 0$ implies OLS estimation is best done with partitioned matrices

$$\hat{\delta}^{(S)} = \begin{pmatrix} Y_i^{(S)'} Y_i^{(S)} & Y_i^{(S)'} Z_i \\ Z_i' Y_i^{(S)} & Z_i' Z_i \end{pmatrix}^{-1} \begin{pmatrix} Y_i^{(S)'} y_i \\ Z_i' y_i \end{pmatrix} \qquad (30)$$

since $Y_i^{(S)'} Z_i$ may be computed for $Z_i' Y_i^{(s)}$ by symmetry, and both $Z_i'Z_i$ and $Z_i'y_i$ need only be calculated once. We get an operation count under Approach I (g equations estimated), per iteration

$$\sum_g (r_i+s_i)^2 ({}^q/_4 (r_i + s_i + 5) + 1) + T(q + 1) \sum_g r_i(r_i + s_i + 1) \qquad (31)$$

In the case of IV estimation we need $Y_i^{(S)'} Y_i$ and $Z_i'Y_i$ in (30) implying an increase of $\sum_g s_i r_i T(q + 1)$ in initial calculations only. Considering endogeneous iteration from (11.16) and (11.19) we see that EIIV saves $(q+1)T^2 \sum_g r_i$ every iteration, at a cost of $T\sum_g r_i$ extra additions and $2(q+1)T^2\sum_g r_i$ initial calculations. Repeating these two cases for parametric estimators with $Z(\hat{\Pi}^{(s)})_i$ for $Y_i^{(s)}$ from (30) we get the computational burden

$$\sum_g (r_i + s_i)^2 \left({}^q/_4(r_i + s_i + 5) + 1\right) + K(q+1)\left[\sum_g r_i(r_i+s_i+1+K)\right] \qquad (32)$$

with $(q + 1)KT(K_g + \sum_g s_i + g)$ extra initial calculations. The IV version gives an extra $(q + 1)\, T \sum_g [s_i r_i + K(r_i + s_i + 1)]$ initial calculations over (30) but saves $(q + 1)K^2 \sum_g r_i$ in (32). In endogenous iteration the parametric versions of (11.16) saves $(q + 1)K^2 \sum_g r_i$ over (11.19b) with a decrease in initial calculations of $(q + 1)KT(Kg - \sum_g r_i)$.

Any of the comparison theorems which follow in this section 4 are strictly only valid if the estimators need exactly identical numbers of estimation iterations. All the comparisons made are for estimators with identical iteration matrices (chapter 12). The comparison theorems are therefore only approximate since (a) differing numerical algorithms may produce slightly different sequences $\beta^{(j)}$ and hence numbers of estimation iterations - these iteration matrices being only evaluated precisely at convergence (b) the convergence results are only locally valid by theorem 12.1. However prior to estimation this is our best and only information and has therefore to be assumed.

Lemma 4:

(a) For all estimators with direct instruments, endogeneous (but visa versa mixed) IIV estimators are preferable to their FP counterparts under nonparametric iterations.

(b) For the same class of estimators under parametric iteration, IIV is always preferred.

Proof:

The corresponding estimators mentioned differ only in their estimation phase. (a) follows unambiguously including initial calculations since any of these estimators requires more than one estimation iteration. The superiority of FP over IIV is negligable but EIIV over EFP is nontrivial.

(b) the result holds precisely for endogeneous estimators since $Kg > \sum_g r_i$ is a consequence of assuming the model to be identified i.e. $(K+1)g \geq \sum_g (r_i+s_i)$, $g>1, \Sigma s_i \neq 0$ is implied. The result is also precise for mixed iteration if $K(Kg - \Sigma r_i) - \Sigma r_i s_i \geq 0$, which indeed holds since $Kg - \Sigma r_i + g \geq \Sigma s_i$ and $\sum_g s_i(K - r_i) \geq 0$. #

The range of options to be considered is now much reduced.

Comment:

In view of evaluation of operations in matrix multiplication, it is obviously most economical to multiply in sequences starting with the pre- or post - multiplication of matrices with the smallest diemensions (ideally a vector) e.g. in EFP (11.19) we obtain the following sequence

$$M_i Y_i^{(S)},\ Y_i^{(S)'} M_i Y_i^{(S)},\ (Y_i^{(S)} M_i Y_i^{(S)})^{-1},\ Y_i^{(S)} M_i) y_i,\ \hat{\beta}_i^{(S+1)},$$

$$Y_i^{(S)} \hat{\beta}_i^{(S+1)},\ M_i (Y_i^{(S)} \hat{\beta}_i^{(S+1)}).$$ This convention is assumed hereafter.

Lemma 5:

In parametric endogenous iteration (11.19c) is preferred to (11.19a) for first estimated equations. [For (11.19c) compared to (11.19b) - see theorem 6.]

Proof:

$\hat{\beta}_i^{(s+1)}$ is calculated identically in either form for both EPFP and EPIIV. But (11.19c) saves the computation of $M_i Z_i$, M_i and $I - M_i$ at a cost of I_i^* (no computation) and C_i^* (which involves less computation than M_i). Further the multiplication of $(\hat{\Pi}^{(s)})_i$ by $M_i Z_i$ is avoided since I_i^* annihilates the rows of $(\hat{\Pi}^{(s)})_i$ if $\gamma_{ij} = 0$ $j = 1 \ldots M$ and retains the rest. #

Lemma 6:

Identities in endogeneous estimators are best handled as $y_i^{(S+1)} = Y_i^{(S)}\beta_i + Z_i\gamma_i$ rather than $y_i^{(S+1)} = (I - M_i)y_i + M_iY_i^{(S)}\beta_i$, or with $\Pi_i^{(S+1)} = (\Pi^{(S)})_i\beta_i + \gamma^*_i$ in place of $\Pi_i^{(S+1)} = C_i^* y_i + I_i^*(\Pi^{(S)})_i \beta_i$.

Proof:

(a) In nonparametric cases we save $(q + 1)T(M - g)$ every iteration and clearly the initial calculation $Z_i\gamma_i$ is preferable to M_i and $(I - M_i)y_i$.

(b) In parametric cases γ_i^* saves all the initial calculations of $C_i^*y_i$. #

Lemma 7:

(a) For endogenous IIV estimators, (11.16), is computationally efficient relative to EFP.

(b) Lemma 6 still holds for (11.16).

Proof:

(a) Discussion following (31) above, $q \geq 1$.

(b) The saving per iteration is now $T \sum_{g+1}^{M} r_i$ #

Comment:

Other possible forms of (11.16) or (11.19) e.g.

$$y_i^{(S+1)} = (I - (M_iY_i^{(S)}(Y_i^{(S)'}M_iY_i^{(S)})^{-1}Y_i^{(S)'} - I)M_i)y_i$$

or $$y_i^{(S+1)} = (M_iY_i^{(S)}(Y_i^{(S)'}M_iY_i^{(S)})^{-1} - I)M_iy_i + y_i$$

are ruled out by the comment preceding Lemma 5 or in IIV estimators by Lemma 7, when the comparison is made with (11.19).

Lemma 8:

EIIV and EPIIV are never performed using $\hat{\beta}_i^{(S+1)}$ $i = 1 \ldots g$ from IV estimation in the style of (30).

Proof:

(i) EIIV

(a) Getting $\hat{\beta}_i^{(S+1)}$ by IV estimation, discarding $\hat{\gamma}_i^{(S+1)}$ is preferred only if

$$\sum_g (r_i+s_i)^2(q/4(r_i+s_i+5)+1)+T(q+1)\sum_g r_i(r_i+s_i+1)<(q+1)((T+1)\sum_g r_i^2+T\sum_g r_i)+q/4(\sum_g r_i^2(r_i+1))$$

$$\text{i.e. } (q+1)T\sum_g r_i s_i+q/4[\sum_g(2r_i s_i+s_i^2)r_i+\sum_g s_i(r_i+s_i)^2]+5/4 q\sum_g(2r_i s_i+s_i^2)+\sum_g s_i^2+2\sum_g r_i s_i<0.$$

which is impossible. (33)

(b) Given $\hat{\beta}_i^{(S-1)}$ $i = 1 \ldots g$ obtained by IV in the style of (30) then IIV is preferred in any event since comparing the remaining operations, IIV is preferable if $(q+1)T(Tg - \sum_g s_i) \geq 0$ by (11.35), i.e. always. (We need only compare behavioural relations by Lemma 6.)

(ii) EPIIV. Part (i)(a) applies (b) PIIV and EPIIV have the same computation per iteration. Hence the result holds for EPIIV. #

We may now compare methods within the parametric and nonparametric iteration groups.

Theorem 5:

Among nonparametric estimators EIIV is most preferred if $4(q+1)\sum_g s_i \geq g$, or iff

$$T \geq \frac{-\{\sum_g r_i(2r_i+s_i)[3/4 q(r_i+s_i+1)+q+2]+q/2\sum_g r_i^2 s_i\}}{(q+1)\{\sum_g r_i(r_i-1)+\sum_g s_i(4r_i+2s_i+1)\}-\sum_g r_i} \quad (34)$$

Proof:

By Lemmas 4 to 8 we need only compare FP with EIIV. By operation counts EIIV is preferred iff

$$(q+1)\left\{\sum_g (r_i+s_i)[T(2r_i+2s_i+1)+r_i+s_i]\right\} + q/4\sum_g (r_i+s_i)^2(r_i+s_i+1)$$
$$- (q+1)\sum_g r_i(2T+(T_i+1)r_i) - q/2\Sigma r_i^2(r_i+1) = T\sum_g r_i \gtrsim 0$$

which when solved for T yields (34). The numerator of (34) is clearly negative for all models. The donominator is nonegative iff

$$(q+1)\sum_g\left\{(r_i+s_i)^2+s_i(s_i+2r_i+1)\right\} \gtrsim (q+1)\sum_g r_i + \sum_g r_i$$

$$\text{i.e. } (q+1)[\sum_g r_i(r_i+4s_i-1) + \sum_g s_i(2s_i+1)] - \sum_g r_i \gtrsim 0 \qquad (34a)$$

Now using $(k+1)g - \sum_g s_i \gtrsim \sum_g r_i$ from the identification of the first g equations, and the last inequality in the proof of lemma 4 we have

$\left(\frac{\sum_g r_i s_i}{\sum_g s_i} + 1\right)g - \sum_g s_i \gtrsim \sum_g r_i$. The denominator is now nonnegative if

$$(q+1)[\sum_g r_i(r_i+4s_i-1)+2\sum_g s_i^2+\sum_g s_i] - \left(\frac{\Sigma r_i s_i}{\Sigma s_i} + 1\right)g + \Sigma s_i \gtrsim 0$$

which is certainly true if $4(q+1)\sum_g s_i \gtrsim g$ (a sufficient condition). #

Comment:

If the sufficient condition fails then (34a) should be tested for positiveness before (34) itself is used. A simpler form of the sufficient condition (but less accurate) is $K \gtrsim \frac{g}{4(q+1)} \gtrsim g/8$. It appears almost certain that therefore either sufficient condition will be satisfied for any model; especially larger models with a sparse β matrix and significant overidentification. Both are satisfied by all nine models considered here. It is possible to construct models for which (34) does not holds, but (11.36) does, but they seem extraordinarily unrealistic e.g. a cyclic β matrix and one nonzero s_i implies $T < 16$ and more than 14 equations; two nonzero s_i implies $T < 32$ and twice as many equations and so on. There seems no doubt that any plausible econometric model will satisfy (34).

Theorem 6:

Among parametric estimators EPIIV is always preferred.

Proof:

Lemma 4 restricts the options to PIIV and EPIIV; EPIIV is preferred if

$$\sum_g (r_i+s_i)^2(\tfrac{q}{4}(r_i+s_i+5)+1)-\tfrac{q}{4}\sum_g r_i^2(r_i+1)+(q+1)[K\sum_g r_i(s_i+K)-\sum_g r_i^2] \geq 0 \qquad (35)$$

which clearly always holds (Lemma 4(b) proof). #

Theorem 7:

Parametric estimators if available are preferred if T > K.

Proof:

(a) Mixed Iterations; using $Y^{(S)} = Z\Pi^{(S)}$ makes the following difference to estimation by the IV version of (30) ($Z\Pi^{(S)}$ operations less those of $Y^{(S)}$)

$$(q+1)\{(K-T)\sum_g (r_i(2+r_i+s_i)\} \qquad (36)$$

so including the operations in instrument formation PIIV is preferable if (we compare FP and PIIV by Lemma 4 (a)).

$$(T-K)\{(q+1)\sum_M r_i+M+(q+1)\sum_g r_i(s_i+r_i+1+(T+K)r_i)\}+(q+1)T\sum_M s_i \geq 0 \qquad (37)$$

which provides the result. In fact (36) holds if T < K by a very small amount due to $(q+1)T\sum_M s_i$ since $\hat{\gamma}_i^{*(S+1)}$ is not multiplied. We neglect this hereafter.

(b) Endogeneous estimators: (EIIV vs. EPIIV by Lemmas 4(b) and 8). EPIIV is preferred if

$$(T-K)[(M-g)(q+1)\sum_{g+1}^{M} r_i+(q+1)\sum_g r_i(r_i+2)+(q+1)T_g] \geq 0 \qquad (38)$$

which holds exactly (a necessary and sufficient condition).

(c) We check that parametric iteration with $Y^{(S)}$ used in the estimation phase is never preferred. Clearly such a PIIV or EPIIV is only preferable to conventional versions if $T < K$ by (36). #

13.4.2 RF instrument IIV estimaotrs (29):

Lemmas 5 - 8 clearly hold for these estimators also; Lemmas 5 and 8 because they are unaffected by methods of instrument formation; Lemma 6 because it is true for one instrument iteration; Lemma 7 because the instrument iteration from (11.16)

$$Y_i^{(S+1)} = Y_i^{(S+1)} \hat{\beta}_i^{(S+1)} + (I - M_i)(y_i + Y_i \hat{\beta}_i^{(S+1)}) \quad (39)$$

involves less multiplication than (11.33).

Theorem 8:

Parametric estimators, when available, are preferable if $T > K$.

Proof:

(a) Mixed iterations; the estimates are the same as for direct instruments, and by (36) parametric estimates are preferable iff $T > K$. One of the S_2 instrument iterations at step S is preferable by parametric methods iff

$$(q+1)[(T-K)\underset{M}{\Sigma} r_i + T\underset{M}{\Sigma} s_i] + (T - K)M \geq 0 \quad (40)$$

which holds if $T > K$. As in Theorem 7, (40) also holds for some $T < K$ but as $S_2 \to \infty$ the necessary and sufficient condition is $T > T^*$ where T^* provides equality in (40).

(b) Endogeneous iterations: because (38) holds iff $T > K$, the same condition for each instrument iteration to be quicker by parametric methods is $(q+1)(T-K)(\underset{M}{\Sigma} r_i + M)$ using (39) above.

(c) Part (c) of Theorem 7 holds by (36). #

Theorem 9:

(a) Theorem 5 reads RFEIIV is always preferred here.

(b) Theorem 6 holds unaltered for RF instruments.

Proof:

By lemma 6 we look at behavioural relations only at the estimation phase.

(a) Rework Theorem 5 as before and the condition for preferring RFEIIV is given by

$$T \gtrsim \frac{-\{\sum_g s_i(2r_i+s_i)[\frac{3}{4}q(r_i+s_i+1)+q+2] + q/2 \sum_g r_i^2 s_i\}}{(q+1)\{\sum_g r_i(r_i-1) + \sum_g s_i(4r_i+2s_i+1)\} + s_2(q+1)\sum_M s_i} \qquad (41)$$

which clearly holds since $T \gtrsim 0$.

(b) (35) is unaltered if RF instruments are obtained by iteration in each case. #

Comment:

(i) We note this is a neater result than Theorem 5.

(ii) Theorem 9 also holds for QRF instruments.

13.4.3 RF instruments and FP estimators (28):

Once again Lemmas 5, 6 and 8 apply and for the same reasons as in Section 4.2[1]. Therefore we can apply Lemma 4 in comparing corresponding estimators in the groups defined by (28) and (29), their iteration, estimation and instrument, phases being identical with the exception of the factors $\frac{\log \tau_1}{\log \lambda^*}$ and a. However in theory $0 = a \lesssim \frac{\log \tau_1}{\log \lambda^*}$ and in practice, none of the five reasons given for $a \neq 0$ would affect members of the FP family differently from those of IIV family i.e. were $\lambda^* = 0$ or $\lambda = 0$ to be the case, even then a would represent the lower bound of actual estimation iterations. Therefore we conclude (with Lemma 4(a), (b) and (28) and (29) and Lemma 12.1 as proof).

---

1. Note (11.16) is not available without changing the character of EFP, (similarly to the footnote to Lemma 4 (a)).

Theorem 10:

The most preferred method of the group defined by (29) is always locally preferred to the best method of the group defined by (28) (except when $\lambda^* = 0$ or $\lambda = 0$ when all methods in these two groups are equally preferred). #

The reader should be aware of the imprecision of Theorem 10 in that the differences in the estimator phase may influence S_2 and f_5 in such a way as to invalidate it since a different sequence $\beta^{(j)}$ $j = 1 \ldots S_1$ will be encountered. However he is reminded that (27) - (29) were only locally true - the same holds for Theorem 10. Additionally it is likely differences in $\frac{\log \tau_2}{\log \lambda^*}$ and a will be large and will more than compensate with a wide interpretation of "locally".

13.5 Comparing (27) and (29) type estimators:

We now test for the overall most desirable estimator in given circumstances. It is none of group (28) by Theorem 10. From Sections 4.2 and 4.3 we can see the RF instrument estimator, strictly corresponding to the most preferred direct instrument estimator, is also most preferred. Thus (29), IIV estimators with RF instruments, are most preferred if

$$f_6 f_7 S_1 (f_1 + f_2) \gtrsim a(f_1 + f_2 + f_4 S_2 f_5)$$

$$\text{i.e.} \quad \text{if } a \lesssim \frac{f_6 f_7 S_1 (f_1 + f_2)}{f_1 + f_2 + f_4 S_2 f_5} \tag{42}$$

where[1] $S_1 = \frac{\log \tau_1}{\log \lambda}$ and $S_2 = \frac{\log \tau_2}{\log \lambda}$. Thus $\frac{S_1}{S_2} = \frac{\log \tau_1}{\log \tau_2}$. We write $a = \frac{\log \tau_1}{\log \lambda_a}$

---

1. $f_1 = (T(T+1)\Sigma r_i + \Sigma r_i^2)(q+1) + \frac{q}{4}\Sigma r_i^2 (r_i+1)$ and $f_2 = f_4 = T(q+1)\Sigma r_i + TM$ (or K for T in parametric versions) Hence the bound on a and λ_a rises with T, overidentification and with sparseness of β; the typical situation for an econometric model especially the large ones.

as it would be unfair to compare estimators with differing tolerances on the final estimates. Hence we can derive bounds on λ_a (or a) below which RF instruments would be preferable. These bounds may be assessed for their plausability or likelihood of occurrence and compared with the kind of figures observed in practice. In fact (42) when written out is unenlightening (it has not proved possible to derive rules of thumb under which RF instruments are generally preferable) so upper bounds on a and λ_a have been calculated for each of the illustrative models for comparison with the actual observed number of iterations for RFIIV. If no instrument iteration acceleration is used $f_5 = 1$, if none for estimate iterations $f_6 f_7 = 1$, and it is argued that if one takes the trouble to calculate acceleration parameters or reorderings of equations for estimation iterations, then the same would be used for instrument iterations since the iteration matrix is locally (approximately) the same ($f_6 f_7 = f_5$). However instrument accelerations might be used without them being used in estimation iterations ($f_5 < 1$, $f_6 f_7 = 1$) for example in RFEIIV, RFEPIIV ($\alpha^* = 1$). Hence tables are presented with estimation accelerations, $f_6 f_7 = f_5 < 1$ or instruments only accelerated $f_5 < 1$, $f_6 f_7 = 1$. Separate studies of semi-iterative acceleration are not undertaken.

Tables follow in Appendix 1 illustrating the comparison of (27) and (29) in the form of (42) under Approach I, where the effect of several kinds of accelerations are investigated as well as the sensitivity to q, τ_1, τ_2, T, slightly differing β matrices (obtained by using differing estimates as the "true" β), differing error structures, slightly differing specifications and various orderings of the equations.

Three models are nonconvergent under (27) or (29) Michigan, Girschick-Haavelmo and Wharton. The first two may easily be made convergent under (29) by accelerating the instrument formation, and using also fractional and recursive accelerations for (27). Wharton however can only be made convergent by

superimposing SI or GRF accelerations at either stage. In order to provide the most comprehensive set of results, we continue, as if whatever accelerations are necessary are applied to either estimator or instrument accelerations to make $\lambda = .99$. Therefore in dealing with unaccelerated direct instrument estimators the symbol * implies an RF instrument estimator must be used (and any figures are for illustration only, indicating the comparison if λ were .99). In other cases ** implies the estimator is unavailable.[1] Table A.2 is derived from the best reordering of equations by the six algorithms presented by Bodin (results given in Bodin (1974) and Bergstrom (1974)). Any of these algorithms may be applied before the estimation and are not therefore optimal in the sense of Theorem 8.10. They are however practical and easy to apply and involve very little extra computation. Tables A.3 and A.4 enable assessment of the plausability that the bounds on a (and λ_a) are satisfied.

Of the models presented here, the following, having $T > K$, should use parametric iteration; Summers, Klein I, Christ, Klein-Goldberger, Hendry and Girschick-Haavelmo. Meanwhile Dutta-Su, Michigan and Wharton should use endogenous iteration and since (34) holds EIIV is recommended. Endogenous iteration, being often available when parametric is not (chapter 12.6 and 15), is of interest for all models.

Tables A.5(a) and A5(b) show upper bounds below which (locally) RF instruments are more efficient. These bounds, being locally valid, tend to be minimum upper bounds and therefore perhaps exceeding the bounds by a small quantity would still leave RF instruments preferable. It is fairly clear from both tables that the choice of instrument formation is very insensitive to the value of q. It is also clear that increasing T makes RF instruments preferable (compare Klein I, Christ, Girschick-Haavelmo; for specific results

---

1. $\lambda = .99$ is used only in comparisons with unaccelerated estimators

see Table A.15 for similar models).Likewise increasing overidentification and sparseness of β increases the advantage of RF instruments (compare Klein-Goldbeger Dutta-Su and Wharton) i.e. for the typical econometric model. A low proporation of identities has the same effect (contrast Summers,Girschick Haavelmo with Klein I, or Michigan (also with low T), Dutta-Su(lowT)and Wharton).It appears the advantage of RF instruments is generally more pronounced for the inappropriate iteration mode (i.e. when T is relatively large for a given model, or with increasing T - K). However a remarkable feature is the steady nature of the upper bound on a, and it remains to be shown that this holds under various acceleration, changes in the model and tolerances, etc. It is more difficult to decide if RF instruments should be used - on the very imcomplete information of Table 4. Observations of λ_a suggest they should be used; observations on S_1 (halved following the changed tolerance) suggest they should only be used for smaller models (say $M \leq 25$). Further there is a strong suggestion that S_1 departs from the theoretical value of zero by more with increasing model size and over-identification (but not T) which reinforces this rule of thumb. It must be emphasised that the really distinctive feature is that in all cases the bounds on a and λ_a are almost identical with their observed values, or what we might expect for them given the reasons that neither will be zero and both tend to increase with model size. Thus the decision on the instrument form appears extremely unimportant

Tables A6 and A7 may be used to see if these general and tentative conclusions hold when the tolerances are altered. The bounds on λ_a are about halved but those on S_1 virtually identical in every case (as expected). The insensitivity of q remains. The effect of overidentification, sparse β, and proportion of identities therefore also remain. The upper bounds also show the same pattern.

Since the decision on the instrument formation method appears so unimportant on computation criteria and since we consider only IIV family members here, it is clear that RF instruments and specifically RFEIIV or RFPEIIV will be preferred. This is for reasons of statistical properties (theorem 12.3, corollar 12.1), error control, as well as preferable, (a priori) global convergence conditions (chapter 12), independence of arbitrary tolerances and starts conveniently, now, we refer to theorem 9 rather than the clumsy theorem 5 (which perhaps should have accounted for initial conditions anyway).

The choice of tolerance has received no attention in the literature. The convergence of any estimator may be viewed as the convergence of the sequence $Y^{(S)}$, S = 1 ..... (12.22(b) and Theorem 12.1). There is therefore little point in setting tolerances very much less than, say, $^{1}/_{10}$th of the standard error of the regression as a percentage of the total variance of Y^{1}. This suggests a fairly weak tolerance and many might argue too weak for precision at the fixed point or for protection against false convergence, apparent convergence, convergence on the "wrong" solution or oscillation between two solutions(see Chap.14). Hence stricter tolerances are usually used and hereafter we comply with convention and use 10^{-6}.

The advantage of RF instruments increases as λ increases (becoming dominant if $|\lambda| > 1$), as well as with the model's structure described earlier, because S_1 and S_2 rise but a does not. This is illustrated in Tables A.7,A.8 and A9 where accelerations of instrument iterations are studied. In all cases, with the very notable exception of Klein I, with no reordering RF instruments would be preferable usually by a large margin. The accelerations of Tables A.7

---

1. Strictly $\tau_1 \leqslant .1 \dfrac{\hat{U}_j^{(S)\prime}\hat{U}_j^{(S)}(T-r_j)}{Y_j'Y_j(r_j)}$ where $j = \max_i \dfrac{\hat{U}_i^{(S)\prime}\hat{U}_i^{(S)}}{Y_i'Y_i}\dfrac{(T-r_i)}{(r_i)}$

   i = 1 .... g. Interest in a particular equation could turn this max into min.

and A9 are difficult to undertake a priori as they require the calculation of the eigenvalues of I - β. Those models marked "I" are p-cyclic so optimal SOR is no more difficult than optimal JOR; those marked "II" are not so it is very much more difficult. It is emphasised that accelerations of Table A.8 and the best reordering are both possible a priori and represent extremely effective methods of acceleration - likely to be better than optimal JOR, SOR and nearly as good as optimal SOR with reordering, by Table A.3.

It is interesting to compare RF and direct instrument estimators under the same accelerations, i.e. $f_5 = f_6 f_7$. The results are in Tables A.10-14. Three conclusions are clear. Our previous observations on the effects of q,T (see also Table A15),overidentification, sparse β become less pronounced. Secondly the upper bounds are very similar to unaccelerated cases, but uniformly very slightly reduced. Thirdly in no case would RF instruments seem preferable (but for small models the decision is unimportant). The uniformity of the upper bounds is quite remarkable and suggests these conclusions hold for models in general and the upper bound on a is always about 8, whether accelerated estimators are used or not. In addition, re-estimation with an increased sample size or small changes in specification or other small perturbations in β does not change these conclusions (Tables A.15 to A.18).

In summary RF instruments are recommended for any iterative estimators. Computationally they are slightly better placed in unaccelerated versions; and while it is possible they could be preferable , the differences are rather insignificant and are often eliminated when accelerations are introduced. RF instruments are inevitably superior when instrument iterations are accelerated, and the same accelerations can be applied to estimation iterations at no extra cost. The prime reasons for their recommendation where those arrived at in chapter 12.

13.5.1

RF instruments have been obtained by iteration in the above analysis. Is it possible they are better obtained by inversion

i.e. $Y^{(S+1)} = Z\hat{\Gamma}^{(S+1)}(I - \hat{\beta}^{(S+1)})^{-1}$ as suggested by all other authors[1].

An operation count indicates iteration is preferable if

$$T \geq \frac{MS_2 + qM(2S_2 - {}^{M}/_{2}(M+1)) - (q+1)KM^2}{(q+1)KM - S_2(M + (q+1)\sum_{M}(s_i + r_i))} \qquad (43)$$

The bound in (43) is positive, increases with model size, overidentification, and S_2 but decreases with K and q. For our models, $\tau_2 = 10^{-6}$, q = 1 the bound is

Table 3

Klein-Goldberger	1.17
Dutta-Su	1.27
Michigan	.114
Hendry	.92
Wharton	3.62

For the remainder (43) is negative. Therefore iterating instruments are without doubt preferable.

13.5.2

FI and AR estimators:

---

1. Lyttkens (1973), Bodin (1974), Bergstrom (1974), Brundy and Jorgenson (1974).

Lemma 9:

(a) In mixed or endogeneous iteration $\phi^{(S)^{-1}}$ is best calculated with typical element of $\varepsilon^{(S)}$, $\varepsilon_{ij}^{(S)} = T^{-1}\hat{\beta}_i^{(S)'}V_i^{(S)'}V_j^{(S)}\beta_j^{(S)}$ or $\varepsilon_i^{(S)} = T^{-1}V_i^{(S)}\hat{\beta}_i^{(S)}\hat{\beta}_i^{(S)'}V_i^{(S)'}$ for FI and AR transforms respectively.

(b) in parametric iteration $\phi^{(S)^{-1}}$ is best calculated with $\varepsilon^{(S)} = T^{-1}(Y\hat{\beta}^{(S+1)}+Z\hat{\Gamma}^{(S+1)})'\ (Y\hat{\beta}^{(S+1)} + Z\hat{\Gamma}^{(S+1)})$ or $\varepsilon_i^{(S)} = T^{-1}(y_i - Y_i\hat{\beta}_i^{(S+1)} - Z_i\hat{\gamma}_i^{(S+1)})\ (y_i - Y_i\hat{\beta}_i^{(S+1)} - Z_i\hat{\gamma}_i^{(S+1)})'$ respectively.

Proof:

(a) $\hat{u}_i^{(S)'}\hat{u}_j^{(S)} = \hat{\beta}_i^{(S)*'}\hat{V}^{(S)'}\hat{V}^{(S)}\hat{\beta}_j^{(S)*} = \hat{\beta}_i^{(S)'}\hat{V}_i^{(S)'}\hat{V}_j^{(S)}\hat{\beta}_j^{(S)}$

where β_i^* is the i^{th} column of β. This is clearly superior to (11.26) and no other options exist for endogenous iteration. In mixed iteration part (a) holds if $(g+K)g > \sum_g^{} r_i$, and that holds since by identification of the g behavioural equations $\sum_g^{} r_i \leq (K + 1)g - \sum_g^{} s_i$ and since $g \geq 1$ i.e. $[g(g-1) + \sum_g^{} s_i \geq 0]$.

(b) This is the only option since calculating $Y^{(S)}=Z\hat{\Pi}^{(S)}$ is never efficient (Lemma 8). #

13.5.3 QRF instruments:

Theorem 11:

In the notation of (42) theoretically $a = \log \tau_1/\log \tau_2$ in IIV estimators for τ_2 constant all S.

Proof:

By Theorem 12.22 the iteration matrix of IIV estimators with RF instruments is $(I - \beta)^{k_s}$; hence $\lambda_a = \lambda^{k_s}$. But $k_s = \log \tau_2/\log \lambda$. $\therefore$ $\lambda_a = \tau_2$ or $a = \log \tau_1/\log \tau_2$. #

Since we actually iterated the instruments in RF instrument estimators in all comparisons, Tables A5-A18 should be compared with the theoretical values $\lambda_a = 10^{-6}$ and $a = 1$. We argued that for several reasons there was a lower limit on a in practice and therefore one too for the effective λ_a. From Table 4, we guess $a \geq 10$ approximately or rather higher for large models. But in view of the theoretical results, that is Theorems 12.15, 12.22 and 11 above, there is the possibility that if τ_2 is allowed to rise, but remains fairly small then the a observed in practice will change very little but the work done in instrument formation will drop dramatically. The possibility is investigated in the following table A.19.

QRF instruments can definitely be promoted as preferable to RF instruments. It is tempting to look for an optimal k_s or τ_2 to minimise the total work done but the theory (Theorem 12.15 or 11 above) clearly indicates $\tau_2 = 0$ or $k_s = \infty$ is optimal. It will not be possible to optimise since it is not possible to obtain any expression for a in practice. It is also not possible to tell where the total work done for each model is minimised from the evidence available in comparisons such as Table A19 (although the advantage of QRF instruments advances with decreasing τ_2) or without a firm prediction of a in practice.

The conclusion is, therefore, QRF instruments should be used but with a very weak tolerance, say $\tau_2 = 10^{-1}$.

13.5.4 Some observations on Klein I:

The results above are of some interest in the light of the computational difficulties reported by Brundy and Jorgenson(1974)in estimating Klein I by FIRFIV using inversion for instrument formation. It is possible that estimation iterations were prolonged because $\phi^{(S)}$ was used with $Y^{(S+1)}$ as instruments. As a full information method, FIRFIV would be speeded up by calculating $\phi^{(S)}$ as in Lemma 9. There seems to be no short cut to the expensive inversion except by the approximate estimator which uses the same $\phi^{(S)^{-1}}$ for a set number of steps.

However a major computational saving is possible by iterating instruments accelerated (Table A.19),or by using QRF instruments. In fact inversion implies $\tau_2 < \tau_1$ which by Table A.19 is the most disadvantageous arrangement. In addition inversion is only possible if β is nonsingular. For Klein I estimated by RFIIV the smallest root is 2.5×10^{-4}. It must have been very expensive to achieve a satisfactory inversion[1]. A much increased within iteration burden results in no decrease in a, which will be on its effective lower bound. However Klein I is distinguished throughout Tables A5-A14 as the model for which RF instruments are least preferable. Evidently the greatest computational saving comes from using direct instrument estimators. The disadvantage of RF instruments is less pronounced for the other models studied here so it would be unfair to generalise from Klein I. However it is important to compare the best of the iterative estimators with conventional methods.

13.6 Comparison of Iterative Estimators with LIML

To complete our analysis of the problem of estimator selection, we compare the most favourable iterative estimator with LIML in the case of each model. If we compare the most favourable iterative estimator with each of the conventional estimators we have a complete preference ordering for estimator selection. Comparisons with 2SLS or 3SLS are superfluous, as are all comparisons with estimators with prespecified or incomplete sequence of iterations which represent a different kind of estimator. They are dependent on their starts - indeed the starts may influence their statistical properties (Theorem 12.3). Since we are now dealing with IIV estimators, 2SLS, 3SLS or FP estimators are only asymptotically equivalent

---

1. Inversion would be impossible for all the other models except Summers; smallest roots were Christ, Michigan, Hendry 10^{-17}; Klein-Goldberger 10^{-6}; Wharton 10^{-13}; Dutta-Su, Girschick-Haavelmo zero.

to LIML, or IIV. Finally even such IIV estimators could not be claimed to have achieved identity with maximum likelihood. Comparisons would have to be conducted on the reduced number of iterations and would therefore reflect too much the special interests of the user to be of value. Any of these reasons permits discrimination which will supersede computational considerations.

In earlier sections we could ignore the computational burden of starts by comparing only the iterated calculations or (preferably) comparing estimators with identical starts. This is no longer the case, so we simplify matters by taking $Y^{(0)}$ requiring no calculation. Any other start will, of course, reduce the advantage of iterative estimators.

LIML is taken as performed according to the scheme presented at (2.30). λ_i is obtained by QR transformation of $W_{ii}^{*}(W_{ii})^{-1}$. It could be faster to use for example von Mises' theorem and the power method but it is well known to be unreliable. To obtain the eigenvector, it is possible to use the inverse iteration of Householder. Although the number of iterations will be quite small their number cannot be assessed a priori. The order of the matrix $(W_{ii}^{*})^{-1} W_{ii}$ will be so small (2 or 3 typically) that, here, the method of Peters and Wilkinson (1970) will probably be better computationally. This implies

$$(3q/2 + 1) \sum_g r_i^2 + (q/2 + 2) \sum_g r_i - 1$$

operations to obtain the eigenvectors.

Putting the LIML calculations together we find EIIV is preferable if

$$S_1 \leq [(q+1)T\{K(2K+T)+M(M+T)+\sum_g r_i^2+T\sum_g r_i-2(\sum_g r_i+\sum_{g+1}^{M} s_i+T_g)\}+\tfrac{q}{4}[K^2(K+1)+2\sum_g r_i$$

$$+4\sum_g r_i^3+18\sum_g r_i^2]+\sum_g r_i^3+3\sum_g r_i^2+(11q+S)\sum_g r_i-T(g-1)\}/\{(q+1)(T(T+1)\sum_g r_i+T\sum_M r_i+\sum_g r_i^2)$$

$$+\tfrac{q}{4}[\sum_g r_i^2(r_i+1)]+(q+1)T\sum_M r_i+TM\} + (3q/2+1)\sum_g r_i^2 + (q/2+2)\sum_g r_i - 1 \qquad (45)$$

and EPIIV is preferable if

$$S_1 \leq \{(q+1)T[K(2K+T)+M(M+T)+\Sigma_g r_i^2+T\Sigma_g r_i^2+T\Sigma_g r_i-(T+K)(g+\Sigma_g r_i)-KM_g-T\sum_{g+1}^{M} s_i]$$

$$+\tfrac{q}{4}[K^2(K+1)+2\Sigma_g r_i+18\Sigma_g r_i^2+4\Sigma_g r_i^3]+\Sigma_g r_i^3+3\Sigma_g r_i^2+(11q+5)\Sigma_g r_i\} /$$

$$\{(q+1)[K(K+1)\Sigma_g r_i+\Sigma_g r_i^2]+\tfrac{q}{4}[\Sigma_g r_i^2(r_i+1)]+(q+1)K\Sigma_M r_i+KM\} \qquad (46)$$

$$+ (3q/2+1)\Sigma_g r_i^2 + (q/2+2)\Sigma_g r_i - 1$$

under Approach I in either case. (45) and (46) are presented in Tables A.20 and A21 respectively for the models under consideration.

It should be noted that since likelihood functions require T > K + M for their existence LIML can strictly only be applied in such cases. Obviously however physically it can be performed provided only T > K (to ensure the existence of W). Hence Dutta-Su, Michigan and Wharton may not be estimated by LIML in any event. To provide some comparison, T is set at K + M in all models where T < K + M. This arbitrary procedure is used for illustrative purposes only. In either (45) or (46) S_1 is very insensitive to q. The reader is reminded that only Dutta-Su, Michigan and Wharton would best use endogeneous iteration - the rest parametric. Tables A.20 and A21 should be compared with Table A3 on that basis. The results are rather confused. For Summers and Hendry LIML is always inferior. It is possible that this is the effect of no identities or large T - K value in their respective cases (see Tables A23,A24 below). It appears Klein and Christ are best done with recursive estimators if the equations are reordered according to Bodin's algorithms. Klein-Goldberger (and Dutta-Su with the increased sample size) are always better done by LIML whereas Michigan is better done by EPIIV with either of the two accelerations proposed.

It is difficult to generalise on the effects of different model characteristics on (45) or (46). They appear to rise with M, density of β, degree of overidentification, T or T - K, and fall with a rising proportion of identities. These generalisations are however rather unreliable and may only have a significant effect when several of the characteristics appear together.

We can add to these observations that any IIV estimator with RF instruments is theoretically superior to LIML; and that the correct policy for a priori selection of an estimator is to choose such an IIV estimator with RF instruments as the expected best. In practice from Table A.4 we see Klein, Christ and Girschick-Haavelmo are better estimated by such methods rather than LIML. To complete the picture the QRF instrument case is investigated in Table A.22. We expect once again the actual number of iterations will exceed the theoretical 6 but by a small margin but the results are little different from the direct instrument case. Hendry, Summers and Wharton are always better done by iterative methods; as are Michigan, Christ and Klein with suitable accelerations.

We note also that from Chapter 14 these observations can be approximately extended to FIML and full information iterative estimators, whenever FIML is used in its linear form, because it is no more than the full information version of (2.32). The approximation comes about because we do not know how much work is involved in iterating the covariance matrices.

13.6.2 The computation of FIML:

We have four basic types of FIML estimator, and several algorithms for each: (2.36), (2.42), (7.1) and (7.3(a)). As a computational (and notational) simplification, the general K-matrix class estimator is written

$$\hat{X}'\phi^{*-1}(E \otimes I - K \otimes Q)\hat{X}\hat{\delta} = \hat{X}'\phi^{*-1}(E \otimes I - K \otimes Q)y \qquad (47)$$

with ϕ^* from section 2.4, and E the unit matrix of order M.

13.6.2.1

Global convergence is guaranteed for (2.42) and (7.1) by theorems 12.15, 12.25 and 7.6. Local convergence is guaranteed for (2.36) by theorem 2.21. Finally for (47) and (48), taking the case $\phi^* = \Sigma \otimes I$ for simplicity:

Theorem 10:

(a) (47) is globally convergent when K is fixed.

(b) (47) is globally convergent if the condition at (50) below holds, when K varies with each iteration in (7.3a).

Proof:

(a) if K is fixed in (47): we require $\rho\left(\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}^{(s)}}\right) < 1$. For $j=1\ldots\sum_g(r_i+s_i)$

$$\hat{X}'(\Sigma^{(s)^{-1}} \otimes I)(E \otimes I - K \otimes Q)\hat{X}\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}^{(s)}_j} = \hat{X}'\left(\frac{\partial\Sigma^{(s)^{-1}}}{\partial\hat{\delta}^{(s)}_j}\right) \otimes I)(E \otimes I - K \otimes Q)\hat{u}^{(s+1)} \quad (48)$$

The matrix on the left of (48) is nonsingular and the vector on the right vanishes for the same reasons as that in Theorem 7.6

so $\left\{\frac{\partial\hat{\delta}^{(s+1)}}{\partial\hat{\delta}^{(s)}}\right\} = 0$, all $s > 0$ and convergence is expected to need the same number of iterations as (7.1).

(b) if K varies with every iteration in (47), then (48) is

$$\hat{X}'(\Sigma^{(s)^{-1}} \otimes I)(E \otimes I - K^{(s)} \otimes Q)X\frac{\partial \hat{\delta}^{(s+1)}}{\partial \hat{\delta}_j^{(s)}} = \hat{X}'\left[\frac{\partial \Sigma^{(s)^{-1}}}{\partial \hat{\delta}_j^{(s)}} \otimes I\right](E \otimes I - K^{(s)} \otimes Q)\hat{u}^{(s+1)}$$

$$+\hat{X}'(\Sigma^{(s)^{-1}} \otimes I)\left[-\frac{\partial K^{(s)}}{\partial \hat{\delta}_j^{(s)}} \otimes Q\right]\hat{u}^{(s+1)} \qquad (49)$$

The first term on the right of (49) vanishes as before; $\frac{\partial K^{(s)}}{\partial \hat{\delta}_j^{(s)}}$ has typical element $\frac{1}{\sigma^{(s)pq}}\left[\frac{\partial w^{(s)'pq}}{\partial \hat{\delta}_j^{(s)}} - k^{(s)pq}\frac{\partial \sigma^{(s)pq}}{\partial \hat{\delta}_j^{(s)}}\right]$ so $\left\{\frac{\partial \hat{\delta}^{(s+1)}}{\partial \hat{\delta}^{(s)}}\right\}$, whose spectral radius must be less than unity, has elements from

$$-[\hat{X}'(\Sigma^{(s)^{-1}} \otimes I)(E \otimes I - K^{(s)} \otimes Q)\hat{X}]^{-1}\hat{X}'(\Sigma^{(s)^{-1}} \otimes I)\left[\frac{\partial K^{(s)}}{\partial \hat{\delta}_j^{(s)}} \otimes Q\right]\hat{u}^{(s+1)} \qquad (50)$$

which cannot vanish in general. (Local convergence: s^{th} step values are replaced by true values.) #

13.6.2.2

Comparison of (7.1) and (7.3(a)) for computation burden (i.e. (7.3(a)) less (7.1) in operations) yields a difference given by

$$S_1\left\{M^3\,\frac{9}{4}q+2 + M^2\left(q(T^2+\frac{5}{4})+T\right) - MT(qT+1)\right\} - (\frac{7}{4}q+2)\Sigma r_i^{\,3} - [q(\frac{5}{4}+T)+T+2]\Sigma r_i^{\,2}$$

$$-[q(\frac{21}{2}+T^2)+T^2+4]\Sigma r_i - \frac{q}{4}\Sigma s_i^{\,2} - [q(2T+\frac{1}{4})+T]\Sigma s_i^{\,2} - (q+1)T^2\Sigma s_i - T + S_2\{\frac{q}{4}\left(\Sigma(r_i+s_i)\right)^3$$

$$+ [q(\frac{5}{4}+MT)+MT+1]\left(\Sigma(r_i+s_i)\right)^2 + (q+1)M^2T^2\Sigma(r_i+s_i) + M^3[q(\frac{7}{2}+T^3)+3+T^3]$$

$$+M^2[q(\frac{3}{2}+T^2+K)+T+K]\} \qquad (51)$$

where S_1 is the number of iterations required by (7.1) and S_2 the number in excess of that required by (7.3(a)) - Theorem 10 indicates that (7.3(a)) will actually require more iterations.[1] It is easily verified, using $r_i < T$ and $\Sigma r_i^n \leq (\Sigma r_i)^n$ all $n \geq 1$, that if $S_1 = S_2 = 1$ (the least favourable case possible for (7.1)) (51) is positive.

Theorem 11:

Therefore (7.1) is preferred to (7.3(a)). #

The disadvantages of (7.3(a)) are the increased computational burden, the lack of guaranteed convergence, and the nonlinear iteration involved means plural convergence points and therefore a search procedure over possible $K^{(0)}$ values is required, in addition, to select the global maximum.

The remaining comparisons are less firm in their conclusions. For algorithms maximising (2.36), the only result on the number of iterations required - the quadratic termination property - certainly does not apply to the degree of nonlinearity in (2.36). We can safely assume a greater number of iterations than that will be necessary, i.e. greater than the $n = \Sigma(r_i+s_i)+\frac{M}{2}(M+1)$ parameters to be estimated. Even with the savings in computation provided in Section 2.5 and superlinear convergence, there can be no doubt that (2.42) is preferred to (2.36)[2] since using RF instruments (or QRF instruments) a very small number of iterations will be used. This has been recognised to some extent in practical work e.g. Goldfeld and Quandt (1972) emphasise that (2.36) methods are cumbersome and unsuitable except for small models ($M \leq 20$ say). Further searches over all $Y^{(0)}$ still have to be undertaken.

Once again it is impossible to predict a priori how much extra computation and iteration is required by (7.1) and (2.42) over their limited information versions. Hence comparison is only possible by assuming it to be approximately equal for both. On this basis we prefer (2.42) whenever IIV was preferred to LIML in the previous discussion. As this was rather inconclusive, practical

---

1. (51) assumes an identical $\Sigma^{(0)}$ for both estimators.
2. The computational burdens may be rather similar for very small models.

considerations are emphasised such as if $\max_i(r_i,s_i) \lesssim T \lesssim K,M$ then IIV methods only are available. One the other hand the unique global convergence of (7.1) to the global maximum would definitely recommend (7.1) for all models having $T > K,M$.

13.7 OLS Computations and QR Transformation

It has recently been suggested by Belsley that OLS or IV computation is better done by replacing the data matrices by the QR transforms. He argues that procedure is computationally more efficient and numerically stable since inversion is avoided. The importance and desirability of a given degree of numerical stability is of course determined by the researcher, and the ease of obtaining it is a function of a particular data set. It is beyond the scope of this analysis to assess the efficiency of alternative inversion routines and analyse the rounding error problem in each case. But this whole chapter has been based on OLS and IV estimators performed traditionally (with inversions by the product form of the inverse), and since we were concerned from the outset with the computational efficiency, the use of traditional methods must be justified. Comparing operation counts of estimates by QR transform and those by OLS, we have QR transform is more efficient when

$$\frac{n^3q}{4} + n^2\left[T(q+1)+\frac{3q+2}{2}\right] - n\left(T(T+1)\left(\tfrac{q}{2}+1\right) + \frac{q+1}{2}\right) - T\left((T+1)\left(\tfrac{q}{2}+1\right)+10q+3\right)\gtrsim 0 \qquad (52)$$

where n = number of coefficients to be estimated in the particular equation. By inspection this expression is negative for $n = 1$, $T \gtrsim 1$, $q \gtrsim 0$ but quality in (52) implies a unique positive solution n^* for all $T,q > 0$. n^* was found for various T and q in Table A.25.

A good rule of thumb for computational efficiency is therefore to use QR transforms only if $n \geq T/2$ for $T \leq 100$. In iterative estimation when estimates are only obtained for structural coefficients even for small models, T will be large enough and r_i $i = 1 \ldots M$ small enough for QR transforms never to be computationally efficient. Additionally collinearity and other problems aggrevating numerical stability will be less severe in the structural form. Unrestricted reduced form estimation however can be difficult and $n > T/2$ usually holds. Numerical instability is at its worst. In 2SLS, 3SLS the QR versions given by Belsley may well be superior.

Further investigation of (52) confirms our rule of thumb. Denote (52), with $n = T/2$ to eliminate n, as L;then $\frac{\partial L}{\partial T} < 0$ all $T \geq 0$ if $q < 8$. However if $q \geq 8$, $\exists T$ such that $\frac{\partial L}{\partial T} > 0$ so $n^* < T/2$ exists but for $q = 10$ that n^* exceeds 50. For all practical purposes we may obey thus rule of thumb even when working in double precision. We note n^* is less sensitive to q than T and the smaller q or T the more inaccurate the rule of thumb. A more accurate rule appears to be $n^* \geq T(\frac{1}{2} + q/5)$.

APPENDIX 1

Table A.1

Factor improvements from various accelerations:-

Model	λ	f_6 opt. JOR	f_6 Gauss-Seidel	f_6 opt.SOR	Block cyclic β?
Summers	.6708	.2297	.519	.2172	2-cyclic
Klein I	.8122	.653	1.0	.3576	not p-cyclic
Christ	.9742	.0799	.4386	.05246	not p-cyclic
Klein-Goldberger	.9038	.175	.3269	.2066	not p-cyclic
Dutta-Su	.9744	.1697	.1626	.069	5-cyclic
Michigan	-1.0313	.0093	.0082	-	not p-cyclic
Hendry	.5123	.592	.746	-	not p-cyclic
Wharton	+7.43	**	**	-	-
Girschick-Haalvelmo	-1.368	.01922	**	.00922	3-cyclic

** not possible to perform due to $\lambda > 1$

Table A.2

Factor improvements in SOR with the best reordering

	$\alpha = 1$	$\alpha^*$for SOR
Klein I	.2176	.11297
Christ	.02316	.01517
Klein-Goldberger	.1214	.1214
Dutta-Su	.03287	.02512

Summers is invariant to reordering

Table A.3

S_1 or S_2 for τ_1 and/or $\tau_2 = 10^{-6}$. For 10^{-3} halve these figures.

	$f_6=1$	opt.JOR	Gauss-Seidel	opt.SOR	Gauss-Seidel best reordering	opt.SOR best reordering
Summers	34	8	18	8	8	8
Klein I	67	44	67	24	15	8
Christ	531	43	233	28	13	9
Klein-Goldberger	137	24	45	29	17	17
Dutta-Su	536	91	88	37	18	14
Michigan	(1364)	13	12	-	-	-
Hendry	21	13	16	-	-	-
Wharton	(1364)	**	**	-	-	-
Girschick-Haalvelmo	(1364)	27	**	13	-	-

(All figures rounded up to nearest integer)

Bracketed figures; after accelerations to make $\lambda = .99$

Table A.4

Observed a and λ_a under RFIIV taken from Bergstrom

	Observed a	Observed λ_a	
Klein I	7	.1585	$\tau_1 = 10^{-6}$
Christ	11	.218	It is impractical
Klein-Goldberger	12	-	to calculate λ_a
Dutta-Su	36	-	for large models
Girschick-Haalvelmo	5	0	

Table A.5(a)

Upper bounds for λ_a; below these EIIV is superior to RFEIIV. Bracketed figures are Upper bounds for a; below these EIIV (29) is superior to RFEIIV. Nonparametric versions, EIIV and RFEIIV, with $\tau_1 = \tau_2 = 10^{-3}$; the influence of q.

	q = 1	q = 2	q = 3	q = 4	q = 5	q = 10
Summers	.4213(8)	.4415(9)	.4524(9)	.459(9)	.4628(9)	.4762(10)
Klein I	.2878(6)	.3118(6)	.3248(7)	.331(7)	.3386(7)	.3517(7)
Christ	.442(9)	.4677(10)	.4819(10)	.4898(10)	.4927(10)	.5091(11)
Klein-Goldberger	.3956(8)	.4354(9)	.4249(9)	.4315(9)	.4354(9)	.4457(9)
Dutta-Su	.4507(9)	.4015(8)	.4141(8)	.4225(9)	.4278(9)	.451(9)
Michigan	.03795(3)*	.0423(3)*	.0451(3)*	.0468(3)*	.048(3)*	.045(3)*
Hendry	.3631(7)	.3741(8)	.3797(8)	.3832(8)	.3857(8)	.3911(8)
Wharton	.6655(17)*	.6814(18)*	.6893(19)*	.694(19)*	.6972(20)*	.705(20)*
Girschick-Haalvelmo	.6401(16)*	.6676(18)*	.681(18)*	.689(19)*	.6953(20)*	.7084(21)*

Bracketed figures rounded up to nearest integer; no accelerations used.

Table A.5(b)

As Table 6(a) but EPIIV and RFEPIIV

	q = 1	q = 2	q = 3	q = 4	q = 5	q = 10
Summers	.175(4)	.1983(5)	.2115(5)	.22(5)	.226(5)	.224(5)
Klein I	.1113(4)	.127(4)	.136(4)	.1419(4)	.146(4)	.156(4)
Christ	.2557(6)	.2793(6)	.2922(6)	.3004(6)	.3054(6)	.3193(7)
Klein-Goldberger	.3424(7)	.3622(7)	.373(8)	.3792(8)	.3836(8)	.394(8)
Dutta-Su	.4897(10)	.5129(11)	.5225(11)	.5333(11)	.5388(12)	.5504(12)
Michigan	.3116(6)*	.3316(7)*	.3424(7)*	.349(7)*	.3535(7)*	.3639(7)*
Hendry	.2374(5)	.252(6)	.2599(6)	.2647(6)	.268(6)	.276(6)
Wharton	.906(70)*	.911(75)*	.9137(77)*	.9155(79)*	.9162(80)*	.9187(82)*
Girschick-	.1325(4)*	.1553(4)*	.1687(4)*	.1835(5)*	.1835(5)*	.2687(6)*

Table A.6

Upper bounds for λ_a and a (bracketed) exactly as Tables 5(a) and 5(b) but $\tau_1 = \tau_2 = 10^{-6}$

Iteration type:	Endogenous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.263(11)	.04541(5)	.3346(13)	.074(6)
Klein I	.1121(7)	.01513(4)	.1510(8)	.02968(4)
Christ	.2004(9)	.06705(6)	.266(11)	.1047(7)
Klein- Goldberger	.1719(8)	.1295(7)	.2194(10)	.1716(8)
Dutta-Su	.2085(9)	.246(10)	.1995(9)	.3111(12)
Michigan	.0014(3)*	.098(6)*	.0026(3)*	.1338(7)*
Hendry	.2392(10)	.1024(7)	.2783(11)	.1385(7)
Wharton	.4428(17)*	.8208(71)*	.4969(20)*	.844(82)*
Girschick-Haalvelmo	.4127(16)*	.01744(4)*	.415(16)*	.07291(6)*

Table A.7(a)

Upper bounds for λ_a and a when instrument iterations in RFEIIV RFEPIIV are optimally accelerated by JOR, $\tau_1 = \tau_2 = 10^{-3}$.

Iteration:	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.5994(14)	.4899(10)	.6163(15)	.5233(11)
Klein I	.4131(8)	.2221(5)	.4797(10)	.2769(6)
Christ	.9147(78)	.8758(53)	.925(89)	.8913(60)
Klein-Goldberger	.7823(29)	.7633(26)	.7987(31)	.782(29)
Dutta-Su	.8551(45)	.9539(146)	.8519(44)	.8845(57)
Michigan	.9596(168)*	.9795(318)*	.9628(183)*	.9802(347)*
Hendry	.4246(9)	.3325(7)	.434(9)	.3612(7)
Wharton	**	**	**	**
Girschick-Haalvelmo	.9815(372)*	.952(142)*	.9833(415)*	.965(195)*

Table A.7(b)

As Table 7(a) but $\tau_1 = \tau_2 = 10^{-6}$

Iteration:	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.5328(22)	.356(14)	.5633(25)	.3983(16)
Klein I	.2085(9)	.0603(5)	.2812(11)	.09363(6)
Christ	.8586(91)	.7866(58)	.8478(107)	.8153(68)
Klein-Goldberger	.6761(36)	.6434(32)	.7057(40)	.6959(36)
Dutta-Su	.7501(49)	.772(54)	.7445(47)	.8028(63)
Michigan	.928(187)*	.9687(437)*	.9367(212)*	.9716(481)*
Hendry	.3278(13)	.1985(9)	.3585(14)	.2731(10)
Wharton	**	**	**	**
Girschick-Haalvelmo	.973(511)*	.9151(156)*	.977(596)*	.9410(227)*

Table A.8

Upper bounds for λ_a and a when instrument formation only is accelerated by Gauss-Seidel (Recursive). $\tau_1 = \tau_2 = 10^{-6}$

Iteration	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.4087(16)	.1644(8)	.4634(18)	.2116(9)
Klein I	.1121(7)	.01513(4)	.1510(8)	.02968(4)
Christ	.4865(20)	.3012(12)	.562(24)	.3664(14)
Klein-Goldberger	.5254(22)	.4789(19)	.569(25)	.525(22)
Dutta-Su	.7586(51)	.7791(56)	.7528(49)	.8095(66)
Michigan	.7387(218)*	.9714(476)*	.9428(235)*	.9738(521)*
Hendry	.2912(12)	.1546(8)	.3258(13)	.1935(9)
Wharton	**	**	**	**
Girschick-Haalvelmo	**	**	**	**

Table A.9

Upper bounds for λ_a and a with instrument acceleration, only, by optimal SOR and best reordering. $\tau_1 = \tau_2 = 10^{-6}$

Iteration		Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
I	Summers	.539(23)	.368(14)	.568(25)	.409(15)
	Girschick-Haalvelmo	.982(762)*	.954(293)*	.984(848)*	.9665(405)*
	Dutta-Su	.9374(215)	.941(229)	.9365(211)	.947(254)
II	Klein I	.634(31)	.519(22)	.634(31)	.519(22)
	Christ	.9506(277)	.936(208)	.9506(277)	.936(208)
	Klein-Goldberger	.6416(32)	.606(28)	.634(31)	.641(32)

Table A.10

Upper bounds for λ_a and a; optimal fractional estimators; $f_5 = f_6 f_7$ $\tau_1 = \tau_2 = 10^{-6}$

Iteration:	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.06457(6)	.01115(4)	.0822(6)	.01818(4)
Klein I	.0906(6)	.01355(4)	.1355(7)	.02659(4)
Christ	.1486(8)	.0497(5)	.1972(9)	.07756(6)
Klein-Goldberger	.1069(7)	.08045(6)	.1363(7)	.1066(7)
Dutta-Su	.1838(9)	.2169(10)	.176(8)	.2743(11)
Michigan	.0005(2)	.0334(4)	.00088(2)	.0459(5)
Hendry	.1521(8)	.0651(6)	.1768(8)	.088(6)
Wharton	**	**	**	**
Girschick-Haalvelmo	.2499(10)	.01067(3)	.3049(12)	.0438(5)

Note the absence of * on Michigan, Wharton or Girschick-Haalvelmo; and in subsequent tables.

Table A.11

Upper bounds for λ_a and a; Recursive estimators $\alpha = 1$; $\tau_1 = \tau_2 = 10^{-6}$

Iteration	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	.1784(9)	.0308(4)	.2272(10)	.051(5)
Klein I	.1121(7)	.01513(4)	.151(8)	.0297(4)
Christ	.1938(9)	.06486(6)	.2735(11)	.1012(7)
Klein-Goldberger	.1396(8)	.1052(7)	.1782(9)	.1394(8)
Dutta-Su	.1852(9)	.2149(9)	.1747(9)	.2724(11)
Michigan	.0004(2)	.0288(4)	.0008(2)	.0679(6)
Hendry	.1913(9)	.0819(6)	.2224(10)	.1107(7)
Wharton	**	**	**	**
Girschick-Haalvelmo	**	**	**	**

Table A.12

Upper bounds for λ_a and a; optimal recursive estimators, $\tau_1 = \tau_2 = 10^{-6}$

Iteration		Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
I	Summers	.05815(6)	.01005(4)	.074(6)	.0164(4)
	Girschick-Haalvelmo	.1398(8)	.0659(6)	.1702(8)	.0112(6)
	Dutta-Su	.1476(8)	.1738(8)	.141(8)	.2198(10)
II	Klein I	.0698(6)	.0104(4)	.1042(7)	.0205(4)
	Christ	.125(7)	.0419(5)	.1663(8)	.0654(6)
	Klein-Goldberger	.1167(7)	.0879(6)	.1489(8)	.1164(7)

Table A.13

Upper bounds for λ_a and a; best reordering; recursive estimators $\alpha = 1$; $\tau_1 = \tau_2 = 10^{-6}$.

Iteration	Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
Summers	0.5815(6)	0.1005(4)	.074(6)	.0164(4)
Klein I	.0482(5)	.0072(3)	.0719(6)	.0141(4)
Christ	.0669(6)	.0224(4)	.0887(6)	.0349(5)
Klein-Goldberger	0.828(6)	.0624(5)	.1057(7)	.0826(6)
Dutta-Su	.0977(6)	.1153(7)	.0935(6)	.1465(8)

Note: the reordering is of course in both estimation and instrument iterating.

Table A.14

Upper bounds for λ_a and a; best reordering; optimal recursive estimators; $\tau_1 = \tau_2 = 10^{-6}$.

Iteration		Endogeneous q = 1	Parametric q = 1	Endogeneous q = 10	Parametric q = 10
I	Summers	.05815(6)	.01005(4)	.074(6)	.0164(4)
	Dutta-Su	.0767(6)	.0905(6)	.0735(6)	.115(7)
II	Klein I	.02004(4)	.003(3)	.0299(4)	.0059(3)
	Christ	.037(5)	.0124(4)	.0492(5)	.0193(4)
	Klein-Goldberger	.0828(6)	.0624(5)	.1057(7)	.0826(6)

Table A.15

Upper bounds for λ_a and a; the effect of increasing sample size (Summer's Model) for q = 1 only.

	Corresponding to Tables								
	5(a)	6	7(a)	7(b)	8	9	10	11	12
T = 20	.4213(8)	.263(11)	.5944(14)	.5328(22)	.4087(16)	.539(23)	.06457(6)	.1784(9)	.05815(6)
T = 40	.522(11)	.4042(16)	.6298(15)	.5908(27)	.511(21)	.592(27)	.0992(6)	.2742(11)	.0892(6)
T = 60	.5645(13)	.4782(19)	not computed					.3281(13)	.1105(7)
T = 80	.593(13)	.529(22)						.3629(14)	.1224(7)

Table A.16

Upper bounds for λ_a and a; the effect of varying specifications (Christ's Model) for q = 1 only. Versions 17, 15, 8, 13; comparisons for parametric iteration.

Corresponding to Tables	5(b)	6
Version 17	.175(4)	.0671(6)
Version 15	.2307(5)	.053(5)
Version 8	.2495(5)	.06223(5)
Version 13	.1217(4)	.0148(4)

Version 17 is Christ's preferred version, with gross private investment expenditure a function of gross business product, gross private investment the previous year, stock market price at the start of the year and the year before that, and net private domestic capital at the start of the year. Version 15 is his second choice and lacks gross private investment of the previous year but gains property income as explanatory variables in the investment function. Versions 8 and 13 were chosen as they represent the extremes, worst and best, in RFIIV convergence (Bergstrom). Version 8 is as version 15 but without the true stock market price variables. Version 13 is as version 17 but without lagged investment expenditure.

Table A.19

Upper bounds for λ_a, a; QRF instruments with q = 1 and $\tau_1 = 10^{-6}$

Iteration	endogeneous $\tau_2=10^{-3}$	parametric $\tau_2=10^{-3}$	endogeneous $\tau_2=10^{-2}$	parametric $\tau_2=10^{-2}$	endogeneous $\tau_2=10^{-1}$	parametric $\tau_2=10^{-1}$
Summers	.4213(16)	.175(10)	.4844(20)	.2699(11)	.565(25)	.4214(16)
Klein I	.2878(12)	.1113(8)	.4062(16)	.2156(10)	.5749(25)	.4188(16)
Christ	.442 (18)	.2557(12)	.5751(25)	.399 (16)	.7489(48)	.6237(30)
Klein-Goldberger	.3956(16)	.3424(14)	.519 (22)	.473 (19)	.6855(37)	.6537(30)
Dutta-Su	.4507(18)	.4897(20)	.5833(26)	.6159(29)	.7539(49)	.6539(33)
Michigan	.03795(6)*	.3116(12)*	.1126(7)*	.4581(18)*	.3338(13)*	.7749(55)*
Hendry	.3631(14)	.2374(10)	.4003(16)	.3018(12)	.4553(18)	.3954(15)
Wharton	.6655(34)*	.906(140)*	.7596(51)*	.933(201)*	.867(97)*	.96(349)*
Girschick-Haavelmo	.6401(32)*	.1325(8)*	.731(46)*	.259(11)*	.8565(90)*	.5064(21)*

No accelerations of any kind; compare Table 6.

Table A.20

Endogenous iteration:

	q=1	2	3	4	5	10	Comparison figure
Summers	3.6	3.6	3.7	3.7	3.7	3.7	4
Klein I	4.8	4.9	4.9	5.0	5.0	5.0	5
Christ	7.2	7.3	7.4	7.4	7.4	7.5	8
Klein-Goldberger	5.5	5.6	5.6	5.6	5.6	5.7	6
+ Dutta-Su	3.2	3.3	3.3	3.3	3.3	3.3	4
+ Michigan	18.7	19.2	19.5	19.6	19.8	20.0	20
Hendry	6.2	6.2	6.2	6.3	6.3	6.3	7
+ Wharton	21.1	21.3	21.4	21.5	21.5	21.6	22
Girschick-Haalvelmo	1.6	1.6	1.7	1.7	1.7	1.7	2

The comparison figure is that used for comparisons with Table 3.

+ indicates this iteration would best be used.

Table A.21

Parametric iteration:

	q=1	2	3	4	5	10	Comparison Figure
+ Summers	35.5	36.2	36.5	36.8	37.0	37.3	38
+ Klein I	13.4	13.8	14.0	14.1	14.2	14.4	15
+ Christ	14.3	14.6	14.8	14.9	14.93	15.1	16
+ Klein-Goldberger	5.4	5.5	5.5	5.5	5.6	5.6	6
Dutta-Su	5.6	5.7	5.8	5.8	5.8	5.8	6
Michigan	13.7	14.0	14.2	14.3	14.3	14.5	15
+ Hendry	21.5	21.8	22.0	22.1	22.1	22.3	23
Wharton	20.9	21.2	21.4	21.5	21.5	21.6	22
+ Girschick-Haalvelmo	11.3	11.4	11.5	11.6	11.6	11.7	12

+ indicates this iteration would best be used.

Table A.22

Upper bounds for a, where EIIV or EPIIV with QRF instruments is superior to LIML:

		No acceler-ations	Optimal fract-ional	Recursive $\alpha = 1$	Optimal recur-sive	Reordered recursive $\alpha = 1$	Reordered optimal Recursive
Para-metric, LIML avail-able	Summers	17	28	23	28	23	28
	Klein	<6	<6	<6	8	8	10
	Christ	<6	6	<6	8	10	13
	Klein-Goldberger	<6	<6	<6	<6	<6	<6
	Hendry	16	18	17	-	-	-
	Girschick-Haalvelmo	<6	6	**	7	-	-
Endog-eneous, LIML with incre-ased sample	Dutta-Su	<6	<6	<6	<6	<6	<6
	Michigan	<6	16	16	-	-	-
	Wharton	15	**	**	**	-	-

$\tau_1 = 10^{-6}$, $\tau_2 = 10^{-1}$, theoretical a = 6 for EIIV and EPIIV.

Table A.23

Upper bound for a, where LIML is inferior, as a function of sample size, Summers model (EPIIV) and q = 1

T = 10	10.4	T = 30	68.7
T = 20	33.1	T = 40	117.2

Table A.24

S_1 under Approach III (compare Tables 3 and 20, 21).

	Jacobi	opt. JOR	SOR $\alpha = 1$
Summers:	identical to Table 3		
Klein	27(.5974)	14(.3696)	0 (10^{-6})
Christ	45(.7334)	28(.6023)	72(.8243)
Klein-Goldberger	34(.6637)	13(3.665)	** (30.6)
Dutta-Su	24(.5545)	12(.3095)	**(2.632)
Michigan	18(.4474)	12(.2882)	46(.7409)
Hendry	7(.1342)	7(.1342)	13(.333)
Wharton	**(7.495)	**	**(7962.2)
Girschick-Hallvelmo	identical to Table 3.		

Identities removed; $\tau_1 = 10^{-6}$; values rounded up to a whole number of iterations, (bracketed figures are spectral radii). Beside ** is the eigenvalue preventing convergence.

Notes:

1. Rather little appears to be gained by optimal JOR in these cases; rather frequently SOR is inferior to JOR.

2. All the iteration matrices will be full by (11) and hence none will be p-cyclic. No optimal SOR parameters were therefore calculated.

3. Iterative estimators preferable as follows:

Table A.25

n^* for $10 \leq T \leq 100$, $1 \leq q \leq 10$

	q=1	2	3	4	5	10
T = 10	9	8	8	7	7	7
20	15	14	13	13	12	12
30	22	20	19	18	17	16
40	29	26	24	23	23	21
50	36	32	30	29	28	26
60	43	38	35	34	33	31
70	50	44	41	39	38	36
80	57	50	47	45	43	41
90	64	56	52	50	49	46
100	70	62	58	56	54	50

Appendix 2

Below are the precise forms of each of the decision rules presented in Chapter 13, Sections 4 and 5. The symbol (A31), for example, indicates that expression is the precise form of the expression (31) given in the text. Throughout $S_1 = \frac{\log \tau_1}{\log \rho(J_1)}$; J_1 is the relevant iteration matrix of estimation iterations and $S_2 = \frac{\log \tau_2}{\log \rho(J_2)}$; $J_2 = I - \beta$, the iteration matrix of instrument iteration.

(A31) $$\Sigma_g(r_i+s_i)^2(\tfrac{q}{4}(r_i+s_i+5)+1)+T(q+1)\Sigma_g r_i(r_i+s_i+1)+\frac{(q+1)}{S_1}T(\Sigma_g s_i(s_i+1))$$

(A32) $$\Sigma_g(r_i+s_i)^2(\tfrac{q}{4}(r_i+s_i+5)+1)+K(q+1)\Sigma_g r_i(r_i+s_i+D+\frac{(q+1)}{S_1}T(\Sigma_g s_i(s_i+1)+K(K_g+g+\Sigma s_i))$$

Adjustments for parametric and endogeneous iterations are obvious.

Lemma 7:

(a) (14.16) increases initial calculations by $(q+1)T\Sigma_g r_i(2T+1)$.

(b) Therefore $(q+1)T^2 \sum_{g+1}^{M} r_i$ is saved from initial calculations.

Lemma 8:

Precisely, PEIIV is only done with $\hat{\beta}_i^{(S+1)}$ $i = 1 \ldots g$ as described if

(A33) $$(33) < \frac{1}{S_1}(\Sigma_g(q+1)s_i T(2s_i+T)+\Sigma_g \tfrac{q}{4}s_i^2(s_i+1))$$

which would require a very large T and small S_1.

Theorem 5:

Test should be ammended as follows

$$T < T^* \text{ where } T^* = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{where } aT^2 + bT + c \gtrless 0$$

replaces (34) and

$$\text{(A34)} \qquad \begin{cases} a = -\left(\frac{q+1}{S_1}\right)\left(2 + 2\sum_g r_i + \sum_g s_i\right) \\ b = \text{denominator of (34)} - \frac{(q+1)}{S_1}\, 2\Sigma s_i^2 \\ c = \text{numerator of (34)} - \frac{q}{4S_1}\, \Sigma s_i^2 (s_i + 1) \end{cases}$$

note $4ac < 0$ so a real solution is available and since $a < 0$ we need the the positive root for T^*.

Theorem 6:

Add to L.H.S. of inequality (35) the term

$$\text{(A35)} \qquad \frac{1}{S_1}\left\{(q+1)[KT\Sigma_g s_i + T\Sigma r_i s_i - T^2\Sigma(r_i + s_i) - T\Sigma(s_i^2 + 1)] - \frac{q}{4}\sum_g s_i^2 (s_i + 1)\right\}$$

which could be negative for large T or little overidentification.

Theorem 7:

(a) Add to L.H.S. of inequality (37)

$$\text{(A37)} \qquad \left(\frac{q+1}{S_1}\right) T\left[\sum_{g+1}^{M} s_i - K(\Sigma s_i + g(T+1))\right]$$

(b) add to inequality (38)

$$\text{(A38)} \qquad \left(\frac{q+1}{S_1}\right) T\left[\sum_{g+1}^{M} s_i + g(T-2K) - K\Sigma r_i\right]$$

Both will, in most circumstances, be negative, (A37) in particular.

Theorem 8:

(a) and (b) the initial calculations are (A37) and (A38) respectively. Thus the total computational burden according to (29).

Theorem 9:

(a) and (b) initial calculations as (A35) and (A34) using however (41)(in view of (39)).

Theorem 10:

The additional terms for initial calculations are as Theorems 8 and 9 above plus (A31) or (A32). Theorem 10 holds with respect to RFFP and RFIIV because they are computationally equivalent on iterated calculations, and assuming (41) holds then RFIIV actually is superior to RFFP if

$$\left(\frac{\log \tau_2}{\log \lambda^*} - a\right)(f_1 + f_2 + f_4 S_2 f_5) S_1 f_6 f_7 > (q+1)T \sum_g r_i s_i$$

which seems highly likely.

CHAPTER 14: ON THE USE OF ITERATIVE ESTIMATORS

14.1 Introduction

The proceeding chapters complete the review and analysis of iterative estimation for the model (2.5). The following chapter extends the analysis to the general nonlinear model. In this chapter we attempt to tie up some loose ends of the theory for model (2.5). In particular we make some comments on the practical use of iterative estimation. It remains to specify precisely how to guarantee convergence from an arbitrary start. We consider also how to use the tests of convergence, and the results on accelerations and optimal parameters, in practical applications where the objective is to minimise overall computations. The problem of selecting the desired solution from a set of multiple solutions must be solved, together with a test to indicate where multiple solutions may be expected. We also make reference to results on apparent convergence and the sensitivity of convergence to the orderings of equations.

14.2 Guaranteed convergence

14.2.1

Having established the existence of at least one fixed point in $F(\cdot)$ of the iterative process (Theorems 12.25 and 12.26), we have to select a particular algorithm to reach it. The main problem is in adapting the results of Chapters 8, 9 and 10 to the fact that the iteration matrix of any iterative estimator changes with every iteration. The convergence and comparison theorems and all acceleration results of those chapters are strictly only applicable to each step S, separately, and possibly to each RF instrument formation process. In theory this is quite satisfactory but in practice a formidable amount of computation is involved if they are to be applied exactly at each step. It

is desirable to provide some short cuts in the computation which do not detract too much from the formal results.

It is obviously convenient to prove

Theorem 1:

In both FP and IIV families there exist globally convergent iterative estimators for an arbitrary start.

Proof:

(a) Any IIV estimator using RF instruments, sphericality transforms included by Theorem 12.22, satisfies this theorem. We need only show RF instruments are always available. This is obvious if $\beta^{(s)}$ is nonsingular all S. Even if it is singular, or the model nonlinear in the endogenous part, RF instruments are always available by, say, GRF or semi-iterative methods on (11.29) by Theorems 10.7 and 10.8.[1] These theorems are applicable precisely since within the instrument iteration the iteration matrices are fixed for a given S. Such estimators have zero iteration matrices, barring errors in the instrument calculation. But an arbitrarily small tolerance τ_2 forces the instruments to be arbitrarily close to the true RF ones.

(b) Any estimator with a GRF, semi-iterative or second order stationary iterative instrument formation may be used such that Theorem 1 holds. Specifically a GRF, IIV or FP estimator with $P = \text{diag}\,(p_i)$ has the iteration

$$y_i^{(s+1)} = y_i^{(s)}(1+p_i) - Y_i^{(s)} p_i \hat{\beta}_i^{(s+1)} + (I - M_i)(y_i - Y_i \hat{\beta}_i^{(s+1)}) \qquad (1)$$

where $\dfrac{\partial \hat{\beta}_i^{(s+1)}}{\partial y_{tj}^{(s)}}$ is as in Theorem 12.7 or 12.15. Following the steps of those theorems, (1) gives

$$\frac{\partial y_i^{(s+1)}}{\partial y_{tj}^{(s)}} = \frac{\partial y_i^{(s)}}{\partial y_{tj}^{(s)}}(1+p_i) - \frac{\partial Y_i^{(s)}}{\partial y_{tj}^{(s)}} p_i \hat{\beta}_i^{(s+1)} + M_i Y_i^{(s)} (Y_i^{(s)\prime} M_i Y_i)^{-1} \cdot \frac{\partial Y_i^{(s)\prime}}{\partial y_{tj}^{(s)}} M_i \hat{u}_i^{(s)} \qquad (2)$$

1. This was also proved at theorem 11.10.

or its FP counterpart. Thus the iteration matrix is $J^{(s)} = (I+\beta^{(s)}P)\otimes I$. Hence a GRF estimator is made globally convergent if $P^{(s)}$ is chosen such that $\rho(I+\beta^{(s)}P^{(s)}) < 1$ every S. This is possible by Theorem 10.2. For the general iterative estimator with iteration matrix $G^{(s)}$ we have to choose $P^{(s)}$ such that $\rho(I+G^{(s)}P^{(s)}) < 1$.

(c) Similarly a semi-iterative estimator on matrix $G^{(s)}$ has direct instrument formation (10.28), which may be written for model (2.5) as

$$W^{(s+1)} = W^{(s)}H^{(s)} + (Z\Gamma^{(s+1)} \vdots 0) \qquad (3)$$

where[1] $W^{(s)} = (Y^{(s)} \vdots Y^{(s-1)})$ and $H^{(s)} = \begin{bmatrix} G_1^{(s)} & I \\ G_2^{(s)} & 0 \end{bmatrix}$

such that $G_1^{(s)} = \dfrac{2\alpha_{s+1}}{2-(b^{(s)}+a^{(s)})}G^{(s)} - (b^{(s)} + a^{(s)})I$ and $G_2^{(s)} = (1-\alpha_{s+1})I$.

Use $Y^{(s)}$, from $W^{(s)}$, as instruments (or regressors) to obtain $\beta^{(s+1)}$, $\Gamma^{(s+1)}$. We have an estimator formally equivalent to that member of the FP or IIV family indicated by $G^{(s)}$. Regard the estimation phase associated with (3) as

$$d_i^{(s+1)} = \begin{pmatrix} \hat{\delta}_i^{(s+1)} \\ \hat{\delta}_i^{(s)} \end{pmatrix} = \begin{bmatrix} Y_i^{(s)\prime}Y_i^{(s)} & Y_i^{(s)\prime}Z_i & \vdots & & \\ Z_i'Y_i^{(s)} & Z_i'Z_i & \vdots & 0 & \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots\cdots \\ & & \vdots & Y_i^{(s-1)\prime}Y_i^{(s-1)} & Y_i^{(s-1)\prime}Z_i \\ 0 & & \vdots & Z_i'Y_i^{(s-1)} & Z_i'Z_i \end{bmatrix}^{-1} \begin{pmatrix} Y_i^{(s)\prime} \\ Z_i' \\ Y_i^{(s-1)\prime} \\ Z_i' \end{pmatrix} y_i$$

$$i = 1 \ldots\ldots M$$

and consider the convergence process of $d_i^{(s+1)}$ which contains that of $\hat{\delta}_i^{(s+1)}$. Naturally $d_i^{(s+1)}$ can be written in an IV form also. Thus by the results of Chapter 12 the iteration matrix is $J^{(s)} = H^{(s)} \otimes I$. Therefore by Theorem 10.8, if $a^{(s)}$ and $b^{(s)}$ are chosen so that the ellipse of Lemma 10.7 contains the eigenvalues of $G^{(s)}$ in the complex plane, the estimator is globally convergent for every matrix $G^{(s)}$ at every step S. Moreover

$$\lim_{S\to\infty}\left\{\prod_{i=1}^{S} ||H^{(s)}|| = ||P_S \quad (G^{(s)})|| \geq |\rho(G^{(s)})|^S\right\} = 0$$

may be applied to (10.31). #

---

1. RFFP would have $W^{(s+1)} = W^{(s+1)}\ H^{(s)} + (Z^{(s+1)} \vdots 0)$; for IIV with RF instruments there is no change over the form of Chapters 11 and 12 by corollary 3 below. For RFFP $J^{(s)} = -(I-H^{(s)})H^{(s)-1} \otimes I$.

Corollary 1:

We may apply all the results of Section 10.6 to any semi-iterative estimator in view of

$$\prod_{i=1}^{S} ||H^{(s)}|| = ||P_s(G^{(s)})|| \quad \#$$

Corollary 2:

If instrument iteration (11.31) is accelerated by semi-iterative methods then convergence is independent of instrument formation methods in a global sense, whether or not the estimation iteration is accelerated by any particular method, for all iterative estimators.

Proof:

Application to Theorem 10.8 to Theorem 12.2, the iteration matrices being fixed each S. #

Comments:

(i) $G^{(s)}$ depends on the underlying method: $G^{(s)}$ is $\beta^{(s)}$ for direct instrument FP or IIV methods; $(I - \beta^{(s)})\beta^{(s)^{-1}}$ for FP methods with RF instruments; zero for IIV estimators with RF instruments. Fractional accelerations give $\alpha G^{(s)} + (1 - \alpha)I$ in place of $G^{(s)}$ and recursive $(I - \alpha L^{(s)})^{-1}(\alpha U^{(s)}+(1-\alpha)I)$ where α is chosen independent of S and $L^{(s)}$ and $U^{(s)}$ are the strictly lower and upper triangular matrices in $G^{(s)}$. Therefore we would never separately use a semi-iterative acceleration on an IIV estimator using RF instruments any more than we would employ fractional or recursive accelerations with $\alpha \neq 1$. Secondly it will be optimal, but tedious to compute, to obtain α^* anew for each step S. Thirdly, further accelerations may be imposed after the semi-iterative one such as fractional to give $\alpha_o H^{(s)} + (I - \alpha_o)I$, or recursive as in (10.29). These further sophistications are not considered in this study.

(ii) The importance of Theorem 1 is that we can easily guarantee convergence in any model with no prior knowledge of the parameters to be estimated by a method invariant to the current estimate of $J^{(s)}$ and hence the start.

This can only be achieved by IIV estimators using RF instruments.

Corollary 3a:

Comparison theroems will indicate IIV estimators with RF instruments as preferable to other methods for all models at any step S. Moreover $\alpha^* = 1$ in fractional or recursive accelerations at any step S by Theorems 12.19(ii) and 12.20(ii). Any estimator has the number iterations given approximately by $\log \tau_1 / \log \rho(J)$ so at any step S we predict these estimators will be the fastest anyway. Recursive accelerations save slightly on computer storage and are preferred for a second reason below (Section 4). #

Corollary 3b:

An optimal semi-iterative acceleration on an IIV estimator using RF instruments would have $\alpha_s = 1$ and $a^{(s)} = b^{(s)} = 0$ all $s \geq 1$ by (10.29a)[1] since $Z = \infty$. Such an estimator (as seen by Theorem 1) then collapses exactly to the original IIV estimator with RF instruments. Therefore any comparison theorems would indicate such estimators preferable to any other method on a criterion of average as well as asymptotic convergence rates. #

Remark:

Corollary 3 is of considerable importance since there is clearly nothing optimal in selecting α^* or orderings of equations to maximise the asymptotic convergence rate for the iteration matrix $G^{(s)}$ and then only performing one iteration with it. Evidently we want to maximise the average rates of convergence for each S rather than the asymptotic rates which we can never approach even as an approximation. Similarly although a comparison result may be valid for some $G^{(s)}$, it may not be for any other step. The estimators of Corollary 3 however are preferable on both counts at every step and is guaranteed to converge optimally without any additional computational effort.

---

1. The distinction of Lemma 10.7 then vanishes.

14.2.2

We conclude that the results of Chapters 8 - 10 may be applied exactly to IIV estimators using RF instruments.

The following section applies to other estimators which can only use those results as an approximation. The best approximation for all steps taken together would be to test convergence, select the best estimator by the comparison theorems, and calculate the optimal acceleration parameters, from the system of equations (2.5) estimated at the convergence point. This was done in Chapter 12 in local convergence theorems. Thus we can apply the convergence, comparison and acceleration results exactly as local results in the sense of being within a neighbourhood of the convergence point.

Since, in any practical situation, we have no a priori knowledge of the system (2.5) at the convergence point we can only use these local results if estimates within that neighbourhood are first obtained. This suggests a very simple estimator should be used for a few initial iterations to provide those estimates. Those initial iterations might be performed until either (i) the parameter estimates have stabilised, i.e. are changing sufficiently slowly to satisfy an auxiliary tolerance that is very weak (say 1% or even 10%); or (ii) the parameter estimates are consistent and asymptotically efficient in the sense of Theorem 12.3 so that tests, comparisons and acceleration parameters are consistently estimated; or (iii) a small number of iterations have been performed following $Y^{(0)}$ chosen close to Y^*. In the case of (ii), Theorem 12.3 implies a consistent start must be used if suitable estimates are to be achieved for a low value of S. There is also considerable empirical evidence that convergence is reached faster if consistent starts are used (e.g. Bodin (1974)) so we conjecture that a given set of initial iterations will approach the convergence point closer in these circumstances. However we cannot be sure in infinite samples that by relying only on asymptotic properties we are in fact necessarily closer to the convergence point than by other methods. In

the case of (iii) $Y^{(0)} = Y$ suggests itself for well fitting models but it is an inconsistent start. The most sensible strategy would be to use all three conditions.

Having computed estimates "within the neighbourhood" of the fixed point, convergence tests, comparison tests and acceleration parameters may be calculated, according to the results of Chapters 8 - 12, on the basis of these estimates and justified as locally correct, as well as consistent and efficient. If desired, and if computation has a low disutility, convergence and comparison tests and acceleration parameters may be reassessed at regular intervals thereafter. However it is suggested that one assessment is likely to be adequate if the three conditions of the previous paragraph are satisfied since otherwise the gains will be small in comparison with the computational burden of reassessment. Possibly an exception should be made if the spectral radius of the actual iteration matrix is close to unity because the estimate of iterations to convergence (Section 8.4) is highly sensitive to small changes in the spectral radius. In many cases this spectral radius will not have been evaluated precisely but a crude estimate may always be made by monitoring the iterations in $Y^{(s)}$ via von Mises' theorem.

Taking $G^{(s)}$ as any of the iteration matrices of comment (i) in the previous section, then

$$\lim_{S\to\infty} \left\{ \frac{y_{t1}^{(s)}}{y_{t1}^{(s-1)}} \right\} = \rho(G^{(s)})$$

where $y^{(s)}$ is obtained from the associated iterative estimator. Naturally $S = \infty$ is not feasible but despite the fact $G^{(s)}$ alters with each step we can obtain serviceable estimates of the spectral radius if $G^{(s)}$ is changing quite slowly. This is achieved "within the neighbourhood" of the fixed point as defined by conditions (1) - (iii) above.

Monitoring the iterates in von Mises' theorem may be used three ways. Let the spectral radius be assessed after a sufficient number of iterations "within the neighbourhood" of the fixed point. First we may check the estimates of the spectral radius against those obtained when conditions ($_i$) - (iii) were first satisfied. If the spectral radius is seen to increase significantly then alternative estimators or acceleration parameters are called for. A reworking of the convergence and comparison tests and acceleration parameters might be undertaken. If however the spectral radius is constant or decreases then it would be adequate to continue as for earlier iterates. Secondly, we may use this monitoring of the iterates to assess the effects of choosing various acceleration parameters. This was illustrated for the most likely recursive acceleration case in Section 9.3.2, but could equally well be used in other cases such as fractional acceleration or reordering the equations.

Thirdly, and most important, we can use this monitoring device to ensure convergence and in the case of semi-iterative estimators optimal convergence. In general for any method the von Mises' theorem enables us to check that the estimated spectral radius is less than unity. Following Theorem 1 in any circumstances we can guarantee convergence. The GRF case cannot be recommended. We could monitor the iterates to ensure suitable $P^{(s)}$ matrices are used at each step. It was argued (Sections 10.2.2 and 10.2.3) that the lack of any precise relation between the eigenvalues of $I + \beta^{(s)}P^{(s)}$ and $P^{(s)}$ makes the application of GRF methods very difficult. The choice of a suitable $P^{(s)}$ depends on all the eigenvalues of $\beta^{(s)}$ by Theorem 10.3 which implies substantial computation. Any method utilising monitoring however only provides satisfactory information on the dominant eigenvalue and therefore is no help in designing $P^{(s+1)}$. Additionally we have no satisfactory way of designing a "good" $P^{(s)}$ matrix.

Theorem 2:

However any method may be made convergent by superimposing a semi-iterative acceleration with $-a^{(s)} = b^{(s)} = \rho(G^{(s)})$ set by monitoring the iterates in the manner of von Mises. (The rate of convergence could be increased by lemma 10.7.)

Proof:

Theorem 10.9 demonstrates convergence follows at the rates given there if $a^{(s)}$ and $b^{(s)}$ are set as stated. Convergence could however improve as in Theorem 10.12. By Theorem 1 we may monitor the iterates by von Mises to obtain estimates $\hat{\lambda}^{(s)}$ of $\rho(H^{(s)})$ where $H^{(s)}$ contains the iteration matrix $G^{(s)}$ of the general underlying method. Let $\lambda^{(s)} = \rho(H^{(s)})$ have eigenvector $x^{(s)'} = (x_1^{(s)'} : x_2^{(s)'})$. Then

$$\begin{pmatrix} G_1^{(s)} - \lambda^{(s)} I & I \\ G_2^{(s)} & -\lambda^{(s)} I \end{pmatrix} \begin{pmatrix} x_1^{(s)} \\ x_2^{(s)} \end{pmatrix} = 0 \qquad (4)$$

and eliminating $x_2^{(s)}$ we have

$$[\lambda^{(s)}(G_1^{(s)} - \lambda^{(s)} I) + G_2^{(s)}]x_1^{(s)} = 0,$$

$$\text{i.e. } G_1^{(s)} x_1^{(s)} = \mu^{(s)} x_1^{(s)}$$

where $$\mu^{(s)} = \frac{\lambda^{(s)^2} + \alpha_{s+1} - 1}{\lambda^{(s)}}$$

In view of the structure of $G_1^{(s)}$ we have

$$\mu^{*(s)} = \left[\frac{2 - (a^{(s)} + b^{(s)})}{2\alpha_{s+1}}\right] \left[\mu^{(s)} + a^{(s)} + b^{(s)}\right] \qquad (6)$$

where $\mu^{*(s)} = \rho(G^{(s)})$, since $\mu^{(s)}$ is monotonically increasing in $\lambda^{(s)}$ by the

envelope property of Theorem 9.5. We know α_{s+1} so using the monitored estimate $\lambda^{(s)}$ and $a^{(s)} + b^{(s)} = 0$, we calculate $\hat{\mu}^{*(s)} = - a^{(s+1)} = b^{(s+1)}$ for the next step. #

We might set a and b initially arbitrarily, or by direct calculation of the dominant eigenvalue of G, or by the neighbourhood method above.

14.2.3

It may be important use use semi-iterative acceleration for three reasons:

(a) The maximising of average rates of convergence in terms of the iterated variable may reduce the total work in overall estimation iteration. However tolerances here will probably be relatively tight so it is unlikely to make the big contributions to the solution of equation systems that were suggested under decision 5 of Chapter 11 - it may still be significant as $G^{(s)}$ changes every step, especially in the context of the remark following Corollary 3.

(b) It is guaranteed that the iterative estimator will converge for any form of linear or nonlinear model and that it will do so with a stochastic iteration matrix which changes with every extra step S. The guarantee of convergence is bought at the small price of checking that the extreme eigenvalues do not stray outside the complex circle defined by a,b of Chapter 10. This can easily be done by monitoring the iterates in the convergence process. The convergence of any iterative estimator is very much more difficult to analyse than Chapters 8 - 10 would suggest since the iteration matrix is always stochastic and dependent on S. However by these means the convergence may be controlled rather than left to chance on a trial and error basis.

(c) Semi-iterative methods provide a very easy way to obtain a rather fast rate of average and asymptotic convergence for very little computation. Indeed such a method has the speed of optimal 2-cyclic SOR on G by theorem 10.9 or by (5) above.[1] At its simplest G =I-βis taken and evidently such a speed will be considered a satisfactory improvement over Jacobi, especially as it is

---

1. Which is the maximum gain over JOR by Chapter 9.5.

obtained without any of the complex calculations of Chapters 8 and 9 and may be monitored to retain good average rates of convergence. Naturally if other matrices say β_α or R_α can be used for G that have smaller spectral radii, then so much the better.

There are two further points. First since (5) shows the 2-cyclic relation from which maximum improvement over Jacobi is obtained. Because of the logarithmic relation between spectral radii and numbers of iterations (Section 8.4) there is little point in engaging in lengthy computations to find the optimal matrix for G. The main benefit in acceleration has already been obtained by setting G =I-βwith semi-iterative acceleration.

14.2.4

The above discussion covers, in general terms, the use of tests and acceleration parameters in practical work where we wish to minimise the total computational burden. There remains a difficulty in obtaining optimal SOR parameters for recursive acceleration. Bodin reports that on extensive testing he found $\hat{\alpha}^* = 2\left[1 + \left(1 - \rho(\beta)\right)\right]^{-1}$ always to be a very satisfactory substitute for α^* in the case of p-cyclic matrices. The author knows of no proof for this, let alone if it is suitable in a more general context. However it is suggested that this should be used as an initial value for α^* in the general model (which is likely to be nearly p-cyclic on reordering and renormalisation due to the sparse β matrix). Then superior α^* values may be searched for according to the 6 point scheme of Chapter 9. A further short cut is to take $\alpha = 1$ as the approximation to $\hat{\alpha}^*$, for large M, which avoids an initial estimate of $\rho(\beta)$. In any event the lengthy calculations of the true α^* amy be avoided by these means.

No more detailed analysis of how in practice to use the tests and superimpose optimal accelerations at each step of any specific estimator is undertaken because Corollary 3 indicates which estimators are simplest to use. Optimal properties, with guaranteed convergence (Theorem 2) faster than other

estimators, follow with no additional computation. This cannot be matched by other estimators, which are therefore only treated in general terms.

14.3 Plural convergence points

14.3.1

Theorem 3:

IIV estimators have at least $\prod_{i=1}^{M} r_i$ real convergence points.

Proof:

(5.10) and (2.42) show IIV estimators to be the solution to the necessary conditions for a turning point in the likelihood function (2.35) appropriate to the relevant information class. Corollary 7.1 shows each turning point is also obtained by direct differentiation of (2.35) where each eigenvalue μ_i is used in turn. There are $\prod_{i=1}^{M} r_i$ such eigenvalue permutations possible in the system as a whole. #

Corollary 4:

FP estimators have at least asymptotically $\prod_{i=1}^{M} r_i$ real convergence points. This is all the information we need because we are only concerned with asymptotic properties. It is not possible to state if the number of convergence points changes for small samples; but if it does the new ones are not of interest. Also if the small sample properties are indeed reasonably approximated by the large sample ones we can be confident in conjecturing Theorem 3 holds for FP estimators. #

Comment

Theorem 3 and Corollary 4 show the importance of recognising that plural convergence points will exist for all iterative estimators of models with non-cyclic β (although the majority of such points may well in practice be quite implausible or unacceptable). Iterative estimation not only fails to distinguish which of the estimates might be globally maximum likelihood from those

which correspond to local maxima or other turning points in the likelihood function (2.35). It is therefore vital we have some means of selecting the desired estimate set both to get consistent and efficient estimates (see Thoerem 2.15) and preserve the optimality criterion (indeed any rationale for iterative estimation if they are likely to show local minima of (2.35)!) The possibility of plural convergence points is often deliberately ignored as not being a serious problem. However Mosbaek and Wold (1970) and Bodin (1974) report them as a common occurance, especially where residuals are large or interdependence strong.[1] This section confirms that a search must be made by means of various starts $Y^{(0)}$.

14.3.2

Three remedies suggest themselves for the selection of the desired solution among those plural convergence points.

(i) A number of the solutions are likely to show some elements of $\hat{\beta},\hat{\Gamma}$ and $Y^{(s)}$ that are quite implausible or even infeasible. If a solution set contains any such elements, it may be rejected. This policy is necessarily somewhat subjective and contains no statistical justification such as we seek for Theorem 2.15 or in the optimality criterion. Economic theory will probably dictate the sign and/or size of only a few structural coefficients with any certainty. In other cases we will have little or no prior knowledge and some parameters may even be the object of various hypothesis tests. Any restrictions in the form of relationships between parameters may also eliminate some potential solutions. Generally we may have more information on the feasibility or plausibility of elements of $Y^{(s)}$ when compared to Y. We have already argued that $Y^{(s)}$ and Y should not be too dissimilar (following Corollary 4.5) and on that basis we can be confident about the appropriate signs and sizes of components of $Y^{(s)}$. Further we know something of the relative sizes of components of $Y^{(s)}$ (apart from those specified by the identities of the model).

---

1. Recall also differences in estimates in Chapter 3 indicating plural solutions.

(ii) Theorem 2.15 indicates the correct method to select the desired solution is to pick that convergence point which is associated with the global maximum of (2.35). Having obtained all possible convergence points, and eliminated those implausible or infeasible, it is very easy to evaluate (2.35) at each solution set. The global maximum is then obvious. By Theorem 2.15, the estimates thus obtained are the unique, consistent and asymptotically efficient ones by the maximum likelihood criterion, under the conventional assumptions of Section 2.2, and that at least a regular log likelihood function exists (in the sense of Dhrymes Lemma 3.13). However since FP and IIV may be specified independent of the existence of a likelihood function, it is convenient to give a test of the desired solution independent of likelihood functions.

(iii) We chose the convergence point for which $R^2 = \sum_{i=1}^{M} R_i^2$ is a maximum; noting that in either (2.6) or (2.30) at convergence $R_i^2 = 1-\hat{v}_i'\hat{v}_i/y_i'y_i$ by (6.5) and (6.6) where v_i is the observed value at convergence. Because OLS on (2.30) minimises $\hat{v}_i'\hat{v}_i$ (and IV on (2.30) is asymptotically equivalent to it) for FP (and IIV asymptotically) we want that convergence point which maximises R^2. This follows, as Edgerton (1973) shows, since $\bar{R}_i^2 = 1 - v_i'v_i/y_i'y_i$ and plim $R_i^2 = \bar{R}_i^2$. Under classical assumptions OLS on (2.30) provides unbiased predictions (either directly or equivalently through the implied restricted reduced form) and thus the consistently estimated version has $\bar{R}_i^2 \geq \bar{P}_i^2$ where $\bar{P}_i^2$ = plim P_i^2 if P_i^2 is the coefficient of multiple correlation for any other version. Thus plim $R^2 \geq$ plim P^2 where $P^2 = \sum_{i=1}^{M} P_i^2$.

The justification of picking the convergence point with maximal R^2 may be developed further. Since the residuals of (2.30) and the associated restricted reduced form are identical at convergence by (6.5) and (6.6), we may equivalently select the desired solution either from structural or reduced form estimated equations. But the reduced form estimates may be considered as functions of the starts by Theorem 4.7. Therefore we may apply the results

of Nerlove (1960) directly to show choosing the convergence point with maximal R^2 is the same as choosing the global maximum of (2.35). In short, searching over all starts $Y^{(0)}$ to maximise R^2 is the iterative maximum likelihood method of Nerlove on the reduced forms - excepting that the results strictly holds only asmptotically since the estimates themselves are in such a case consistent but not unbiased. Thus the procedure suggested by Edgerton is asymptotically equivalent to paragraph (ii) above and may be used independently of the specification of any likelihood function.

There are some theoretical disadvantages in the Edgerton method. The first is that it is valid only asymptotically and therefore we assume that in finite samples it is an adequate approximation in the sense that as T increases unboundedly we have a decreasing probability of making a wrong choice. But there is no guarantee that the convergence point with maximal R^2 in some finite sample will also be that with maximal $\bar{R}^2$. Secondly it is possible, but perhaps unlikely, that there is in some finite samples no convergence point having both the largest R^2 as well as R_i^2 each i = 1 .... M. Then the method loses validity and it is not clear which of the competing convergence points should be selected. In either difficulty, the only remedy is to return to (ii) above and seek the global maximum of (2.35); or equivalently take that version for which the ratio of total to explained sum of squares is a maximum.

14.3.3

Plural convergence points may also be analysed by means of Algebraic Fixed Point and its I.V counterparts, (Bodin (1974)). Although Bodin is able to provide some particular results for Summer's model by such analysis, he demonstrates that it is impossible to provide any general results even in this simple case. This approach was therefore avoided here.

14.4 Apparent convergence problems

14.4.1

The notation of apparent convergence was introduced by Bodin (1974) who provided some detailed analysis of the phenomenon. We adopt that analysis, and the definitions givern there. This aspect of iterative estimation is only treated for the bearing it has on our recommendations on how iterative estimators are best used.

Bodin demonstrates that any early tendency to converge is due to convergent properties of the "error" process $\eta^{(s,k_s)}$ of (12.22). True convergence at some later point comes about according to the appropriate convergence condition of Chapter 12; decisive is the spectral radius of the appropriate function of β. We are naturally interested that when we observe convergence that (i) it is not apparent convergence in the iterative estimation of a model which in fact is divergent for that estimator and (ii) it is not apparent convergence of a process that has true convergence on some fixed point not identical to the apparent convergence point.

To avoid apparent convergence the following action should be taken for any iterative estimator:

(a) RF instruments be used with small tolerance τ_2 so that errors $\eta^{(s,k_s)}$ are very small (if not zero) and cannot increase (c.f. (12.23)).

(b) Consistent starts only are used. Then by Theorem 12.3 $\eta^{(0,0)} = 0$ as iterates start constrained to be in Z-space, and being RF instruments are prevented from straying more than negligibly from that space. Then if convergence is observed, it is true convergence to a fixed point because, following Bodin, the iterates, being eigenvectors of the Z-space, correspond to eigenvalues within the unit circle. Any other eigenvectors are negligibly small corresponding to equally negligible eigenvalues.

(c) If convergence is observed, iteration should proceed with a rather smaller τ_1 to check if convergence is true or apparent.

14.4.2

An additional problem that may be encountered is that of apparent or no convergence due to mixing[1] of convergence points where plural convergence points exist (as they will). This problem was reported by Lyttkens (1966 and 1968). Bodin offers recursively accelerated instruments as a way of preventing such mixing since each component of the new instrument set is calculated in a prescribed sequence so long as at least the initial step takes $\alpha = 1$. Although his example is invalid, being that of a cyclic β matrix which by Theorem 3 has only one convergence point, this procedure would be helpful in any model.

14.5 Optimal equation orderings

We know from Chapter 8 that the convergence of recursively accelerated iterative estimators are influenced by $^{M!}/_{2!}$ different equation orderings. In that chapter we showed how the optimal ordering might be obtained in theory, minimising the spectral radius of the iteration matrix (Theorem 8.15) various rules for "good" orderings were produced, e.g. Theorem 8.14. However all such results depended on a precise knowledge of β which was appropriate in the context of solving a system of equations. In the iterative estimation context we have no knowledge of β a priori. And these results would hardly be useful "within a neightbourhood" of the solution, as defined in Section 2.2, because they are complicated to implement and we have no information on their sensitivity to small changes in β elements.

Bodin (1974) has treated this problem in detail for RFP. He proposes a scheme of analysis comprising 6 algorithms to select a "good" ordering a priori. In applications of FP, or IIV estimators with direct instruments, with recursive acceleration it is obviously sensible that Bodin's procedure be employed before estimation commences. The empirical results reported by Bodin show

---

1. Alternating between convergence points, each containing some elements of, say, two possible convergence points.

considerable improvements to the speed of convergence are obtained. It is therefore recommended here that his procedure be used. We note two further points.

(i) Bodin's empirical results show in every case that α^*, after a good ordering is chosen, is closer to unity than with the original ordering. It is suggested that this will very frequently be the case although we cannot prove it in general since the spectral radius cannot be shown to a monotonic function to the ordering vector (except within a sufficiently small neighbourhood of α^*).

(ii) Theoretically, IIV estimators with RF instruments have convergence properties invariant to the ordering of equations. Thus another advantage of such estimators is that any problems on ordering of equations may be ignored. However as an empirical matter to the extent there are errors in the instrument process and to cover those few iterations required, if Bodin's procedure is available on the computer, then it would be sensible to use it.

CHAPTER 15: NONLINEAR MODELS

15.1 Introduction

The development of the estimation theory set out so far is, in this chapter, completed by extension to the treatment of nonlinear models with additive disturbances such as may be derived from (2.2). Additive disturbances, at least under transformation, are assumed for three obvious reasons:

(i) that is the traditional assumption used in nearly every operative econometric model;

(ii) such theoretical analysis of the estimation of models (2.2), as has previously been undertaken, use this assumption;

(iii) it is not clear how estimators, other than maximum likelihood itself, would be formulated if this assumption were not made.

As a convenience, this will be assumed to be at least an adequate approximation to the true nature of the economy in question.

We note that the problem of nonlinearities in the predetermined variables (not their parameters) has been handled throughout previous chapters. The only point of significance is that the matrix Z contains quantities being the evaluations of those functions of predetermined variables rather than the variables themselves. Assumption (e) of section 2.2 must be interpreted accordingly.

Using the notation of chapter 2 we consider the model

$$F(X,D) + U = 0 \tag{1}$$

It is convenient to assume a normalisation of each equation and indeed this will be found in every nonlinear econometric model when finally specified. Thus (1) becomes

$$Y = F^{*}(X,D) + U \tag{2}$$

where the i^{th} equations, in (1) and (2), are respectively

$$f_i(X_i, \delta_i) + u_i = 0 \qquad \text{or } y_i - f_i^*(X_i, \delta_i) = u_i \tag{3a-b}$$

Note the functional values are time dependent even if D is constant.

The estimation problem of a general system of nonlinear simultaneous equations has received very little attention in econometric theory. Although a considerable body of practical experience with various estimator methods has been built up (e.g. Goldfeld and Quandt (1972) or Eisenpress and Greenstadt (1966)) there have been very few analytical results on the properties, statistical or computational, of such estimators. To rectify that situation is a very large job, requiring a complete study in itself. It certainly has no place in this work but would be a most important sequel to the results presented here. However it is necessary to provide the nonlinear complement to the foregoing chapters. There is therefore no attempt to cover all possible estimators of (2), nor to discuss the details of those estimators that are analysed in very general terms. Estimators are treated in their broad groups as before and we look for a structure in the families of estimators; in particular how our estimators of the linear model (LIML, FIML, 2SLS, 3SLS, FP and IIV family members) carry over to the nonlinear case. We attempt to answer the same questions: can we find an estimator that always exists under the availability criterion that is optimal with desirable statistical properties? We also aim to see how far the results of the previous chapters can be used in the estimation of (2). Particular attention therefore is paid to generalising the FP and IIV family members for (2) - also, and more importantly, to provide generally available estimators which avoid the necessity of making the assumption of independently and identically normally distributed disturbances in the model. Finally some general observations are made on the problem of selecting the most desirable estimator for estimating (2).

In section 2 we consider the form of analysis. In sections 3 and 4 some special cases of nonlinearities are treated. Section 5 carries the analysis of the general model (2). Statistical properties, and other properties, of the estimators considered are reviewed in section 6. The conclusions of this particular view of nonlinear estimation are in a final section.

15.2 The background analysis

The structural form (2) gives rise to a reduced form which may be expressed either in a transformed structural form

$$Y = F^*(X^*, D) + V \tag{4}$$

providing the definition $V = Y - F^*(X^*, D)$; where Y^* is defined by

$$Y^* = F^*(X^*, D) \tag{5}$$

Alternatively the reduced form may be written equivalently in the conventional style

$$Y = H(Z, U, D) \tag{6}$$

In general there will be plural solutions $Y^*_{(i)}$, $i = 1 \ldots p$ say, satisfying (5) due to its nonlinear nature and hence $H_{(i)}(.,.,.)$, $i = 1 \ldots p$ functions in (6) all of which, as the solutions of simultaneous nonlinear equations, will be nonanalytic except in the simplest cases. Since only one Y_t vector may be observed for each t, only one of the p solutions can be the true reduced form. It is assumed that we have selection rules[1] to pick out that desired solution

---

1. Such rules may include $Y^*_{(i)}$ on a dynamically stable growth path, certain components being nonnegative or within certain "reasonable" bounds or nondecreasing, etc. indicated by what is possible (acceptable) within the short run development of the economic system being modelled. Such an assumption is conventionally employed, e.g. Goldfeld & Quandt (1972), Edgerton (1974). We have already used it in section 14.3.2 (i), i.e. that $Y^*_{(i)}$ should not be too far from Y for all $t = 1 \ldots T$. See also the selection of the global maximum, or best fit, solutions

and with it the true reduced form. Naturally, therefore, such rules place restrictions on the transformation between structural and reduced forms and are thus part of both the model specification and its identification. If we follow Fisher's (1966) principles of identification in nonlinear models, we can see these rules are bound up in the identification criteria (as Kelejian (1970) argues, they may help identification). Since chapter 2 we have avoided discussion of the model specification and its identification which were, there, taken as given. Thus this assumption was implicit. (Without it, p plural solutions would be available for each plural solution that results from the nonlinear iterations of FP and IIV estimators, as discussed in chapter 14).

The main problem in extending the previous analysis to the estimation of (2) will be in constructing a REID formulation.[1] In particular $E(Y|Z) \neq F^*(X,D)$ since expectations are linear operations and therefore not invariant to non-linear transformations. Similarly, since $E(Y|Z) \neq H(Z,0,D)$, REID formulations cannot be obtained from (6), also a nonlinear function of U. Thus estimates of that conditional expectation, which is the REID specification, cannot be obtained by (5) or (6) as counterparts to (11.29) and (2.8); and it is of no consequence that (6) is nonanalytic. However, from (4), $Y = Y^* + V$ may still be obtained mutually orthogonal. Similarly since $F^* = F^*(X,D)$ is a stochastic matrix, it can[2], by Wold's decomposition theorem (Wold (1938)), also be split into orthogonal deterministic and stochastic parts. Thus write $E(F^*|Z) = H^*$ and $F^* = H^*+\varepsilon$. We should seek a method of estimating H^* to replace the REID specification (and the reduced forms used to estimate it). This would at least help, if F^* were linear in D, legitimising OLS or IV in the estimation phase. Otherwise there is apparently no call for a REID substitute.

---

1. Naturally where REID does not exist, neither will GEID, chapter 6 nothwithstanding.

2. We assume the expectations exist in what follows.

The additional difficulty in extending FP and IIV methods for (2) is that nonlinearity in parameters precludes the use of OLS or IV normal equations at the estimation phase. For those models where linearising transformations or approximations are not available or suitable, further modifications are necessary. In the case of 2SLS and 3SLS the same difficulty has been overcome by the use of distance functions. We adopt a similar procedure.

It can be seen that FP and IIV methods would require substantial modification for (2). A major difficulty is, in fact, to decide if some of the modifications proposed yield a method that can actually be termed an extension of FP and IIV. The following criteria were therefore employed:

(a) the estimators should have an OLS or IV normal equation structure even if their solution is not obtained by those means.

(b) They should degenerate into FP and IIV family members if the model is linear.

(c) FP methods should have a REID specification or a substitute for it. IIV methods should also have a REID specification, or a substitute, as the method of forming valid instruments (2.19). Alternatively the minimisation of a REID distance function would be adequate (as suggested at (2.17)).

(d) The methods should have the same properties of availability and optimality as their linear counterparts, plus, hopefully, the same statistical properties of consistency, asymptotic efficiency.

(e) The methods should be related to one another with the same family structure as their linear counterparts.

The above five points, while preserving the distinctive features of FP and IIV estimators, may see rather restrictive for selecting their nonlinear generalisations. Certainly many other nonlinear estimators - all of them iterative - are available and these are also considered for their ability to answer those questions posed earlier. Chiefly, we look at maximum likelihood methods and others related to them which are computationally quicker. It is

conceivable that a more liberal definition of FP and IIV would include some such methods.

15.3 Nonlinearity in the parameters

15.3.1

A point of considerable importance is that of incidence of nonlinearities in the specification of most simultaneous equation econometric models is very low. Very often, at a few points in a few equations nonlinearities occur only in variables. Typical examples are ratios of variables such as in specifying the real value of an economic quantity or per capita values or porportional rates of change or relative prices. Other examples occur with powers of variables or products of them such as with wage bills or man-hours for labour input. Alternatively it is very common to find econometric models with nonlinearities only in the parameters. Two classic examples are models with autoregressive errors,and models with partial adjustment, adaptive expectations or distributed lag variables. Again in other models combinations of these two cases may be observed; equations being specified as linear combinations of products of such nonlinear terms.

Therefore in this and the next section we consider estimators specifically for these cases, covering, it is argued, most econometric models. Additionally if, as has usually been the case, nonlinearities are so few in relation to the size of the complete model, then linear approximations at those points of nonlinearity will very likely be adequate for models not covered by those two cases.

On the other hand, there are models that are nonlinear in both variables and parameters in a way more complicated than either above. An example is the CES production function. Linearising transforms may not be available; and linear approximations may be difficult and thought to be quite inadequate(as Verdoon and Post (1964) argued for the dynamic effects of their capacity variable on the Dutch Economy). We therefore also review estimators for the general nonlinear case. Throughout we will specify estimators in their full information versions.

15.3.2

An IIV estimator for models nonlinear in parameters only has been derived by Hausman (1975) in a manner similar to (2.42). We write (2) as

$$XD(\delta) = YB(\delta) + Z\Gamma(\delta) = U \tag{7}$$

where $B(\delta)$ and $\Gamma(\delta)$ are nonlinear functions of an r dimensional vector of parameters to be estimated, δ. The jacobian of the likelihood function is constant with respect to time, so following the steps of the derivation of (2.42) we have the estimator

$$\delta^{(s+1)} = \left(W^{(s)\prime}\tilde{X}\frac{\partial D^{(s)}}{\partial \delta}\right)^{-1} W^{(s)\prime}y \tag{8}$$

with instruments[1]

$$W^{(s)} = (\Sigma^{(s)} \otimes I)^{-1}\hat{X}^{(s)}\frac{\partial D^{(s)}}{\partial \delta}\ ; \quad \tilde{X} = \text{diag}\ (X_i)$$

with $\hat{X}^{(s)} = \text{diag}\ \hat{X}_i^{(s)}$, $\hat{X}_i^{(s)} = [Z\hat{\Gamma}(\delta^{(s)})\hat{B}(\delta^{(s)})^{-1} : Z_i]$, $\frac{\partial D^{(s)}}{\partial \delta}$ being $\left.\frac{\delta D}{\partial \delta}\right|_{\delta^{(s)}}$

and $\Sigma^{(s)} = T^{-1}[XD(\delta^{(s)})]'\ [XD(\delta^{(s)})]$

It is clear that (8) occupies the same position to (7) as RFIIV does to (2.6); with asymptotic covariance matrix obtained at convergence δ^*, (partial derivatives being evaluated at δ^*),

$$\left(\frac{\partial D'}{\partial \delta}\hat{X}^{*\prime}(\Sigma^* \otimes I)^{-1}\tilde{X}\frac{\partial D}{\partial \delta}\right)^{-1} \tag{9}$$

---

1. Note both $\hat{X}^{(s)}$ and $\frac{\partial D^{(s)}}{\partial \delta}$ change with every iteration in this estimator. Hausman also considers nonlinear IV estimation where (8) holds but only $\frac{\partial D^{(s)}}{\partial \delta}$ alters with every iteration from a consistent start.

Similarly members of the IIV family defined for (7) may be obtained exactly analogously as IIV estimators for (2.6). Finally the convergence conditions are again identical except that β becomes $\frac{\partial B}{\partial \delta}$.

As may therefore immediately propose an FP estimator for (7), namely

$$\hat{\delta}^{(s+1)} = (W^{(s)'} \hat{X}^{(s)})^{-1} W^{(s)'} y \tag{10}$$

The same remarks apply with respect to specifying members of the FP family, and their relations to IIV family members. Convergence follows as in chapter 12 with $\frac{\partial B}{\partial \delta}$ in place of β. The covariance matrix for FP estimation is as (9). Clearly these FP and IIV estimators satisfy the criteria of section 2.

We note in passing that 2SLS and 3SLS estimators of (7) are given by Hausman (1975). For 3SLS we can use

$$\hat{\delta} = (\bar{w}' \tilde{X} \frac{\partial D}{\partial \delta})^{-1} \bar{w}' y \tag{11}$$

where

$$\bar{w} = \hat{Z}(\hat{\Sigma} \otimes Z'Z)^{-1} \hat{Z}' \frac{\partial D}{\partial \delta} \qquad \text{and } \hat{Z} = \text{diag } \{Z_i\}$$

and we iterate over $\frac{\partial D}{\partial \delta}$ only making it a special case of nonlinear IV estimation. $\hat{\Sigma}$ is some consistent estimate of Σ.

15.4.1 Nonlinearity in the variables

In the light of the remarks of section 1 on predetermined variables we can treat the model

$$Y = G(Y,Z)\beta + Z\Gamma + U \tag{12}$$

Such a model has been treated by Kelejian (1971) for 2SLS and Edgerton (1973) for FP. Considering the second approach we obtain the REID form as

$$E(Y_t|Z_t) = Y_t^* = E\{G(Y_t, Z_t)|Z_t\}\beta + Z_t\Gamma \tag{13}$$

The expectation in (13) may be approximated by a Taylor's series approximation about Y_t^*

$$E\{G(Y_t,Z_t)|Z_t\} = G(Y_t^*,Z_t) + \sum_{q=2}^{\infty} \sum_{n_q} \Delta(q,Y_t^*,U_t) \left(\frac{E\left(\sum_{k=1}^{m} (v_{tk})^{n_k} | Z_t \right)}{\prod_{k=1}^{M} n_k!} \right) \tag{14}$$

with $n_q = [n_1 \ldots n_M; \sum_{k=1}^{M} n_k = q]$

and

$$\Delta(q,Y_t^*,U_t) = \left. \frac{\partial^q G(Y_t, Z_t)}{\partial y_{t1}^{n_1} \ldots \partial y_{tM}^{n_M}} \right|_{Y_t^*} \tag{14a}$$

and naturally we set $U_t = E(U_t) = 0$, and remember $E(v_{tk}|Z_t) = 0$. Since the partial derivatives are known functions, say Δ_t^*, we can write (13) as

$$Y_t = G(Y_t^*, Z_t)\beta + \Delta_t^*\psi + Z_t\Gamma + V_t \tag{15}$$

where ψ are linear combinations of the conditional moments of V_t post multiplied by β. Now FP estimates can be obtained as

$$\hat{\delta} = (w^{*\prime}X^*)^{-1} w^{*\prime}y \tag{16}$$

where $w^* = (\Sigma \otimes I)^{-1}X^*$ and $X^* = \text{diag}\,[G_i(Y_{ti}^*, Z_{ti}) \vdots \Delta_{ti}^* \vdots Z_{ti}]$ in which Y_t^* is the prediction from (14). Obviously these estimates can be obtained iteratively for some start $Y_t^{(0)}$. Similarly then we can propose IIV estimators for this case as

$$\hat{\delta} = (w^{*\prime}X)^{-1} w^{*\prime}y \tag{17}$$

with $X = \text{diag}\,[G_i(Y_{ti}, Z_{ti}) \vdots \Delta_{ti} \vdots Z_{ti}]$; Δ_t being (14a) evaluated at Y_t. In practice we will have to truncate the expansion in (14) after say $q = q^*$ terms. Thus these estimates must be approximate and as such asymptotic efficiency will not be a property of them. Further there is an element of arbitrariness in the selection of q^*. However, from this basis all members of the IIV or FP families can be generated; RF instrument versions by (11.29).

Rather similar 2SLS and 3SLS estimators based on Kelejian (1971) are available for this case. Systematic parts of $G(Y_t, Z_t)$ are obtained by an unrestricted reduced form estimation of $G(Y_t, Z_t)$ elements on (low) order polynominals (all of the same order) of Z_t. Then either estimator continues with those systematic parts in the usual way. It is shown by Amemiya (1974) that these estimators are exactly minimum distance estimators with distance function

$$\phi(\delta) = (y - f^*)' \, \Sigma^{-1} \otimes H(H'H)^{-1}H'(y - f^*) \qquad (18)$$

where $f^* = \text{vec}\,(F^*)$ and H consists of those low order polynomials of Z_t.

15.4.2

Combined use of the estimators of sections 4.1 and 3.2 give IIV estimators or FP estimators in another kind of nonlinear specification previously mentioned. That is for the model (2) where F^* may be written in the form

$$Y = F(X_t)\, g(\delta) + U \qquad (19)$$

$F(X_t)$ and $g(\delta)$ are both matrices of nonlinear functions of their respective arguments. The estimators would be (8) for IIV with $w^{(s)}$ given by $w^{(s)} \frac{\partial D}{\partial \delta}^{(s)}$ of (16) evaluated at step s; or (10) with $w^{(s)}$ as in the IIV case. If, however, the model has a more general form than (19) we have to estimate Δ_t^* at each step in addition to the parameters in (15) which would require further approximations. One solution is a two part iteration within each step.

15.4.3

Various transformations and approximations of (2) may be possible such that one of the foregoing methods may be directly applied.

(a) the whole model may be linearised. In this case we have

$$y_{ti} = C_i + \sum_{j=1}^{M+K} (x_{tj} - x_{tj}^{(0)}) \left. \frac{\partial F_i^*}{\partial y_{tj}} \right|_{X^{(0)}} + U_t \qquad (20)$$

for (2) linearised about an arbitrarily selected point $X^{(0)}$, $D^{(0)}$, and

$$C_i = F_i^*(Y^{(0)}, Z^{(0)}, D^{(0)}) + \sum_{j=1}^{M} (D_j - D_j^{(0)}) \left. \frac{\partial F_i^*}{\partial D_i} \right|_{D^{(0)}} .$$

The selection of $X^{(0)}$ is arbitrary[1] and as such makes (20) rather unsatisfactory and even worse whenever it is felt that the approximation of the linear form is inadequate. However (20) may be estimated by any method of section 4.1. In particular if either (8) or (10) is used the latest estimates of $X^{(s)}$ may be used to revaluate the approximation, which is an improvement over the use of (11).

(b) Naturally transformations in some special cases may exist which put (2) into the separated form (19). Alternatively we can approximate (2), rather better than (20), by separation in the form (19) when a first order Taylor approximation is taken using $F(X_t)$ and $g(\delta)$ as variables.

(c) For some special cases of (2) transformations may make it linear in the parameters. The most obvious examples are models that are "log-linear". Although the result is not linear in variables it is then such that the REID substitute $F^* = H^* + \varepsilon$ may be used. The problem with the traditional REID formulation is that even if plim $T^{-1} Y^{(s)\prime} U = 0$ it is not the case that plim $T^{-1} F^*(Y^{(s)}, ., .)' U = 0$, in general. If however the model can be made linear in the parameters by one transformation then we can calculate $H^{(s)}$ in any of the ways of the conventional FP or IIV methods, or if there is only one

---

1. The sample means of X would be a good choice for $X^{(0)}$. $D^{(0)}$ is not actually used.

nonlinearity in the RHS endogenous part then the nonlinear elements of $H^{(s)}$ may be obtained as residuals, i.e. $h_j^{(s)} = f_j^{*(s)}(y_{ij}) = y_j^{(s)} - f_j(Z_j)$ which clearly leaves plim $T^{-1}H^{(s)'}U = 0$. Again any FP or IIV method can be manufactured for these circumstances. This was treated in Chapter 12.6.

15.5 General nonlinearities

We now consider (2) as it stands without the possibility of linear approximations or any transformation. We note that very few econometric models that have been proposed actually fall into this category, rather than one of the proceeding. Also it proves difficult to specify satisfactory FP and IIV estimators here. Our treatment is therefore brief.

15.5.1

Some methods for (2) have been discussed in the literature, specifically by Jorgenson and Laffont (1974) and Berndt et. al (1974).

The most obvious method is maximum likelihood and we can obviously proceed by (2.38) although this does not tell us the best computational method. Alternatively we can proceed as (2.42) without concentrating the likelihood function, or even by direct maximisation of the likelihood function as suggested in Chapter 2.5. The last mentioned has been fully discussed. A log likelihood function of (2) is

$$L = C + \tfrac{T}{2} \log \det \Sigma^{-1} + \sum_{t=1}^{T} \log \det J_t$$

$$- \tfrac{T}{2} \operatorname{tr}(T^{-1} \Sigma^{-1} (Y - F^*)'(Y - F^*)) \qquad (21)$$

where $J_t = \dfrac{\partial U_t}{\partial Y_t}$. Thus the equivalent to (2.41) is the solution D to the matrix equation

$$\left\{\left[\sum_{i=1}^{M}\sum_{j=1}^{M+K}\left(\sum_{t=1}^{T}(J_t')^{-1}\right)\frac{\partial J_t}{\partial D_{ij}}\right]F' - \left[\sum_{i=1}^{M}\sum_{j=1}^{M+K}\frac{\partial F}{\partial D_{ij}}\right]'\right\}F\ \Sigma^{-1} = 0 \qquad (22)$$

or

$$H'(Y - F^*(X,D))\Sigma^{-1} = 0 \qquad (23)$$

in obvious notation. Evidently to solve for D, iterative methods must be used. However (23) is an instrument variable form (c.f. (2.40)) and further it satisfies our criteria for being accepted as an IIV method. Thus we can create the IIV family by applying suitable algorithms to (23), as were discussed in Chapter 5.

A gradient method of maximising (21) when concentrated has been given by Berndt et al. We iterate

$$\delta^{(s)} = \delta^{(s-1)} + \lambda^{(s-1)}\ (R_T^{(s-1)})^{-1}\ (P^{(s-1)} - q^{(s-1)}) \qquad (24)$$

where $\delta' = (\delta_1' \ldots\ \delta_M')$

and

$$R_T^{(s-1)} = T\sum_{t=1}^{T}(P_t^{(s-1)} - q_t^{(s-1)})(P_t^{(s-1)} - q_t^{(s-1)})',$$

$$P = \sum_t\sum_k\sum_\ell (J_t')^{-1}{}_{k,\ell}\left[\frac{\partial J_t}{\partial D}\right]_{k\ell} \text{ and } q = \sum_t\left(\frac{\partial f_t^*}{\partial D}\right)\left(\sum_t f_t^{*\prime}\ f_t^*\right)^{-1} f_t^*$$

reevaluated at each step. $\lambda^{(s-1)}$ is chosen by a line search. However the line search is not necessary if (2.47) is observed. Similarly a conjugate gradient algorithm (without line searches) could be used with a probable reduction in computational burden. We refer of course to the analysis of Section 2.5 and to this extent we improve the results of Berndt et al, who point out convergence is assured and $T^{-1}\ R_T^{-1}$ estimates the variance covariance matrix of the parameters.

15.5.2 Minimum distance estimators

Such estimators may be used based on (18). The problem is the arbitrary choice of the instruments H - if we consider the general approach of Amemiya (1974) rather than the two special cases of 3SLS and Kelejian (1971). The algorithm proposed by Berndt et al to minimise $\phi(\delta)$ is the simplest to apply, being again a gradient method.

$$\delta^{(s)} = \delta^{(s-1)} - \lambda^{(s-1)} \left[G^{(s-1)\prime} Q^* G^{(s-1)}\right]^{-1} G^{(s-1)\prime} Q^* f^* \qquad (25)$$

where $G^{(s-1)} = \frac{\partial f^*}{\partial \delta}\Big|_{\delta^{(s-1)}}$ and Q^* is the distance metric; here $\Sigma^{-1} \otimes H(H'H)^{-1}H'$. The method (25) on (18) is convergent with $(G' Q^* G)^{-1}$ being the parameters' variance covariance matrix. Again line searches for $\lambda^{(s-1)}$ is not necessary and we could also apply a conjugate gradient method to reduce computations.

The interesting development in this section would be to see if any minimum distance estimator would be acceptable as an IIV or FP estimator. In view of criterion (6) of section 2, this would require $Q^{*(s-1)}$ of (25) to be reevaluated at each iteration. The obvious candidate for H would be $[H^* : Z]$, if it were known. If it were not, then estimates $H^{(s)}$ would have to be made at each step towards covergence at H^*. In view of the idempotency of Q^* it might be thought that FP estimates would result. However it will be noted that such a distance metric results in a 3SLS estimator, not FP, when we consider the linear model case in order to satisfy criterion (b). Thus even if some procedure were invented to calculate F^* and then split it into $H^* + \varepsilon$, no FP estimator, acceptable on our criteria will result. Put another way, to satisfy the necessary criterion (b) we must find a metric Q^* such that

$$Q^* F = (-Z\Gamma\beta^{-1} : Z) \qquad \text{or} \qquad (H^* : Z) \qquad (26)$$

and it may not be possible to find any H to satisfy (26) even when all the other matrices are known, let alone also find a method to estimate it. We could obtain Q* by a generalised inverse procedure (arbitrarily chosen)

$$Q^* = (H^* \vdots Z)F^-, \quad \text{(in practice } Q^{*(s)} \text{ using } F^{(s)}, H^{(s)} \text{ given below)} \qquad (27)$$

and (abandoning H) use (25) and (18) as FP provided Q* in (27) is symmetric and idempotent. Thus we conclude that an FP estimator generally does not exist for model (2) as evidently the condition[1] for its existence is that Q* in (27) should be symmetric and idempotent. However no such restriction is necessarily possible.

Alternatively we may abandon the symmetry and idempotency of Q*. The minimum distance estimator would then become an IV style estimator if (27) were used. Further since H* is unknown but would have to be estimated iteratively this could be called IIV. We still need a REID substitute and for this purpose the following (possibly not the best procedure) may be used. Calculate $F^{(s)} = F^*(Y, X, D^{(s)})$ at step S and then regress column-vectors from $F^{(s)}$ on suitable order polynomials of the predetermined variables to obtain $H^{(s)}$ as the systematic parts of

$$F^{(s)} = H^{(s)} + \varepsilon \qquad (28)$$

Naturally we assume, in general, nonlinear functions of the predetermined variables alone are not in (2); hence the use of polynomial approximations to (6) of an order dictated by the available degrees of freedom. Further since the predetermined nonlinear functions are by assumption asymptotically orthogonal to the structural errors so will therefore $H^{(s)}$ be uncorrelated

---

1. Criterion (a) and (b).

with them in the limit. An estimator using $Q^{*(s)} = \hat{H}^{(s)}(\hat{H}^{(s)'}\hat{H}^{(s)})^{-1}\hat{H}^{(s)'}$ with $\hat{H}^{(s)} = [H^{(s)} : Z]$ would also be classified as IIV; using it as FP would not be a distinct estimator in small samples, contravening criterion (e).

15.6 Statistical Properties

In this section we draw on the results of Jorgenson & Laffont (1974) for the estimators of section 5.

It is well known that the maximum likelihood estimator of the parameters of (21) is consistent, asymptotically efficient and normal. This is proved in general by theorem A2.7 (also theorem 2.15) and in many points in the literature (e.g. Jorgenson & Laffont, who include an explicit expression for the Cramer-Rao lower bound in the general case). The optimality of such estimates are obvious. Their availability is less obvious in that (21) must exist. The requirement is as previously $T > K + M$ for the existence of $T\Sigma^{-1} = F(X,D)'F(X,D)$. The same requirement holds for (24) to ensure the eixstence of $(R_T)^{-1}$. However to solve (22) iteratively appears to involve no such restrictions. A unique solution, that is one in which no generalised inverse is implied (chapter 4.3) either for $H'F^*$ or Σ in the notation of (23), does impose some conventional restrictions: $T > \max_i (r_i + s_i, M)$. We are therefore interested in estimators which satisfy the availability criterion as well as not requiring the assumptions underlying (21).

If it is assumed that plim $T^{-1} H \frac{\partial f}{\partial \delta}$ is finite with rank r, then Jorgenson & Laffont show that the minimum distance estimators here are consistent. Further if plim $T^{-1} H' \frac{\partial^2 f_i}{\partial \delta_j \partial \delta'}$ is finite all i = 1 .... M and j = 1 .... r and with normality of errors (predetermined variables having, in the limit, no distribution) the estimates are asymptotically normally distributed with the given means and variance-covariance matrices. However, asymptotic efficiency is more difficult. It is also shown that apart from estimators of the type in section 3.2 these estimators are not generally asymptotically efficient. Indeed improvements appear impossible from Jorgenson & Laffont's result of the asymptotic

equivalence of the minimum distance and efficient nonlinear IV estimators. Their optimality of course is unquestioned.

Estimators of section 4 are shown by Edgerton (1974) to be consistent and asymptotically normal but we have already concluded that they are inefficient.

Equation (25) is available if the inverse exists which generally follows whenever $TM > K + M$. Minimum distance estimators, solved directly by iterative methods, can always exist implying the generalised inverses of various matrices (chapter 4.3); also theorem 6.3 Theil (1971). More conventionally, clearly the requirement for a unique solution in such methods came from the inverses in $\left(\frac{\partial f^*}{\partial \delta}\right)' \Sigma^{-1} \otimes Q(y - f^*) = 0$, i.e. $T > \max_i (r_i + s_i, M)$.

15.7 Conclusions

On these statistical properties we can conclude that the use of maximum likelihood methods is preferred whenever they are available, (22) is obviously the most widely available. Otherwise, we should use (25) as it stands (with the suggested computational improvements) and avoid the arbitrariness of $(F)^-$ in (27). Following Chapter 10, we can guarantee that any semi-iterative estimator correctly applied to (2) will be convergent - but, depending on which algorithm presented here is so accelerated, we cannot necessarily have desirable statistical properties.

CHAPTER 16: THE USE OF FINITE SAMPLE PROPERTIES

16.1 Introduction

The scheme for selecting the desired estimator for models (2.2) or (2.5), based on asymptotic statistical properties and the analysis of computational burdens, is rather cumbersome. The most obvious point is that asymptotic properties have provided no discriminatory power at all except with respect to OLS. We have found many estimators which are consistent and asymptotically efficient in their information class. The immediate problem is, therefore, should we not discriminate between them on the basis of the moments of their finite sample distributions (in general, or failing that, particular cases)?

A companion point is a justification for the lexographic preference ordering[1] for asymptotic unbiasedness over efficiency. A superior estimator might be OLS, because although, for T large enough, OLS bias exceeds that of other methods, its variance might, at least in some cases, be smaller. Indeed it might be, for T small enough, that OLS had smaller bias than other methods. A third point therefore is to find if any of these estimated have minimum mean square error for finite T and as $T \to \infty$, and therefore could be ordered on this criterion. In addition we argued from the theorems of Sargan (1975), in Chapter 3, that asymptotic sampling theory is a good approximation to exact, even when models become very large and/or T - K becomes small. But being approximations, we must improve this theory, and estimator comparisons, with finite sample results where possible. To the extent distribution functions of estimates are functions of unused degrees of freedom, how much practical benefit accrues to estimators that save on degrees of freedom? Finally the central thesis of this analysis - that the estimator

---

1. Recall a lexographic ordering of statistical properties over computational was justified by t requirement to relax the statistical maintained hypothesis (Chapter 3).

selection requirement follows from the necessity of relaxing the statistical maintained hypothesis - compels the introduction to estimator comparisons of as many finite sample results as possible.

In Section 2 we seek minimum mean square error estimators. In Section 3 we review the existence of moments of finite sample distributions of estimates. In Sections 4 and 5 we resort to proof for general propositions by counter example regarding finite sample distributions. Finally Section 6 attempts to review how approximate distributions might help. The object is to find a sharper estimator selection procedure, thus taking the weight off computational burden analysis which is of secondary interest and somewhat model dependent.

16.2 Mean Square Error Estimation

16.2.1

The argument of this whole small chapter is dominated by the fact that even if we had all the finite sample properties, there are no differences between finite sample distributions of any estimator of the IIV or maximum likelihood families within a given information class.[1] Statistically all such estimators are identical if a common convergence criterion set $\{\tau_1, \tau_2\}$ is used (Chapters 5,11) - ignoring problems of rounding errors which might make this remark approximate. Thus no discrimination is possible between these estimators on finite sampling theory grounds and the job is correctly done by analysis of computational burdens.

Classifying 2SLS, 3SLS, LIVE, FIIV, the traditional principle components estimators, 2SPC, etc. as iterative estimators with a specified finite number of iterates, to answer any of the foregoing questions the model builder must, himself, trade off the benefits of any differences in finite sample statistical properties with his disutility of computation and the desirability to him

---

1. Considering of course convergence to the same parameter set.

of the maximum likelihood interpretation at final convergence.

Thus we have answered all our questions in the negative for IIV, and maximum likelihood estimators within an information class. Within this group we cannot discriminate further or improve the sampling theory. In sections following this one, we treat this group as a whole. Similarly if the disutility or desirability mentioned are low or high respectively we have answered the questions for the finite iterate iterative estimators. If not the analysis of the subsequent section applies.

The position of OLS, FP and minimum mean square error estimators are then the immediate concern.

16.2.2.

It is frequently observed that OLS, having some bias, is relatively efficient compared to simultaneous estimators. The suggestion that it might be the minimum mean square error (MMSE) estimator in the class of linear estimators is, however, false for both finite T and as $T \to \infty$.

Theorem 1:

A linear estimator for (2.5) always exists which dominates OLS, GLS or their full information versions, on the mean square error criterion in each information class. If $T \to \infty$ the MMSE estimator is the FP,IIV or maximum likelihood estimator of the appropriate information class and anything asymptotically equivalent to it.

Proof:

The general estimator (2.15) will be used. Proof for specialisations of that is immediate. We use a technique due to Chipman (1964). Consider the general linear estimator

$$\hat{\delta} = Ay + b \tag{1}$$

and select A, and B, a matrix and vector of constants, to minimise the mean square error

$$R = E(\hat{\delta} - \delta)\,(\hat{\delta} - \delta)' \tag{2}$$

Allow some prior information $\bar{\delta}$ with mean δ and a certain variance. Define $\bar{\bar{\delta}} = \bar{\delta} - \delta$ and assume $E(\delta - \bar{\delta})'u = 0$. Then we have by Chipman

$$R = E\{(I-A\hat{X})\bar{\bar{\delta}} + Au + [b-(I-A\hat{X})\bar{\delta}]\}\,\{(I-A\hat{X})\bar{\bar{\delta}} + Au + [b-(I-A\hat{X})\bar{\delta}]\}' \tag{3}$$

Now full information GLS has $A = (\hat{X}'\phi^{-1}\hat{X})^{-1}\hat{X}'\phi^{-1}$ and $b = 0$. Hence (3) implies $R = E(Auu'A')$ but in this specification

$$E(Au) \neq 0 \tag{4}$$

And for the same A, choosing $b = -E(Au)$ we have

$$R^{*} = E(Auu'A) + bb' + E(Au)b' + bE(u'A'); \quad \text{thus } R^{*} - R < 0 \tag{5}$$

$\delta = (\hat{X}'\phi^{-1}\hat{X})^{-1}\hat{X}'\phi^{-1}y - E(Au)$ dominates (2.15) on mean square errors, whether or not we ignore[1] the prior information, being independent of T and var$(\bar{\delta})$. Equally we obtain estimates $\hat{\delta}$ from the transformation of (2.5), the REID system (2.30). Here the MMSE estimator is immediate from Chipman,

$$\hat{\delta} = (\hat{X}^{*}{}'\phi^{-1}\hat{X}^{*} + \Omega^{-1})^{-1}\hat{X}^{*}{}'\phi^{-1}y + [I-(\hat{X}^{*}{}'\phi^{-1}\hat{X}^{*} + \Omega^{-1})^{-1}\hat{X}^{*}{}'\phi^{-1}\hat{X}^{*}]\bar{\delta} \tag{6}$$

where var $(\bar{\delta}) = \Omega$. Thus ignoring prior information, $\Omega \to \infty$ and (6) becomes the FIML estimator (2.41). But plim $\Pi^{+} = \Pi = -$ plim $(\hat{\Gamma}^{(s)}\hat{\beta}^{(s)-1})$, $S \geq 3$ which completes the theorem. #

---

1. We can ignore prior information by letting Var $(\bar{\delta}) \to \infty$.

Corollary 1:

We know the asymptotic MMSE estimator.[1] We do not know it for finite T. It is neither (2.15) nor (6) if $\hat{X}^*$ is unknown since replacing $\hat{X}$ by $X^{(s)}, \hat{X}^+$ throughout (our only estimates) in Theorem 1 leads to the same result. It is none of the estimators in this work. #

We consider if OLS or its other versions have a lower mean square error than the simultaneous estimators of the same information class in Section 4.

16.3 The existence of moments

We now restrict our attention to estimators considered in previous chapters, and take all IIV members as identical to the maximum likelihood estimator of their information class. To the extent the moments of finite sample distributions do not exist we can answer the remainder of the foregoing questions in the negative.

Theorem 2:

The moments of finite sample distributions of structural form estimates exist as follows:

(a) OLS group iff $K - s_i - r_i$ exceeds the order of the next higher even order moment.

(b) 2SLS, 3SLS group iff $T - s_i - r_i$ exceeds the order of the next higher even order moment.

(c) Maximum likelihood group, no moments exist of integer order.

Proof:

(a) (b) Mariano (1972) and extensions by Sargan (1975), (c) Sargan (1970) with specialisations of the general estimator that are obvious. #

Theorem 3:

The moments of integer order of finite sample distributions of restricted reduced form estimates of groups (a) and (b) above do not exist, and for (c) they exist up to order T - M - K, i.e. only if the likelihood function exists.

---

1. The result is obvious given consistency. It is stated, not proved, by Sargan (1975).

Proof:

(a) Sargan (1975); (b) McCarthy (1972), Sargan (1973); (c) Sargan (1973). #

Corollary 2:

Maximum likelihood group estimators are preferred for forecasting purposes: when T > K + M; group (b) for structural inference purposes.

Proof:

Since the distribution functions of all estimates must exist, nonexistence of second order moments implies thicker tails compared to distributions whose second order moments do exist. It follows the probability of observing large deviations from the central tendancy is relatively higher but that for small deviations relatively lower. (We eliminate group (a) on asymptotic MMSE grounds.) #

Remarks:

(a) We have one ground for discrimination between estimator groups. Otherwise we can answer the outstanding foregoing questions in the negative in comparing estimators of the maximum likelihood group with those not in it.

(b) Corollary 2 is borne out again and again in the superior forecasting results of the maximum likelihood group. In particular this is noted for FP, IIV families - see Bergstrom (1974), Bodin (1974), Agren (1972), Mosbaek & Wold (1970) and Lyttkens (1973) for references to extensive testing.

Comment:

The validity of Theorem 2 for nonlinear models is uncertain, since the proof follows from the singularity of the transformation from reduced form to structural form parameter estimate density functions.

16.4 Exact bias and variances:

We compare OLS and 2SLS as representatives of their groups for bias and variance, in some special models, where they are known exactly. If i^{th} equation has $r_i = 1$, normally distributed errors and Z nonstochastic;

Lemma 1:

$|E(\hat{\delta}_i - \delta_i)| \lesssim |E(\tilde{\delta}_i - \delta_i)|$ where $\hat{\delta}_i$ is the 2SLS estimate and $\tilde{\delta}_i$ is the OLS one ($q_i \gtrsim 2$).

Proof:

From Sawa (1972) Corollary 1, OLS bias less 2SLS bias may be deduced to be

$$\Delta b_i = \theta(\beta_i - \rho_i)e^{-\theta} \sum_{h=0}^{\infty} \frac{\theta^h}{h!} \frac{-n}{(m+h)(m-n+h)} \gtrless 0 \text{ if } (\beta_i - \rho_i) \lessgtr 0 \qquad (7)$$

where $\rho_i = \dfrac{\beta^*_i{}' \Sigma \beta^*_j}{\beta^*_j{}' \Sigma \beta^*_j}$, β_j^* = column j of β if $\beta_{ji} \neq 0$ defines j,

and $\theta = \dfrac{\Pi_{.j}{}' \bar{Z}_i' \bar{Z}_i \Pi_{.j}}{2\beta^*_j{}' \Sigma \beta^*_j}$; with also

$n = \dfrac{T - K}{2}$ and $m = \dfrac{T - s_i}{2}$ implying $m > n \gtrsim 0$; and $\bar{Z}_i \cap Z_i = \{0\}$

but $\bar{Z}_i \cup Z_i = Z$ with $\Pi_{.j}$ conformable from $\Pi_{..}$. If $\rho_i = \beta_i$ or $T = K$ then equality of biases holds, otherwise the theorem follows with inequality since (Sawa (1972) Corollary 4) both biases are of same sign, opposite to that of $(\beta_i - \rho_i)$. Extensions to γ_i since $\dfrac{E(\hat{\gamma}_{ji} - \gamma_{ji})}{E(\tilde{\gamma}_{ji} - \gamma_{ji})} = \dfrac{E(\hat{\beta}_i - \beta_i)}{E(\tilde{\beta}_i - \beta_i)}$ $j = 1 \ldots s_i$ is given by Richardson & Wu (1971). #

Comment:

Δb_i is an increasing function of T, and model size, in so far as they affect θ; and a decreasing function of q_i. It is unlikely to be sensitive to T otherwise, except when T - K is small, or to model size.

Lemma 2:

A sufficient condition that OLS has larger mean square error, when T is finite, is

$$\frac{w^2}{\xi^2}(\beta_i - \rho_i)\theta\left\{2\rho_i - \frac{(\beta_i - \rho_i)(2\theta - 1)}{2}\right\} - \frac{\theta}{2} + \frac{q_i - 1}{4} \lesssim 2 \qquad (8)$$

where $w^2 = \beta_j^{*\prime} \Sigma \beta_j^*$ and $\xi^2 = \beta_i^{*\prime} \Sigma \beta_i^* - (\beta_i^{*\prime} \Sigma \beta_j^*)^2 w^{-1}$.

Proof:

Following Sawa (1972) Corollary 2, OLS mean square error in $\tilde{\beta}_i$ less that of 2SLS in $\hat{\beta}_i$ may be deduced to be

$$n \left\{ \lambda_1 \sum_{h=0}^{\infty} \frac{\theta^h (-1)}{h!(m+h)(m-n+h)} + \lambda_2 \sum_{h=0}^{\infty} \frac{\theta^h}{h!} \frac{[\mu_1(m+h-1)(m-n+h-1) - \mu_2(m+h+1)(m-n+h+1)]}{(n+h-1)(m+h+1)(m-n+h-1)(m-n+h+1)} \right.$$

$$\left. + 2\lambda_2 \sum_{h=0}^{\infty} \frac{\theta^h}{h!} \frac{1}{(m+h-1)(m+h)} \right\} \qquad (9)$$

$$\text{where } \lambda_1 = \left[2\rho_i + \frac{\xi^2 (\mu_1 - \mu_2)}{w^2 . 2(\beta_i - \rho_i)} \right] \theta(\beta_i - \rho_i) e^{-\theta} \; ; \; \lambda_2 = \frac{\xi^2}{2w^2} e^{-\theta} > 0 \qquad (10)$$

$$\text{and } \mu_1 = \theta + \frac{2w^2}{\xi^2} (\beta_i - \rho_i)^2 \theta^2 > 0 \text{ and } \mu_2 = m-n + \frac{w^2}{\xi^2} (\beta_i - \rho_i)^2 \theta > 0. \qquad (11)$$

Combining the terms in (9), a sufficient condition for it to be positive is that the resulting numerator should have every term positive. After much tedious manipulation this sufficient condition is certainly satisfied when

$$\frac{\lambda_1}{\lambda_2} \leq \frac{(m+h)(\mu_1 - \mu_2) + 2(m-n+h+1)}{(m-n+h-1)} \quad \text{for all } h = 0,1, \ldots. \qquad (12)$$

Further (12) holds if (8) is satisfied. Extensions by Richardson and Wu (1971) who show if $E(\hat{\beta}_i - \beta_i)^2 \leq E(\tilde{\beta}_i - \beta_i)^2$ then $E(\hat{\gamma}_{ji} - \gamma_{ji})^2 < E(\tilde{\gamma}_{ji} - \gamma_{ji})^2$ and by lemma 1. #

Lemma 3:

The difference in mean square errors is zero if T = K (since n = 0 in (9)). #

Lemma 4:

Under this specification, OLS variances are not generally less than those of 2SLS.

Proof:

We need only a counterexample. OLS variance less that of 2SLS is (9) - $(E(\tilde{\beta}_i - \beta_i))^2 + (E(\hat{\beta}_i - \beta_i))^2$. By lemma 1 the effect of the last two terms is negative. Thus a counterexample is any model satisfying (8) by a suitable margin where the difference in squared biases is sufficiently small. An obvious case is where T is large enough by theorem 1 and lemma 1. #

Comments:

(a) (8) is sufficient but far from necessary for lemma 2. Two insignificant approximations were employed loosening the bound. But of real importance is that negative terms in the summation over h in (9) still give the result if they are outweighed by positive terms. Obviously, by (9) negative terms will result from high h values - but each term is $O\left[\frac{\theta^h}{h^2(h!)}\right]$ so convergence even for large θ will be very rapid as such negative terms are likely to be insignificant. The necessary condition will be considerably less stringent than (8) and depends chiefly on θ and m - n. It does not seem worthwhile to work out such a complicated expression for this special case.

(b) Since the finite sample distributions are identical if T = K (Swamy & Holmes (1971)) and 2SLS is the MMSE estimator if $T \to \infty$, OLS cannot generally dominate 2SLS. Even if the difference in mean square error is negative for some T, $\exists\ T_o < \infty$ such that $T > T_o$ implies the difference is a positive and nondecreasing function of T.

(c) (8) is sample size and model dependent. It also involves the true parameters of the model and could at best only be used approximately with $\hat{\beta}_i$ and elements from $V^{+\prime}V^{+} = Y'QY$ for unbiased tests of the superiority of 2SLS. The more complicated necessary condition is therefore unwarrented.

(d) From (8) it seems very probable that, for most econometric models of this kind, 2SLS will be superior in mean square error. The advantage of 2SLS increases with increasing θ (this is almost certainly implied by increasing T although strictly $\tilde{Z}_i$ is a set of stationary variables. We require Z

bounded for asymptotic theory but here T is finite so we may now recognise that in practice Z almost certainly contains nonstationarities). Unless the j^{th} reduced form is insignificant and an extraordinarily bad fit, $\theta > 1$ and is large. But $|\rho_i| \leq 1$ so the left of (8) is dominated by

$$-\left[\Pi_{jj}' \bar{Z}_i'\bar{Z}_i \Pi_{jj} \frac{(\beta_i - \rho_i)^2}{\xi^2} (2\theta - 1) + 1\right]\frac{\theta}{2} + \frac{q_i - 1}{4} \quad (\beta_i \neq \rho_i) \qquad (13)$$

Further ξ^2 is likely to be quite small especially if $|\rho_i|$ is relatively large. Thus (13) decreases, and (8) is more likely satisfied, with increasing T, increasing $|\beta_i - \rho_i|$, decreasing ρ_i but increasing goodness of fit in specification. The effect of model size is uncertain since as it increases both θ and q_i will probably increase. If they were to move roughly in proportion it benefits 2SLS since (13) is $- O(\theta^2) + O(q_i)$ possibly (8) is insensitive to model size. In fact (8) appears highly likely to be observed unless both T and $\beta_i - \beta_i$ are small (when biases are very similar) and both reduced forms are extremely bad fits with high positive correlation in ρ_i (i.e. an extremely bad specification) and the model is heavily overidentified. If the specification is indeed so bad, the inevitable misspecification biases will in any case render (8) irrelevant. But one conclusions is that OLS should be used for any experimental and preliminary testing of potential specifications and the other is large models with small sample size and dubious specifications are best estimated by OLS or non-2SLS methods.

(e) It is conjectured that these results will hold for general model specifications but no results are available yet.

(f) Simpler tests than (8) or (13) were not found since they all involve μ_1, μ_2, λ_1, λ_2 preventing simplification. See Corollary 4 below.

(g) There is no possibility of no unique preferences in the i^{th} equation because 2SLS is superior on one parameter and OLS on another (lemma 2). But it is possible that 2SLS is preferred for one equation and OLS for another. Thus attempts to generalise these results to full information estimators are unlikely to be successful.

It is also conjectured (for reasons given in Section 6) that the inclusion of lagged dependent variables will not materially affect these results if the errors u_i, u_j are serially independent. #

Results in section 3 indicate that no further comparisons of this kind with other estimators are necessary, or are even possible. The exact sampling distribution of LIML[1] models covered in this section is given by Mariano and Sawa (1972) so that probabilities of certain deviations from an observed mean could be compared for a given model. The distribution however is very complicated and depends as usual, on knowledge of the underlying parameters of the model. We therefore continue with corollary 2.

The results of this section only are exact for a particular subgroup of models. 2SLS has always lower biases (this is a justification of lexographic preferences in that no other precise results are available) and we expect frequently lower mean square error. At least by contradiction,

Theorem 4:

OLS does not generally dominate 2SLS on mean square error for any finite T. #

16.5 Approximate finite sampling theory

Many approximations to the very complicated finite sample distributions (and hence to their moments) have been given. We may improve our estimator comparisons to some extent in this approximate sense.

16.5.1

We use large θ asymptotics, which implies the small σ asymptotics of Kadane (1971). We assumed that $\lim_{T\to\infty} T^{-1}Z'Z \leq \chi^{*2} < \infty$ and $\theta \to \infty$ would contradict this assumption because a nonstationary Z is required unless $E(v_j'v_j) \to 0$. Thus $\theta \to \infty \Rightarrow T \to \infty$ but not conversely. Alternatively we can establish the legitimacy of this approach by treating the results as approximations of arbitrary accuracy for an arbitrary but suitably large T. The

---

1. thus FP, IIV families, approximately, as well.

latter will be necessary here since we already know exact results for $T \to \infty$.

All the estimators under consideration are k-class with k known, if we ignore FP, 2SPC and the restricted class. We have

Theorem 5:

For the model of section 4, the MMSE k-class estimator has k given by

$$k^* \begin{cases} = \dfrac{E(v_j'v_j)(\beta_i-\rho_i)^2(2n+q_i-2)}{2E(v_j'v_j)(\beta_i-\rho_i)^2(n+1)+\xi^2} + O(\theta^{-3}) \text{ if } q_i < 4 + \xi^2(E(v_j'v_j)(\beta_i-\rho_i))^{-1} \\ = 1 + O(\theta^{-3}) \text{ otherwise} \end{cases} \tag{14}$$

Proof:

Sawa (1972) Theorem 4 Corollary 1 in my notation. #

Corollary 3:

$0 \leq k^* \leq 1$ to order θ^{-2}.

Proof:

The first term in (14) is bounded above and away from [illegible]

$\left[\dfrac{2(n+1)}{(2n+q_i-2)} + \dfrac{q_i - 4}{(2n+q_i-2)}\right]^{-1} = 1$ whenever q_i satisfies the inequality given.#

Corollary 4:

To order θ^{-2}, OLS has lower mean square error than 2SLS iff either $T + q_i - s_i \leq 7$ or both $0 \leq \dfrac{w^2(\beta_i-\rho_i)^2}{\xi^2} < \dfrac{1}{T+q_i-s_i-7} < 1$ and $T + q_i - s_i > 7$.

Proof:

Sawa (1972) Theorem 5 corollary 2. #

Remarks:

(a) Neither LIML, FIML, nor any member of the IIV, FP families is a MMSE estimator among k-class estimators, let alone all linear estimators, for this model and large θ. That this is a completely general result follows from Theorem 2.

(b) (8) and lemma 2 will always hold whenever Corollary 4 might apply with any reasonable accuracy. Moreover since $\theta \to \infty \Rightarrow T \to \infty$ corollary 4 will certainly be negated when it can be held to apply. We are left only with lemma 2, unless the approximation of corollary 4 may be tolerated for finite T; but in the absence of further information its conditions may be used as necessary ones, with sufficiency from Lemma 2, for discrimination between OLS and 2SLS.

(c) Under corollary 2 OLS is MMSE if $\beta_i = \rho_i$, $T - s_i \leqslant 3$ and/or the reduced forms are a very poor fit and are mutually uncorrelated. Thus OLS is only superior in very extreme cases. To the extent these results may (approximately) generalise, it is suggested the arguments in favour of OLS are adequately refuted. It is not recommended for estimating the general model (2.5).

16.5.2

Approximations to density functions are useful as an approximate method of generalising the analysis of sections 4 and 5.1 to a general model. Those results may be treated only as circumstantial evidence since their validity in general specifications is not proven. Estimators with nonexistent finite sample moments may also be included since good approximate distribution functions have been found for the true distributions in such cases. Approximate moments for any given T may then be compared - in particular mean square errors - as well as straightforward comparisons of percentage points of those calculated distribution functions.

It is beyond the scope of this analysis to discuss the various approximations. Generally they are not amenable to analytical comparisons. In any particular case their method of employment is obvious for the objectives of this chapter, and their computation is straightforward. Any light shed on estimator comparisons however is exclusively restricted to situations of a given model.

The most general approximations are those of Sargan and Mikail (1971) for IV estimates - approximations of the standardised distribution functions of estimates to order $T^{-3/2}$. The approximation while completely general is rather complicated and depends critically on the true parameters of the model. Saragan (1970) develops a parallel approximation for FIML estimates with an exactly similar outcome. In either case the approximation is superior in accuracy to those of Nagar (1959,1962) which are expressed as functions of covariances between exogenous variables and structural disturbances.

Mariano (1973), Anderson and Sawa (1973), and Richardson and Wu (1970) provide approximations in terms of θ, which are closely related to the small σ approach of Kadane (1971). It appears the Anderson and Sawa version performs better than the others.

Some observations on the results are: (i) they are almost all completely model dependent and offer no general inferences, (ii) they are complicated and may involve nearly as much computation as numerical integration of the exact distributions, (iii) they may be poorer approximations at large T than the asymptotic sampling results traditionally used (Anderson & Sawa), (iv) they do not allow for lagged dependent variables, (v) in some cases the approximation deteriorates with increasing overidentification, (vi) they depend critically on parameters a priori unknown which prevents their effective use in practice.

Sargan (1974) investigates the general validity of his approximations, in the sense of when they may be used with reasonable confidence, with results for all the estimators considered here.

16.5.3

Approximations of the finite sample distribution of the traditional "t ratio" test statistic are also available. If the exact distribution of a endogenous variable coefficient is non-normal, then the exact distribution of the test statistic may take some equally difficult form. Moreover the

exact distribution of the coefficient is observed in earlier results to be significantly skewed, so the wisdom of using the asymptotic result of normality is questionable. For the model of Section 4 Richardson and Rohr (1971) provide the exact distribution of the "t ratio" in the case of 2SLS. It tends to t_{T-K} as $\theta \to \infty$ (and a standard normal as $T \to \infty$) or if $\beta_i = \rho_i$ but is rather complicated. The distribution is relatively insensitive to $|\beta_i|$ if $\theta > 9$ but for $\theta \lesssim 9$ they conclude significant errors will result from the use of asymptotic results. Further if T - K is very small then the distribution is highly skewed. The suggestion is that t_{T-K} may be a better approximation than the standard normal for all finite T.

For the general case, approximations for the distributions are introduced by Sargan (1975). Gram-Charlier approximations may be used whenever $\hat{\beta} - \beta$ and $\hat{\sigma}_\beta$ can be written as two independent stochastic variables, functions of $\hat{\Pi}$ and $\hat{V}$ respectively. For all our estimators we may evidently refer to the REID specification and they are therefore suitable for such approximations.

16.6 Monte Carlo evidence:

Further circumstantial evidence may be obtained from the many Monte Carlo studies of finite sample distributions. The results obtained (e.g. Summers (1965) Cragg (1965, 1967, 1968) Nagar (1960) Quandt (1965) etc.) support the foregoing analysis but it is clear that these results are most sensitive to the actual model and true parameters. (In many cases no unique ordering of preferred estimators is obtained and it appears that factors external to this study (e.g. degree of collinearity, misspecifications the stochastic specification, q_i, sparseness of β) are the factors which cause the preference ordering to diverge from that reached in this chapter. Except for very small T, asymptotic sampling theory is usually claimed adequate for consistent estimators, but not OLS. The OLS variance is often small but the mean square error is not. Consistent estimators have poor variances due to

thick tails rather than a wide spread about the mean. Quandt (1965) reports that, particularly, near triangularity of β improves the statistical performance of estimators - this will come about naturally if Bodin's reordering algorithms are used, which provides another reason to encourage their use.

Monte Carlo results are also useful in assessing the effects on finite sampling theory of the inclusion of lagged dependent variables, since all exact results assume their exclusion. Wagner (1958) and Nagar (1960) both present results which indicate little change in any of the foregoing conclusions except that biases and mean square errors of OLS are particularly increased. Basman, Richardson and Rohr(1974a,b) test the conjecture that exact distributions of estimators and their traditional test statistics in (2.5) are unaffected by lagged dependent variables - and the conjecture is accepted in nearly every case. Some evidence is given that any rejections are a result of the marginal densities of coefficients of lagged dependent variables being nonstandard normals rather than standardised as in asymptotic theory. Also their results, and Maddala (1974), support the suggestion of using the t distribution as an approximation to the "t ratio" test statistic as the highest powered test. None of these results are surprising since it was known that the effects of including lagged endogenous variables vanish very rapidly with increased T (Malinvaud (1970)).

16.7 Conclusion:

Finite sampling theory might have provided a sharper discrimination procedure for estimator selection and a more complete sampling theory of our estimators. We dismiss further consideration of finite sampling results for those purposes because: (a) there are no general exact results, (b) such results would not provide any comparisons within the maximum likelihood groups, (c) such exact results are too complex to implement analytically, (d) moments

often do not exist, making comparisons ineffective if they must make use of approximations, (e) unique orderings are very unlikely, (f) numerical approximations are usually available, (g) all distributions, and approximations, depend on model parameters a priori unknown making their effective use likely.

In addition to those conclusions we have the Corollary 2 and Theorem 4 with Section 5.1, but beyond that it appears impossible to relax further the statistical maintained hypothesis.

CHAPTER 17: THE ESTIMATOR GENERATOR FUNCTION AND CONCLUSIONS

17.1 Introduction

The results of this research are best drawn together by the summary descriptive device of an estimator generating function. This function spans the entire estimator space that has been considered, but since there are many methods obtaining any particular estimator it is presented in association with a specification of the structure of algorithms which may be used to obtain each estimator within this estimator space. Estimators may then be rapidly classified and analysed. The findings on the preferred estimator are summarised and finally some remarks made on the contributions, and shortcomings, of the research.

17.2 The estimator generating function

The estimator generating function is defined as

$$\hat{X}'\phi^{*-1}(A \otimes I_T - K^* \otimes Q)\hat{X}\delta = \hat{X}'\phi^{*-1}(A \otimes I_T - K^* \otimes Q)y \qquad (1)$$

Estimators, obtained as the solution to the system (1), contain without approximation[1] all those mentioned in this study by appropriate choices of A, ϕ^* and K^*. The switching matrix A selects the estimator class. Setting $A = E$ gives the K-matrix class; setting $A = I_M$ gives the k-class.[2] K^* has typical element $k^*_{ij} = k_{ij}a_{ij}$ $i,j=1 \ldots M$; the matrix K and remaining notation is that of Chapters 2 and 7. The choice of ϕ^* sorts estimators into information classes. Setting ϕ^* as $(\text{diag } \sigma_{ii}) \otimes I$, $\Sigma \otimes I$, $\text{diag } \Sigma_i$ and $Q^*(R^* R^{*\prime} \otimes I)Q^*$, (as Chapter 2), we have limited information, full information and their auto-regressive versions respectively.

---

1. But note comment (a) below.

2. There is a possibility of generalisation to other choices of A remembering K^* now stands for K of Chapter 7. Further generalisation is not relevant here however.

The choice of K* - equivalently K in this analysis - selects the estimator family. Considering $A = I_M$, $k_{ii} = 0$ $i = 1 \ldots M$ gives the least squares family (OLS, GLS, Zellner (1962) or Parks (1967)). For $k_{ii} = 1$ $i = 1 \ldots M$ we have the simultaneous least squares and IV family (2SLS, 3SLS etc.). If $k_{ii} = \hat{\lambda}_i$ $i = 1 \ldots M$ with $\hat{\lambda}_i$ from (2.32) we have the maximum likelihood and IIV (with the most desirable fixed point selected as in Chapter 14) families.

All these estimator families, plus additionally the general K-matrix estimator and the FP family, may also be obtained when $A = E$ for appropriate K.

Theorem 1:

System (1) is identical when $A = E$ and $A = I_M$ iff either

(a) $\phi^* = \text{diag}(\sigma_{ii}) \otimes I$, or $\text{diag}\ \Sigma_i$, or

(b) $K = \pm E$ and $k_{ii} = \pm 1$, $i = 1 \ldots M$.

Proof:

(a) Obvious from the diagonal or block diagonal structure of ϕ^*.

(b) Is the only case when $I \otimes I - \hat{k} \otimes Q = E \otimes I - K \otimes Q$ in view of the block diagonality of the former. #

Corollary 1:

The simultaneous least squares and IV family is obtainable from the K-matrix class with $K = E$ (see Scharf (1976)). #

Theorem 2:

The set of all k-class estimators is a subset of the set of K-matrix estimators, i.e. identical estimates may be obtained in the K-matrix class for any given k-class but not conversely. The k-class is a proper subset except for limited information versions or models having $r_i = 1$ $i = 1 \ldots M$.

Proof:

The estimates (rather than estimating equations) are identical if

$$\hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I-K\otimes Q)y = \hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I-K\otimes Q)\hat{X}[\hat{X}'(\Sigma^{-1}\otimes I)(I\otimes I-\hat{k}\otimes Q)\hat{X}]^{-1}\cdot (\hat{X}'(\Sigma^{-1}\otimes I)(I\otimes I-\hat{k}\otimes Q)y \qquad (2)$$

in the full information case. (2) Holds where

$\hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I-K\otimes Q)u_k = 0$ when $u_k = y-\hat{X}\hat{\delta}_k$

i.e. $\hat{X}'(\Sigma^{-1}\otimes I_T)(K\otimes Q)f = e$ (3)

when $f = u_k$ and $e = \hat{X}'(\Sigma^{-1}E\otimes I)u_k$.

A typical subset of equations in (3) is

$$\sum_{j=1}^{M}\sigma^{ij}\sum_{\ell=1}^{M}k_{j\ell}\sum_{t=1}^{T}(X_i'Q)_t f_{t\ell} = e_i \qquad i = 1 \ldots M. \qquad (4)$$

Write the r_i column vector

$$P_{i\ell} = \sum_t (X_i'Q)_t f_{t\ell} = \sum_t (Y_i'Q)_t f_{t\ell} \qquad i, \ell = 1 \ldots M \text{ since } Z'Q = 0.$$

Now (4) becomes

$$[\sigma^{i1}P_i \ldots \sigma^{iM}P_i]\text{vec}(K) = e_i \quad \text{where } P_i = [P_{i1} \ldots P_{iM}]$$

and (3) becomes

$$\text{Diag}(P_i)(\Sigma^{-1}\otimes I_M)\text{ vec }(K) = e \qquad (5)$$

We chose M^2 elements of K (with no restrictions at all) to solve (5) comprising $\Sigma r_i < M^2$ equations. This is always possible but the K correponding to a given $\hat{k}$ is by no means unique. Conversely we would seek $\hat{k}$ such that $\hat{X}(\Sigma^{-1}\otimes I)(\hat{k}\otimes Q)f = e$ given $f = u_k = y - \hat{X}\hat{\delta}_K$ and $e = \hat{X}(\Sigma^{-1}\otimes I)u_K$. Eventually

(5) becomes

$$\text{Diag}\,(P_i)(\Sigma^{-1}\otimes I_M)\,\text{vec}\,(\hat{k}) = e \qquad (6)$$

$\text{Vec}(\hat{k})$ has only M elements a priori nonzero but there are $\Sigma r_i \geq M$ equations. Since $r_i \geq 1 \quad i = 1 \ldots M$ a solution is only possible with equality. The final statement follows from that and theorem 1. #

Remark:

The K-matrix class is a generalisation over the k-class for full information estimators in models having at least one $r_i > 1 \quad i = 1 \ldots M$. We are therefore considering rival classes of estimators, one being a slightly more general class. A enables us to switch to the higher level of generality. However this is something of a luxury since all the estimators we are concerned with (excepting FP) are contained within the k-class which is almost certainly the preferred representation in view of Chapter 13.6. Some suggestions have been made for selecting $\hat{k}$ for other estimators, e.g. with improved small sample properties (Nagar's almost unbiased estimator, etc.), but there is no knowledge yet of sensible choices for K other than for FIML and 3SLS. Whereas (1) is presented for completeness, for practical purposes the k-class version is probably much preferred. #

Theorem 3:

Full information FP estimates are obtainable from (1) with A = E; only if $r_i = 1 \quad i = 1 \ldots M$ can they be obtained from (1) with A = I in full or limited information versions.

Proof:

Similar to (2) we have

$$\hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I - K\otimes Q)y$$
$$= \hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I - K\otimes Q)\hat{X}[\hat{X}^{*\prime}(\Sigma^{-1}\otimes I)\hat{X}^{*}]^{-1}\hat{X}^{*\prime}(\Sigma^{-1}\otimes I)y \qquad (7)$$

i e. $$\hat{X}'(\Sigma^{-1}\otimes I)(E\otimes I - K\otimes Q)u_{FIFP} = 0 \qquad (8)$$

where $u_{FIFP} = y - \hat{X}\hat{\delta}_{FIFP}$. Now writing $f = u_{FIFP}$ and $e = \hat{X}'(\Sigma^{-1}E\otimes I)u_{FIFP}$ we have (3) and the result follows Theorem 2. #

Comments:

(a) A lacuna in this theory, common to all K-matrix estimators requiring more than a diagonal K, appears because the limited information versions may not exist. K can be obtained from diag $(\sigma_{ii}^{-1}P_i)$ vec(K) = e but in strictly limited information specification off diagonal K elements will not be used. Thus limited information FP estimates are not contained exactly within the space spanned by (1) if $r_i > 1$, some i. However in practice estimates $\hat{\Sigma}$ will always indicate some, possibly very small, off diagonal elements so the lacuna is only a theoretical drawback to (1). Moreover FP is asymptotically equivalent to IIV, so on either count we may regard (1), for some K, a "good" approximation to FP.

(b) It is not suggested that (1) be used as a method of computation for FIFP estimates. It is merely a classification - as was the case for IIV.

Remarks:

(a) Since identification implies $\Sigma r_i < M^2$ any estimator obtainable as a k-class, plus FIFP, is not uniquely obtainable from (1) in the full information version. Not only are K-matrix estimators substantially rivals to k-class estimators, but also as a class they contain infinite duplications. In particular the FIML and 3SLS versions indicated are not unique. There will be other K matrices yielding identical estimates which may or may not be more difficult to obtain. The lack of uniqueness further detracts from (1) as a computational device.

(b) Any econometric estimator considered here is a special approximation to FIML since that is a particular case of (1) either for A = E or A = I. However no particular preferences on optimality criteria are used here.

(c) (1) obviously provides estimates but does not cover all the different algorithms which may be used to get those estimates. Clearly elaborations such as the double k-class or h-class (and their analogues in K-matrix terminology) may be included in (1). Clearly also estimators such as those not iterated to convergence, unrestricted estimators, those using different starts (e.g. 2SPC, LIVE/FIVE and estimators in Chapter 4.3 - 4.6) can be fitted into (1) for some K. Neither this last group of estimators, nor FP, IIV (or maximum likelihood) estimators using different algorithms to achieve convergence, can be sensibly individually distinguished in (1). The estimator generating function's purpose is therefore classification and analysis of the statistical properties of the components of the estimator space. To this end, recall that consistency holds for $\hat{\delta}$ iff plim K* = E and asymptotic efficiency iff plim $T^{\frac{1}{2}}(K^* - E) = 0$, (Scharf (1976)).

17.3 Algorithm structure specification

In conjunction with the estimator generating function, which is used to summarise the statistical characteristics of the estimators, we must take the algorithm specification which summarises the computational characteristics of any particular method selected to obtain given estimates. The algorithm specification is defined as the 5-tuple mentioned in Chapter 7.4.2. This algorithm specification will enable the computational properties to be classified or analysed.

The algorithm specification is composed of (in order): {the normal equation type; the information class calculation; starts; required number of iterations; algorithm type including any accelerations}. It spans the algorithm space for estimators included in this study. Each estimator may now be separately distinguished if this together with (1) are used as the summary discription.

The normal equations used here have been OLS, IV, k-class, K-matrix class, or direct maximisation of a likelihood function. The starts have been $Y^{(0)}$ or $K^{(0)}$ and $\Sigma^{(0)}$ as appropriate. The algorithm type might by any from Chapters 2.5 or 8-10 as required.

17.4 Preference ordering among estimators

The final conclusions on a preference ordering over the members of (1) is straightforward. The details may be invoked from earlier chapters. To satisfy our requirements on statistical properties, we consider maximum likelihood, simultaneous least squares and IV or iterative estimator groups. Other values of K* may be used for particular purposes, especially to obtain certain small sample properties, but that is beyond the scope of the present analysis - except that OLS was rejected on small sample properties (as well as asymptotic ones) in Chapter 16.

Starts for iterative estimators should be consistent in order that statistical properties should in no way be impaired[1] (being virtual properties). They should not be arbitrary or idiosyncratic, if possible, in order not to influence the estimates unduly when iteration is terminated by a finite tolerance. Finally they should be simple computationally. (4.26) is suggested as the best start.

These estimators are asymptotically equivalent in an information class (small sample differences are not available) so that further ordering is based on optimality, availability, computation and relaxing of assumptions.

Broadly there are now two sets of estimators. The maximum likelihood set (comprising maximum likelihood, FP and IIV) having the maximum likelihood optimality - a much favoured statistical criterion and arbitrary only in the assumption of normality in the errors - satisfying availability in the IIV form of (2.41) or (5.10) where also the maximum possible relaxation of the statistical maintained hypothesis. FP estimators are discarded in favour of

---

1. see also Chapter 14.4, Chapter 12.

IIV since they are only asymptotically as opposed to exactly equivalent to maximum likelihood. Further IIV estimates are clearly preferred for computational reasons: both for certain global convergence and that optimal SOR with semi-iterative accelerations and independence of the normalisation and equation order may all be achieved with no preliminary computation[1]. Against this, if T > K, the computational saving of "immediate" convergence to the global maximum may outweigh the relaxation of the statistical maintained hypothesis, for (7.1). Other methods of FIML computation are not recommended. In any event flexibility is preserved in this preference ordering.

The second set comprise 2SLS, 3SLS and their autoregressive versions. There optimality, in the form of minimising a distance function, is to some extent an arbitrary criterion and less generally accepted. The availability criterion is not satisfied and no relaxation of the statistical hypotheses is possible. Computationally they are extremely simple and have a clear advantage in robustness and numerical stability over the maximum likelihood set. On these criteria therefore they are recommended but lower in the ordering, unless either computational saving or corollary 16.2 are of the essence. Restricted estimators of various kinds swap availability for optimality but are clearly inferior to this set in computational simplicity for arbitrary starts, and therefore are lower still in the ordering.

17.5 Contributions and shortcomings

It was intended that this study should provide a coherent analysis of the structure of the estimator and algorithm spaces, and include in the process the newer econometric estimators. The statistical properties of these estimators and their computational properties together with those of conventional estimators has been established. Relations between them all were investigated and a preference ordering produced. A number of new results in convergence, tests and comparison theorems for the iterative schemes were

[1] as well as actual rather than virtual asymptotic efficiency.

obtained - as well as some new interpretaions of some of the estimators.

A major shortcoming is the inadequacy of the analysis of nonlinear estimators. However the purpose was to see how far our results extended to nonlinear estimation but not provide a nonlinear estimation theory itself. An analysis of numerical stability and error control possibilities in any of these estimators would be useful, especially in relation to its effects on computational efficiency. I should have liked to report more empirical results on the acceleration of the iterative algorithms and more experience with the actual computation of the estimators - particulary IIV. Finally the most glaring gap is the lack of small sample results which can be employed. But to overcome any one, let alone all, of these shortcomings would be to open yet another thesis.

APPENDIX: SOME MATRIX THEORY AND STATISTICAL REFERENCES

A.1 Matrix Factors which Preserve Stability

In this section new results are obtained which specify the eigenvalues of matrices which are products of other matrices with known eigenvalues. The results will be used frequently as it often turns out that iteration matrices of iterative estimators are simple functions of other matrices whose eigenvalues are known or may be obtained.

Lemma 1[1]:

For an arbitrary n x n matrix Q and arbitrary n-vector $V \neq 0$, if $QV > \lambda V$ where $\lambda \geq 0$ is a scalar then Q has an eigenvalue $\lambda_Q > \lambda$. Conversely if $QV < \lambda V$ then $\lambda_Q < \lambda$.

Proof:

By hypothesis $QV = \lambda_Q V = (\lambda + \varepsilon)V$ which implies, if $\lambda_Q = \lambda + \varepsilon$ defines ε, $\left(Q/(\lambda + \varepsilon)\right)^M V = V$ any $M \geq 0$. Given $\lambda \geq 0$ a scalar then

either $Q^M V < \lambda^M V$ follows if $\varepsilon < 0$ i.e. $\lambda_Q > \lambda$

or $Q^M V > \lambda^M V$ follows if $\varepsilon > 0$ i.e. $\lambda_Q < \lambda$. #

Theorem 1:

Let A and B be arbitrary n x n matrices and J=AB, such that A has eigenvalues $|\theta_i| < 1$ and B has eigenvalues $|\lambda_i| < 1$ $i = 1 \ldots n$. Define X and Y, n x n matrices, by $A = X\Theta X^{-1}$ and $B = Y\Lambda Y^{-1}$ where $\Theta = \text{diag}(\theta_i)$ and $\Lambda = \text{diag}(\lambda_i)$ (or the Jordan canonical forms where appropriate). Then $\rho(J) < \min\{\rho(A), \rho(B)\}$, and if $|J|$ has eigenvalues μ_i^* then $|\mu_i^*| < \min\{|\lambda_i|, |\theta_i|\}$.

---

1. Lemma 1 is a generalisation of Lemma 7, Ostrowski (1964).

Proof:

Let $U = Y^{-1}X$. Then $JY = ABY = AY\Lambda$ so $JX = AY\Lambda U$. Suppose B has simple roots i.e. $\Lambda = \text{diag } \lambda_i$

$$JY = \begin{bmatrix} \lambda_1 \Sigma a_{1j} y_{j1} & \cdots & \lambda_n \Sigma a_{1j} y_{jn} \\ \vdots & & \vdots \\ \lambda_1 \Sigma a_{nj} y_{j1} & \cdots & \lambda_n \Sigma a_{nj} y_{jn} \end{bmatrix} \qquad (1)$$

this JX has typical element, in the (p,q) position, $\Sigma_k \lambda_{kj} \Sigma_{} a_{pj} y_{jk} U_{kq}$.

$$\text{Hence } |JX| \leq \begin{vmatrix} |\Sigma_k|\lambda_k| \Sigma_j a_{1j} y_{jk} U_{k1}| & \cdots & |\Sigma_k|\lambda_k| \Sigma_j a_{1k} y_{jk} U_{kn}| \\ \vdots & & \vdots \\ |\Sigma_k|\lambda_k| \Sigma_j a_{nj} y_{jk} U_{k1}| & \cdots & |\Sigma_k|\lambda_k| \Sigma_j a_{nk} y_{jk} U_{kn}| \end{vmatrix} . \qquad (2)$$

But for any index $i = 1 \ldots n, Ax_i = \theta_i x_i$ (3)

x_i is i^{th} column of X.

Also $x_{ji} = \Sigma_k y_{jk} U_{ki}$ and therefore from (3) we have

$$|\Sigma\Sigma_{kj} a_{qj} y_{jk} U_{ki}| = |\theta_i x_{qi}| = |\theta_i| \, |x_{qi}| \quad \text{for } q = 1 \ldots n \qquad (4)$$

since θ_i and x_{qi} are scalars, (4) contains a summation over the index k with weights of unity. Given $|\lambda_k| < 1$, $k = 1 \ldots n$, then

$$\left|\sum_k \lambda_k \sum_j a_{qk} y_{jk} U_{ki}\right| \leq \left|\sum_k |\lambda_k| \sum_j a_{qk} y_{jk} U_{ki}\right| < |\theta_i| \, |x_{qi}| \quad i,q = 1 \ldots n \qquad (5)$$

Thus from (5) and (2)

$$|JX| < |X| \, |\theta|, \text{or } |Jx_i| < |\theta_i| \, |x_i|, \; i = 1 \ldots .n \qquad (6)$$

By Lemma 1 $|\mu_i^*| < |\theta_i|$ $i = 1 \ldots . n$. Further $\rho(|J|) \geq \rho(J)$[1] so $\rho(J) < \rho(A)$.

We may now reverse the argument. J and J' share the same eigenvalues so we can repeat the above steps on J' and B'A', assuming A has simple roots, writing θ_i for λ_i and visa versa; b_{ji} for a_{ij}; x_{ji} for y_{ij} or y_{ji} for x_{ij} throughout. We define $V = X^{-1}Y$ and replace U_{ij} with V_{ij} throughout. We conclude $|\mu_i^*| < |\lambda_i|$ $i = 1 \ldots . n$ and $\rho(J) < \rho(B)$ which completes the proof when A and B have simple roots.

Finally suppose B is not symmetric and has repeated roots, say

$$\Lambda = \begin{pmatrix} \lambda_1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_n \end{pmatrix} \quad \text{then } JY = \begin{bmatrix} \lambda_1 \Sigma a_{1j} y_{j1}, & \Sigma a_{1j} y_{j1} + \lambda_2 \Sigma a_{1j} y_{j2} \ldots\ldots \\ \vdots & \\ \lambda_1 \Sigma a_{nj} y_{j1}, & \Sigma a_{nj} y_{j1} + \lambda_2 \Sigma a_{nj} y_{j2} \ldots\ldots \end{bmatrix} .$$

i.e. $\lambda_1 = \ldots\ldots\ldots = \lambda_n$

Consequently (2) must be ammended. In order to produce the inequalities (6) it is now necessary for $q = 1 \ldots . n$ and all i where the root λ_i is repeated that

$$\left|\sum_{k=1}^{n} \lambda_k \sum_j a_{qj} y_{jk} U_{ki} + \sum_{k=1}^{n-1} \sum_j a_{qj} y_{jk} U_{k+1\, i}\right| \leq \left|\sum_{k=1}^{n} \sum_j a_{qj} y_{jk} U_{ki}\right| . \qquad (7)$$

---

1. This is a well known result derived from the Perron-Frobenius theorem: proof given by Theorem 2.8 Varga(1962).

Let Y_{ij} be the cofactor of y_{ij} and y^{ij} be a typical element from Y^{-1}, then the extra term in (7) may be written

$$\sum_{k=1}^{n-1} \sum_{j} a_{qj}\, y_{jk} \sum_{\ell} y^{k+1\ell}\, x_{\ell i} = \sum_{k=1}^{n-1} \sum_{j} \sum_{\ell} a_{qj} y_{jk}\, Y_{\ell k+1}\, x_{\ell i} \qquad (8)$$

where the inverse is written in cofactor form, and det Y = 1 since Y is a unitary matrix. The summation over index k is an expansion by alien cofactors in Y, so the expression in (8) has value zero. Thus (7) is satisfied as was (5), and inequalities (6) hold . If, in addition, non-symmetric A has repeated roots (3) and (4) continue to hold unaltered. As above we can reverse the argument and conclude, with the aid of an expansion by alien cofactors in X, that $|Jy_i| < |\lambda_i|\ |y_i| \quad i = 1 \ldots n$. Thus the theorem holds for arbitrary matrices A and B. #

By theorem 1 we have obtained relations between μ_i^*, λ_i, θ_i $i = 1 \ldots n$, but only a relation between $\rho(J)$, $\rho(A)$ and $\rho(B)$. It would be attractive to derive relations involving μ_i similar to those for μ_i^*, but unfortunately we have no relationship between μ_i and μ_i^*, beyond $\rho(|J|) \geq \rho(J)$ from Perron-Frobenius. The obvious possibility that $|\mu_i^*| \geq |\mu_i| \quad i = 1 \ldots n$ does not follow. Consider the counter-example $J = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$. Then $|\mu_1| = |\mu_2| = \sqrt{5}$ but $\begin{pmatrix} |\mu_1^*| \\ |\mu_2^*| \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

We must complete the story for μ_i by another method (theorem 2 below). Meanwhile some corollaries follow from theorem 1.

Corollary 1:

If $\lambda_i = 1 \quad i = 1 \ldots n$ then $\mu_i^* = \mu_i = \theta_i$, by equality in (5) and there being no need to consider absolute values throughout. This is a well known result since B is then a unitary matrix. #

Corollary 2:

If for some i $\lambda_i = 0$ and $B \neq 0$ then either $y_{ji} = 0$ $j = 1 \ldots n$ or $\sum_k a_{jk} y_{ki} = 0$ $j = 1 \ldots n$ so either $\theta_i = 0$ or $x_{ji}=0$, $i=1 \ldots n$. It follows from (6) that $\mu_i = \mu_i^* = 0$. #

Corollary 3:

If, say, B is idempotent the non-zero eigenvalues of J are precisely those of A. This statement follows since then λ_i are all zero or unity and we apply corollaries 1 and 2. #

Corollary 4:

If $|\lambda_i| > 1$, $|\theta_i| > 1$ $i = 1 \ldots n$ then $|\mu_i^*| > \max \{|\lambda_i|, |\theta_i|\}$ and $\rho(J) > \max \{\rho(A), \rho(B)\}$. #

Corollary 5:

If $|\lambda_i| \leq 1$ and $|\theta_i| \leq 1$ we replace all "<" by "≤" above. #

For situations other than the above for λ_i, θ_i and for a relation involving μ_i we need:

Theorem 2:

In the notation of theorem 1 let H be defined by $H - AHA^* = I$, where A^* is the conjugate of A, and in this theorem λ_i do not necessarily lie in the unit circle $i = 1 \ldots n$. If h_i are the eigenvalues of H then $\rho(AB) < 1$ iff $|\lambda_i| < \sqrt{h_i/(h_i-1)}$ $i = 1 \ldots n$.

Proof:

By Steins theorem (e.g. Taussky (1964)) $\rho(AB) < 1$ iff $\exists H_1 \in \Pi$ such that

$H_1 - ABH_1B^*A^* \in \Pi$ where Π is the set of positive definite matrices. Now let T be the unitary matrix such that $THT^{-1} = \hat{H} = \text{diag}\,(h_i)$. Then $\hat{H} - E\hat{H}E^* = I$ from the definition of H, where $E = TAT^{-1}$, so $E\hat{H}E^* = \text{diag}\,(h_i - 1)$ and hence $E = \text{diag}\,(\sqrt{(h_i - 1)/h_i})$. Also by Steins theorem in this case, $H \in \Pi \Rightarrow h_i \geq 1 \quad i = 1 \ldots n$. Further by theorem 5 Taussky (1964) iff $\rho(AB) < 1$ then $H_1 - ABH_1B^*A^* = P_1$ can be solved with $H_1 \in \Pi$ for <u>every</u> $P_1 \in \Pi$. We chose P_1 such that $\hat{H}_1 = H$. Then $\hat{H} - E\hat{F}\hat{H}F^*E^* \in \Pi$ iff $\rho(AB) < 1$, where $F = TBT^{-1}$, or

$$E^{-1}\hat{H}E^{*-1} - F\hat{H}F^* \in \Pi \,. \tag{9}$$

Using the conventional definition of a positive definite matrix, for an arbitrary n-vector $X \neq 0$, $X'(E^{-1}\hat{H}E^{*-1} - F\hat{H}F^*)X > 0$ satisfies (9). Rewritten this becomes

$$\frac{h_i^2}{h_i - 1} > \sum_j \sum_k h_k f_{jk} f_{ik} = \sum_k h_k f_{ik}^2 + \sum_{j \neq i} \left(\sum_k h_k f_{jk} f_{ik} \right) \quad i = 1 \ldots n.$$

i.e. $\dfrac{h_i^2}{h_i - 1}$ must lie outside the union of the Gerschgorin circles of $F\hat{H}F$

or $\dfrac{h_i^2}{h_i - 1}$ must be larger than all the roots of $F\hat{H}F$ for each i, which is

a sufficient condition on B. For a necessary condition we use a theorem proved by Dhrymes[1]: if $C \in \Pi$, $D \in \Pi$ then $C - D \in \Pi$ iff $|D - \gamma C| = 0 \Rightarrow \gamma_i < 1$, $i = 1 \ldots n$. Let $E^{-1}\hat{H}E^{*-1} = C$ and $F\hat{H}F^* = D$. We have $\hat{H} \in \Pi$ so $C \in \Pi$ and $D \in \Pi$. Note $C = \text{diag}\left(\dfrac{h_i^2}{h_i - 1}\right)$. Thus, given A, a necessary and sufficient condition on B such that $\rho(AB) < 1$ is D to have roots $d_i < \dfrac{h_i^2}{h_i - 1}$ $i = 1 \ldots n$, since an eigenvector matrix T_γ exists where

---

1. Dhrymes (1970) proposition 18 p.582.

$C^{-1}DT_\gamma = \hat{\gamma}T$ and $\hat{\gamma} = \text{diag}\,(\gamma_i)$ and therefore $DT_\gamma = \hat{\gamma}^*T$ where $\hat{\gamma}^* = \text{diag}\left(\frac{\gamma_i h_i^2}{h_i - 1}\right)$ and $0 \leq \gamma_i < 1$. Moreover D has an eigenvector matrix T_d where $\hat{F}\hat{H}\hat{F}^*T_d = \hat{d}T_d$ with $\hat{d} = \text{diag}\,(d_i)$ and $|\hat{F}\hat{H}\hat{F}^* - d_i I| = 0$. We may then write $FF^*T_d = \text{diag}\,(d_i / h_i)T_d$. Thus F has roots within the complex circle, radius $\left(\frac{h_i}{h_i - 1}\right)^{\frac{1}{2}}$; and T being a unitary matrix this implies for $i = 1 \ldots n$. $|\lambda_i| < \left(\frac{h_i}{h_i - 1}\right)^{\frac{1}{2}}$ as the necessary and sufficient condition for (9). #

We note $h_i \geq 1$ so $|\lambda_i| \geq 1$ is possible. Alternatively given a divergent matrix B, then $\rho(AB) < 1$ iff A is such that $|\theta_i \lambda_i| < 1$ $i = 1 \ldots n$. It is in fact easy to find H given only A. We have $h_i = \frac{1}{1 - \theta_i^2}$ $i = 1 \ldots n$. Indeed H itself may be obtained easily as in the Meyer-Spasche test (equations 8.18 and 8.18a) though this would not be very helpful. Clearly a good rule for the suitability of B is that $\rho(B) < \frac{1}{\rho(A)}$ and choosing a diagonal B (if the problem be one of choice) simplifies the use of the theorem.

A.2 Cramer's Generalised Theorem and an Extension:

It is feasible to derive and analyse only asymptotic statistical properties of the estimators under review. Therefore any test of significance on individual parameters, sets of parameters, goodness of fit of a particular equation, or any other hypothesis derived from the model, is only strictly valid as $T \to \infty$. It may, as it always has been in the past for finite T, be treated as an approximation to the exact test. To perform any tests of hypothesis we need the asymptotic distributions of the traditional test statistics. For models of the form (2.1) and (2.4) it is claimed, under all the estimators considered, the asymptotic distribution of any particular traditional test statistic is identical; is independent of the distribution of U; and is the limiting distribution of the corresponding test statistic for a single equation regression where Gauss-Markov assumptions hold.

---

1. There is a suggestion (Maddala (1974)) that to use the t distribution with degrees of freedom appropriate to OLS estimation is a superior approximation. See also Chapter 16.

These results allow us to consider hypothesis testing once only for all estimators and require only that the formulae used in these tests be produced for each estimator in turn. These results are derived here for reference, in a general form that covers (2.1) and (2.4). They are however well known in a variety of special forms.

The estimators under review are three sorts - either least squares, or instrumental variable, or maximum likelihood estimators. Consider first a model of the form (2.4).

Theorem 3:

Let A be a n x T matrix and y a t-column vector of random variables, arbitrary n, such that for any integer $r \geq 3$

$$\frac{M_{r,j}(a)}{[M_{2,j}(a)]^{r/2}} = O(1) \quad \text{and} \quad \frac{M_r(y)}{(M_2(y))^{r/2}} = O(1)$$

where $M_{r,j}(a) = \frac{1}{T} \sum_{i=1}^{T} (a_{ji} - \bar{a}_j)^r$ and $\bar{a}_j = \frac{1}{T} \sum_{i=1}^{T} a_{ji}$

and $M_r(y) = \frac{1}{T} \sum_{i=1}^{T} (y_i - \bar{y})^r$ and $\bar{y} = \frac{1}{T} \sum_{i=1}^{T} y_j$.

Then $L_j = \sum_{i=1}^{T} a_{ji} y_i$ has a normal asymptotic distribution, N(0,1) for $j = 1 \ldots n$.

Proof

Wilks (1962) Theorem 9.5.1. #

Theorem 4

Let $x_1^{(T)} \ldots x_n^{(T)}$ be a n-dimensional stochastic vector process converging in probability on $(x_1 \ldots x_n)$ and $g(x_1 \ldots x_n)$ be a continuous function of $(x_1 \ldots x_n)$ in E_n. Then

$$\lim_{T\to\infty} p(g(x_1^{(T)} \ldots x_n^{(T)}) = p(g(x_1 \ldots x_n))$$

where $p(\cdot)$ is an n-variate probability density function.

Proof:

Wilks (1962) Theorem 4.3.6.#

Traditional test statistics for finite T fall into three groups. Those based on the t distribution are testing for hypotheses on a particular parameter or any linear combination of parameters; those based on an analysis of variance (F distribution tests) for hypotheses on the significance of a regression or a subset of explanatory variables; and those based on the χ^2 distribution for hypotheses on misspecification and validity of restrictions imposed on the estimates.

Theorem 5:

For the model (2.4) estimated by any version of least squares or instrumental variable estimators the asymptotic distribution of the traditional test statistics is as follows:

(i) $\dfrac{\hat{\delta}_{ij} - \delta_{ij}}{\sigma(\hat{\delta}_{ij})} \overset{a}{\sim} N(0,1)$ or $\dfrac{c'(\hat{\delta} - \delta)}{(c'(\Sigma_{\hat{\delta}})c)^{\frac{1}{2}}} \overset{a}{\sim} N(0,1)$

$j = 1 \ldots r_i+s_i$; $i = 1 \ldots M$.

(ii) $$\frac{\hat{\delta}_i' X_i' X_i \hat{\delta}_i / (r_i + s_i)}{\hat{u}_i' \hat{u}_i / (T - r_i - s_i)} \overset{a.}{\sim} \chi^2_{(r_i+s_i)}$$

$i = 1 \ldots M$

(substitute k for $r_i + s_i$ and ammend the numerator for a k member subset of X_i)

(iii) $$[\hat{u}'_i \hat{u}_i - \hat{\hat{u}}'_i \hat{\hat{u}}_i] \overset{a.}{\sim} \chi^2_{(k)} \qquad \text{some } i = 1 \ldots M$$

with the following notation:

"^" denotes the estimated value;

c is a $\sum_{i=1}^{M} (r_i + s_i)$ - column vector creating a linear combination of $\delta = \text{vec}(\delta_i)$ elements;

$\hat{\Sigma}_{\hat{\delta}}$ is the covariance matrix of estimates;

"$\hat{}$" denotes an alternative set of estimates from another specification or restricted estimates;

k = number of restrictions (or specification changes).

Proof:

By assumption $|x_{ti}| < \chi^*$; for least squares estimation

$$\hat{\delta}_i = [X_{i(\alpha)}' X_{i(\alpha)}]^{-1} X_{i(\alpha)}' y_i$$

where $X_{i(\alpha)} \neq 0$ denotes the explanatory variables have possibly been transformed; for instrumental variable estimation $\hat{\delta}_i = (N_i' X_i)^{-1} N_i' y_i$. We require plim $T^{-1} N_i' X_i < \infty$ so $|n_{ti}| < \chi^*$ is implied. Let $A = (X_{i(\alpha)}' X_{i(\alpha)})^{-1} X_{i(\alpha)}'$ or $A = (N_i' X_i)^{-1} N_i'$ for the two estimators.

Part (i) then follows directly from Theorem 3 since the bounds χ^* imply finite moments. The order requirements on A and y

clearly hold. The second formula then follows from the first as the sum of independent normals (e.g. Johnston (1972) p.153).

Part (ii) follows the direct application of Theorem 3 and then 4, noting $\hat{u}_i = [I - X_i A] u_i$, (see Wilks (1962) Theorem 9.3.2) making numerator and denominator χ^2 variates. They are independent quadratic forms for all T (by Cochrane's theorem; Hogg & Craig (1970) p.393) since $y_i'y_i = \hat{\delta}'_i X_i' X_i \hat{\delta}_i + \hat{u}_i'\hat{u}_i$. Further $\lim_{n\to\infty} F_{m,n}(u) = \chi^2_{(m)}$ since

$$F_{m,n}(q) = \frac{c\, q^{m/n - 1}}{(1 + m/n\, q)^{\frac{n+m}{2}}} \quad \text{and} \quad \lim_{n\to\infty} (1 + m/n\, q)^{\frac{n+m}{2}} = e^{mq}.$$

Part (iii) follows by Theorems 3 and 4 and the distribution of two independent χ^2 variates (a special case of Theorem 9B(v) Parzen (1960)).#

Corollary 6:

Often in sumultaneous models the test in (ii) above is performed as $\dfrac{y_i'y_i/T}{\hat{v}_i'\hat{v}_i/(T-r_i-s_i)} \sim F(T, T-r_i-s_i)$ exactly, if $U \sim N(0, \Sigma \otimes I)$ is assumed. This is a possible improvement to the tests of Theorem 5 which is available to 2SLS, 3SLS only. We may apply the standard results of the normality of y_i and $\hat{v}_i$ and independence of their quadratic forms from the unrestricted reduced forms. Any other form of that test, or that test from any other iterative estimator involves the distribution of $\hat{\delta}_i$ which is known only asymptotically.#

Comment:

(a) No further improvements in the tests can be made even if normally distributed structural errors are assumed since asymptotic results only are available for the distribution of $\hat{\delta}_i$.

(b) Part (iii) could be rewritten as a ratio if desired.

Let model (2.1) be written $y_{ti} = f_i(X_{ti}, \delta_i) + u_{ti}$.

Theorem 6:

Theorem 5 holds also for model (2.1) where $\hat{\delta}_i$ denotes the minimum distance estimate and $X_i\hat{\delta}_i$ is replaced by $f_i(X_{ti}, \hat{\delta}_i)$ provided $\hat{\delta}_i$ is consistent and

(a) The distance function $G(S_T,\delta) = \sum_{t=1}^{T} (y_{ti}-f_i(X_{ti},\delta_i))'S_T(y_{ti}-f_i\ \delta_i)$ implies $\lim_{T\to\infty} S_T = S$ a positive definite finite matrix.

(b) $\exists$ a neighbourhood of $\delta^{(0)}$ the true values, in the parameter space.

Proof:

Define a linear pseudo-model;

$$y_{ti} - f_i(X_{ti}, \delta^{(0)}) = \sum_{k=1}^{r_i+s_i} X_{tik}(\delta_{ik} - \delta_{ik}^{(0)}) = u_{ti}$$ and then by

Theorem 3 the vector δ^* which minimises

$$\sum_{t=1}^{T} \{[y_{ti}-f_i(X_{ti},\delta_i^{(0)})]-X_{ti}(\delta_i-\delta_i^{(0)}\}'S_T\{[y_{ti}-f_i(X_{ti},\delta_i^{(0)})]$$

$$-X_{ti}(\delta_i-\delta_i^{(0)})\}$$

is distributed such that $\dfrac{\delta_{ij}^* - \delta_{ij}^{(0)}}{\sigma(\delta_{ij}^*)} \overset{a.}{\sim} N(0,1)\ j = 1 \ldots\ldots r_i + s_i$

But (by an identical argument to Theorem 1 p.334 Malinvaud (1970)) $\sqrt{T}(\hat{\delta}_i - \delta_i^{(0)})$ and $\sqrt{T}(\delta_i^* - \delta_i^{(0)})$ have identical limiting distributions. The remainder of the theorem follows Theorem 5 identically.#

Finally for models (2.1) or (2.6) we have

Theorem 7:

If $p(x_i,\delta_i)$, a probability density function, with regularity conditions[1] summarised in Definition 13 and Lemma 13 of Dhrymes (Chapter 3), then $\sqrt{T}(\hat{\delta} - \delta) \sim N(0 , \Sigma_{\hat{\delta}})$

where $\Sigma_{\hat{\delta}} = -E\left[\frac{\partial^2 \log p(x_i,\delta_i)}{\partial\delta_i \partial\delta_i'}\right]$ and "^" denotes the maximum likelihood estimator.

Proof:

Lemma 3.14 Dhrymes (1970); also Kendall & Stuart (1968) Theorem 18.16.#

Comment:

(a) Theorem 7 is well known. To the extent the estimators under review are asymptotically maximum likelihood ones it confirms Theorems 5 and 6.

(b) Subject to these general regularity conditions the asymptotic distributions are independent of the distribution of U; they are the same for any of the estimators reviewed here; they are the asymptotic versions of the small sample tests.

Remark:

In this context it will be useful to have a precise definition of efficiency for use in the information class hierarchy.

---

1. These conditions are very general but equally detailed, see also Wold (1949).

Definition:

Let $f(\cdot,\delta)$ be any regular density function with parameter set δ (in the sense of Dhrymes (1970) Chapter 3) and $L(\cdot,\delta)$ be its associated log likelihood function. An estimator is said to be efficient if its covariance $\Sigma_{\hat{\delta}}$ equals the Cramer-Rao minimum variance bound, or failing that if no other estimator $\tilde{\delta}$ exists such that $\Sigma_{\hat{\delta}} - \Sigma_{\tilde{\delta}}$ is positive semidefinite. $\Sigma_{\hat{\delta}} - \Sigma_{\tilde{\delta}}$ positive semidefinite implies $\tilde{\delta}$ efficient relative to $\hat{\delta}$.

Lemma 2:

Any unbiased estimator $\hat{\delta}$ implies a Cramer-Rao lower bound of $-TE\left[\frac{\partial^2 L}{\partial\delta\partial\delta'}\right]^{-1}$; any asymptotically unbiased $\hat{\delta}$ implies a bound which is the limit of that as $T \to \infty$. Then we speak of asymptotical efficiency. #

Lemma 3

$\hat{\delta}$, a maximum likelihood estimator, is efficient among all consistent and uniformly asymptotically normal estimators.

Proofs:

Dhrymes (1970) Lemma 3.16 (Lemma 2) and Theorem 3.7 (Lemma 3). #

Estimators appropriate for other information classes conform to these results if they are treated as limited information estimators with data transformed with the sphericality transformation matrix of Chapter 2.4.

A.3: Some Results for Nilpotent Matrices

Lemma 4:

Let M be a general nilpotent matrix of order n. $M^P = 0$ for $p \geq n - 1$.

Proof:

Write M in Jordan canonical form;

$M^P = \theta L^P \theta^{-1}$ where L has $\{L_{k\ell}\} = \left.\begin{array}{ll} 1 & \text{if } \ell = k + 1 \\ 0 & \text{otherwise} \end{array}\right\} \ell,k = 1 \ldots n,$

and θ is the associated matrix of eigenvectors. $L^P = 0$ iff $p \geq n - 1$. #

Lemma 5:

Let A, a matrix of order np, have n^2 submatrices A_{ji} of order p each of which is nilpotent and such that $A_{jj} = 0$, for $i,j = 1 \ldots n$. Then A is nilpotent.

Proof:

$|A_{ji}| = 0$ $i,j = 1 \ldots n$. Eigenvalues λ_k $k = 1 \ldots np$ of A are the solutions to $|A - \lambda I| = 0 = |B|$ say. $|B| = \Sigma \pm b_{1k} b_{2\ell} \ldots b_{np,m}$ (10)

summing over all permutations of the second subscripts (Johnston (1972)). Thus each term in (29) contains exactly one element only from each row and column of B. Hence $|B| = \Sigma \pm |B_{1i}| |B_{2j}| \ldots |B_{n,q}|$ (11)

where $B_{ji} = \left.\begin{array}{ll} A_{ji} & j \neq i \\ A_{jj} - I & \text{otherwise} \end{array}\right\}$ summing over all permutations of the second subscripts. (i.e. we select elements for the terms of (10) in two stages, from within one submatrix using it up entirely and then each submatrix is taken in turn). The only nonzero term in (11) is that with no $|B_{ji}|$ $j \neq i$ so $|B| = \prod_{j=1}^{n} |A_{jj} - \lambda I| = 0$ yielding the characteristic equation of A to be $\lambda^{np} = 0$. #

Lemma 6:

Let A and B be arbitrary square matrices of fixed finite order n where B is nilpotent. Define $J_1 = A + B$ and $J_2 = AB$. J_1 is convergent iff $\rho(A) < 1$ but J_2 is always convergent with $\rho(J_2) = 0$.

Proof:

(a) J_1 is convergent iff $\lim_{S\to\infty} J_1^S = 0$. For $S < n-1$,

$$J_1^S = \sum_{i=0}^{S} \binom{s}{i} A^{s-i} B^i \tag{12}$$

By Lemma 4 above $B^i = 0$ all $i \geq n-1$ so if $S \geq n-1$

$$J_1^S = \sum_{i=0}^{n-1} \binom{s}{i} A^{s-i} B^i \tag{13}$$

and $\lim_{S\to\infty} J_1^S = 0$ iff $\lim_{S\to\infty} A^{s-n+1} = 0$ given n fixed and finite

i.e. iff $\rho(A) < 1$.

(b) By theorem 1 above $\rho(J_2) \leq \min\{\rho(A), \rho(B)\} = 0$ directly every eigenvalue of B is zero. #

Comment:

(a) A first order stationary linear iterative process with iteration matrix J_1 has global convergence criterion $\rho(A) < 1$. However it is false to suggest $\rho(J_1) = \rho(A)$ and an expression for $\rho(J_1^S)$ $S \geq 1$ is unavailable in view of the summations in (12) and (13). All we know is $\lim_{S\to\infty} \rho(J_1^S) = \lim_{S\to\infty} \rho(A^S)$ from (13) and for increasing S $\rho(J_1^S) \to \rho(A^S)$ with an approximation which is arbitrarily good for sufficiently large S. Their difference is unlikely, in general, even to be a monotonically decreasing function of S as eigenvalues are highly sensitive to very small perturbations in the elements of their matrix (Wilkinson (1965)). That difference depends on B and may vary unboundedly from zero at S = 1 (the null matrix).

Corollary

(i) In asymptotic convergence rates $R_\infty(J_1) = R_\infty(A)$.

Proof:

Since $\lim_{S\to\infty} \rho(J_1^S) = \lim_{S\to\infty} \rho(A^S)$ and definitions in Chapter 8. #

Corollary:

(ii) Iteration methods, acceleration parameters, reorderings or renormalisations which are improvements or optimal on the asymptotic convergence rate criterion for A are also for J_1.

Proof:

Corollary (i); or for any norm and $S \geq n-1$

$$||J_1^S|| \leq \sum_{i=0}^{n-1} \binom{s}{i} ||A^{s-i}|| \; ||B^i|| \qquad (14)$$

where $R_\infty(J_1) = \lim_{S\to\infty} \frac{1}{S} \log ||J_1^S||$ so on this criterion improvements or optimisation on A and J_1 are equivalent since n is fixed and finite. #

(b) Since $\rho(J_1)$ is unknown and $\rho(J_1^S)$ varies with $S \geq 1$, we cannot accurately assess the number of iterations to convergence. If B = 0 then $\frac{\log \tau}{\log \rho(A)}$ are required as usual.

At best B can offset A to imply $\rho(J_1) = 0$. We might use the approximation indicated by S* to get the summation of A's in (13) sufficiently small i.e. s* such that $\sum_{i=0}^{n-1} \binom{s}{i} [\rho(A)]^{s*-1} = \tau$. Clearly

$$S* > \frac{\log \tau}{\log \rho(A)} + n - 1 \qquad (15)$$

with or without that approximation.

(c) Because of the inequality in (15), and more importantly because to reduce $||A^{s-i}||$ is not the only way to reduce $||J_1^S||$ for finite S, schemes to improve or optimise average rates of convergence for A do not do so for J_1 (unless B = 0). Any such schemes must be calculated directly for J_1 itself for each S. (15) suggests if n is large and no prior information is available on B semi-iterative acceleration to improve on S* may be a valuable "fail safe" policy.

(d) J_2 is always globally convergent being a zero iteration matrix. It cannot be further improved on average and asymptotic convergence rate criteria.

BIBLIOGRAPHY

Agren, A. (1972) "Extensions of the Fix-point method: theory and applications", Ph.D. Thesis University of Uppsala (Institute of Statistics), Sweden.

Agren, A. (1973) "A short note on the GEID specification and the GEID Estimator" (eds) Lyttkens, E. and Wold H.O.A. (1973)

Agren, A. (1975) "The Consistency of GEID - FP estimates" Journal of Royal Statistical Society (B), 37, pp. 293-5.

Agren, A. and Wold, H.O.A. (1969) "On the Structure and Estimation of Interdependent Systems" in (ed) P.R. Krishnaiah: "Multivariate Analysis - II", Academic Press, New York.

Amemiya, T. (1974) "The Nonlinear Two Stage Least Squares Estimator", Journal of Econometrics, 2, pp.105-110.

Anderson, T.W. (1958) "An Introduction to Multivariate Statistical Analysis", Wiley, New York.

Anderson, T.W. (1959) "On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations", Annals of Mathematical Statistics, 30, pp.676-687.

Anderson, T.W. and Sawa, T (1973) "Distributions of Estimates of Coefficients of a Single Equation in a Simultaneous System and their Asymptotic Expansions", Econometrica, 41,pp.683-715.

Basman, R.L., Richardson,D.H. and Rohr, R.J. (1974a) "An Experimental Study of Structural Estimators and Test Statistics Associated with Dynamic Econometric Models", Econometrica, 42, pp.717-731.

Basman, R.L., Richardson, D.H. and Rohr, R.J. (1974b) "Finite Sample Distributions Associated with Stochastic Difference Equations", Econometrica, 42, pp.825-840.

Bergstrom, R. (1973) "Convergence Properties, Comparative Studies and Applications of the FP and IIV methods" in (eds) Lyttkens, E. and Wold H.O.A. (1973)

Bergstrom, R. (1974) "Studies in the Estimation of Interdependent Systems, especially the Fix-point and Iterative Instrumental Variables Methods". Ph.D. Thesis, University of Uppsala (Institute of Statistics), Sweden.

Berndt, E.K., Hall, B.H., Hall, R.E. and Hausman, J.A. (1974)

"Estimation and Inference in Nonlinear Structural Models", Annals of Economic and Social Measurement, 3, pp.653-667.

Birkholf, G., Varga R.S. and Young, D.M. (1962)

"Alternating Direction Implicit Methods" in Alt, F.L. and Rubinoff (eds) "Advances in Computers", Vol.III, Academic Press, New York.

Blesley, D.A. (1974)

"Estimation of Systems of Simultaneous Equations, and Computation Specifications of Gremlin", Annals of Economic and Social Measurement, 3, pp.551-615.

Bodewig, E. (1956)

"Matrix Calculus", North Holland, Amsterdam.

Bodin, L. (1973)

"Improved FP estimtion of GEID systems by the RFP method" in (eds) Lyttkens,E. and Wold, H.O.A. (1973)

Bodin, L. (1974)

"Recursive Fix-point Estimation: theory and applications", Ph.D. Thesis, University of Uppsala (Institute of Statistics), Sweden.

Brauer, A. (1952)

"Limits for the Characteristic Roots of a Matrix", Duke Mathematical Journal, 19, pp.75-91 and 553-562.

Brauer, A. (1964)

"On the Characteristic Roots of Non-negative Matrices", in Schneider (1964).

Brundy, J.M. and Jorgenson, D.W. (1971)

"Efficient Estimation of Simultaneous Equations by Instrumental Variables", Review of Economics and Statistics, 13, pp. 207-224.

Brundy, J.M. and Jorgenson, D.W. (1974a)

"Consistent and Efficient Estimation of Simultaneous Equations by Means of Instrumental Variables", in P.Zarembka (ed) "Frontiers in Econometrics", Academic Press, New York.

Brundy, J.M. and Jorgenson, D.W. (1974b)

"The Relative Efficiency of Instrumental Variables Estimators of Systems of Simultaneous Equations", Annals of Economic and Social Measurement, 3, pp.679-701.

Businger, P. and Golub, G.H. (1965)

"Handbook Series Linear Algebra: Linear Least Squares Solutions By Housholder Transforms", Numerische Mathematik, 7, pp.269-276.

Chipman, J. (1964) "On Least Squares with Insufficient Observations", Journal of the American Statistical Association, 59, pp.1078-1111.

Chow, G. (1964) "A Comparison of Alternative Estimators for Simultaneous Equations", Econometrica, 32, pp.532-53.

Chow, G. (1968) "Two Methods of Computing FIML Estimates in Simultaneous Equation", International Economic Review, 9, pp.100-112.

Chow, G. (1974) "A Family of Estimators for Simultaneous Equation Systems", International Economic Review, 15, pp.654-666.

Chow, G. (1975) "Analysis and Control of Dynamic Economic Systems", Wiley, Interscience, New York.

Christ, C.F. (1966) "Econometric Models and Methods", Wiley, New York.

Coen, P.J., Gomme, E. and Kendal, M.G. (1969) "Lagged Relationships in Economic Forecasting", Journal of Royal Statistical Society (A), 132, pp.133-163.

Cragg, J.G. (1966) "On the Sensitivity of Simultaneous Equation Estimators to the Stochastic Assumption of the Models", Journal of the American Statistical Association, 61, pp. 136-151.

Cragg, J.G. (1967) "On the Relative Small-Sample Properties of Several Structural Equation Estimators", Econometrica, 35, 89-110.

Cragg, J.G. (1968) "Some Effects of Incorrect Specification on the Small Sample Properties of Several Simultaneous Equation Estimators", International Economic Review, 9, pp. 63-86.

Daniel, J.W. (1970) "A Correction Concerning the Convergence Rate for the Conjugate Gradient Method", SIAM Journal of Numerical Analysis, 7, pp.277-9.

Dhrymes, P.J. (1970) "Econometrics: Statistical Foundations and Applications", Harper and Row, New York.

Dhrymes, P.J. (1971) "A Simplified Structural Estimator for Large Scale Econometric Models", Australian Journal of Statistics, 13, pp. 168-175.

Dhrymes, P.J. (1973) "Small Sample and Asymptotic Relations Between Maximum Likelihood and Three Stage Least Squares Estimators". Econometrica, 41, pp.357-364.

Dhrymes, P.J., Berner, R. and Cummins, D. (1974) "Estimators for Dynamic Simultaneous Equation Models with Autocorrelated Errors", Econometrica, 42, pp.311-332.

Dhrymes, P.J. and Pandit, V. (1972) "Asymptotic Properties of an Iterate of the Two Stage Least Squares Estimator", Journal of American Statistical Association, 67, pp.444-448.

Dixon, L.C.W. (1973) "Conjugate Directions without Linear Searches", Journal of the Institute of Mathematics and its Applications, 11, pp.317-329.

Dixon, L.C.W. (1975) "Conjugate Gradient Algorithms: Quadratic Termination without Linear Searches", Journal of the Institute of Mathematics and its Applications, 15, pp. 9-19.

Durbin, J. (1960) "Estimation of Parameters in Time Series Regression Models", Journal of Royal Statistical Society (B), 22, pp.139-153.

Durbin, J. (1963) "Maximum Likelihood Estimation of Parameters of a System of Regression Equations", Copenhagen Meeting of the Econometric Society, August 1963.

Dusenberry, J.S., Fromm, G., Klein, L.R. and Kuh, E.(1965) "The Brookings Quarterly Econometric Model of the United States", North Holland, Amsterdam. (Vol.I)

Dusenberry, J.S., Fromm, G. Klein,L.R. and Kuh, E.(1968) "The Brookings Quarterly Econometric Model of the United States, Volume II: some further results", North Holland, Amsterdam.

Dutta, M. and Su, V.(1969) "An Econometric Model of Puerto Rico", Review of Economic Studies, 36, pp.319-333.

Edgerton, D. (1973) "Nonlinear ID Systems and Other Problems", in (eds) Lyttkens, E. and Wold, H.O.A. (1973).

Edgerton, D. (1974) "Nonlinear Interdependent Systems", Ph.D. Thesis, University of Uppsala (Institute of Statistics), Sweden.

Eisenpress, H. & Greenstadt, J. (1966) "The Estimation of Nonlinear Econometric Systems", Econometrica, 34, 851-861.

Evans, M.K. (1969) "Macroeconomic Activity", Harper & Row, New York.

Evans, M.K. and Klein, L.R. (1967) — "The Wharton Econometric Forecasting Model", University of Pennsylvania, Philadelphia.

Farebrother, R.W. (1974a) — "The Graph of a k-class Estimator", Review of Economic Studies, 41, pp.533-538.

Farebrother, R.W. (1974b) — "Recursive Relations for the Samuelson Transformation Coefficients", International Economic Review, 15, pp.805-807.

Fisher, F.M. (1965) — "Dynamic Structure and Estimation in Economy Wide Econometric Models", in Dusenberry, J.S., Fromm,G. Klein, L.R., and Kuh, E. (1965).

Fisher, F.M. (1966) — "The Identification Problem in Econometrics", McGraw-Hill, New York.

Fisher, F.M. & Mitchell, B.M. (1970) — "The Choice of Instrumental Variables", International Economic Review,

Fisk, P. (1973) — Comments in Lyttkens (1973)

Fletcher and Jackson (1974) — "Minimisation of a Quadratic Function of Many Variables Subject Only to Lower and Upper Bounds", Journal of the Institute of Mathematics and its Applications, 14, pp. 159-74.

Fromm, G. and Klein, L.R. (1968) — "Solutions of the Complete System", in Dusenberry, J.S., Fromm,G., Klein, L.R. and Kuh, E. (1968).

Gantmacher, F.R. (1959) — "Applications of the theory of Matrices", Interscience Publishing Co., New York.

Girschick, M.A. and Haavelmo, (1947) — "Statistical Analysis of the Demand for Food: Examples of Simultaneous Estimation of Structural Equations", Econometrica, 15, pp.79-110.

Goldberger, A.S. and Duncan, O.D. (eds) (1973) — "Structural Equation Models in the Social Sciences", Seminar Press, New York.

Goldfield, S.M. and Quandt, R.E. (1972) — "Nonlinear Methods in Econometrics", North Holland, Amsterdam.

Golub, G.H. (1959) — "The Use of Chebychev Matrix Polynomials in the Iterative Solution of Linear Equations Compared with the Method of Successive Over-relaxation", Ph.D. Thesis, University of Illinois, U.S.A.

Golub, G.H. and Varga, R.S. (1961/2) — "Chebychev Semi-iterative methods, Successive Over-relaxation Methods and Second Order Richardson Iterative Methods", Numerische Mathematik, 3, pp.147-156 and 3, pp.157-168.

Grilliches, Z. and Rao, P. (1969) "Small Sample Properties of Several Two Stage Regression Methods in the Context of Autocorrelated Errors", Journal of the American Statistical Association, 64, pp.253-272.

Hausman, J.A. (1974) "Full Information Instrumental Variables Estimation of Simultaneous Equation Systems", Annals of Economic and Social Measurement, 3, pp.641-653.

Hausman, J.A. (1975) "An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Nonlinear Econometric Models", Econometrica, 43, pp.727-739.

Hendry, D.F. (1972) "Manual for GIVE", London School of Economics, Mimeo.

Hendry, D.F. (1973) Comments on Lyttkens (1973).

Hendry, D.F. (1974) "Stochastic Specification in an Aggregate Demand Model of the United Kingdom", Econometrica, 42, pp.923-37.

Hendry, D.F. (1976) "The Structure of Simultaneous Equation Estimators", Journal of Econometrics, 4, pp.51-89.

Hogg, R.V. and Craig, A.T. (1970). "Introduction to Mathematical Statistics", (3rd Edition) Collier-Macmillan, London.

Hughes Hallett, A. (1973a) Comments on Lyttkens (1973).

Hughes Hallett, A. (1973b) "Multicollinearity as a problem in Misspecification", Unpublished paper, Oxford University.

Johnston, J. (1972) "Econometric Methods" (2nd Edition) McGraw-Hill, Tokyo.

Jorgenson, D.W. and Laffont, J-J. (1974) "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances", Annals of Economic and Social Measurement, 3, pp.615-641.

Kadane, J. (1971) "Comparison of k-class Estimators when the Disturbances are Small", Econometrica, 39, pp.723-739.

Kelejian, H.H. (1970) "Identification of Nonlinear Systems: an Interpretation of Fisher", Princeton University, Econometric Research Programme, Research Paper No. 22(Revised).

Kelijian, H.H. (1971) "Two Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in Endogenous Variables", Journal of the American Statistical Association, 66, pp.373-4.

Kendall, M.G. and Stuart, A. (1966) "The Advanced Theory of Statistics", Vols. I-III, Charles Griffen & Co., London.

Klein, L.R. (1950) "Economic Fluctuations in the United States 1921-41", Wiley, New York.

Klein, L.R. (1969) "Estimation of Interdependent Systems in Macroeconometrics", Econometrica, 37, pp.171-192.

Kleok, T. and Mennes, L.B.M. (1960). "Simultaneous Equation Estimation Based on Principal Components of Predetermined Variables", Econometrica, 28, pp.45-61.

Leontieff, W. (1947a) "Introduction to a Theory of the Internal Structure of Functional Relationships", Econometrica, 15, pp.361-373.

Leontieff, W. (1947b) "A Note on the Interrelations of Subsets of Independent Variables of a Continuous Function with Continuous First Derivatives" Bulletin of American Mathematical Society, 55, pp.343-350.

Lyttkens, E. (1966) "A Modification of the Symmetric Iterative Method for Estimating Interdependent Systems", Seminar Paper, Institute of Statistics, University of Uppsala,Sweden.

Lyttkens, E. (1968) "The Fix-point Method with Iterated Reduced Form Coefficients and the Fix-point Method with Pseudovariables", Seminar Paper, Institute of Statistics, University of Uppsala, Sweden.

Lyttkens, E. (1970) Chapter 3.8 in Mosbaek and Wold (1970).

Lyttkens, E. (1971) "The Iterative Instrumental Variables Method and the Full Information Maximum Likelihood for Estimating Interdependent Systems", Barcelona Meeting of the Econometric Society, September 1971.

Lyttkens, E. (1973) "The Fix-point Method for Estimating Interdependent Systems with Underlying Model Specification", Journal of Royal Statistical Society (A), 136, pp.353-394, - including Discussion.

Lyttkens, E. and Wold, H.O.A. (1973) (eds) "FP(fix-point), IIV(Iterative Instrumental Variables) and related approaches to ID (Interdependent Systems): Theory and Applications", Oslo Meeting of the Econometric Society, August 1973.

McCarthy, M.D. (1972) "A Note on the Forecasting Properties of 2SLS Restricted Reduced Forms", International Economic Review,

Madansky, A. (1964) "On the Efficiency of Three-Stage Least Squares Estimation", Econometrica, 32, pp.51-56.

Maddala, G.S. (1971) "Simultaneous Estimation Methods for Large and Medium Size Econometric Models", Review of Economic Studies, 38, pp.435-45.

Maddala, G.S. (1974) "Some Small Sample Evidence on Tests of Significance in Simultaneous Equation Models", Econometrica, 42, pp.841-52.

Malinvand, E. (1970) "Statistical Methods of Econometrics", (2nd Edition) North-Holland, Amsterdam.

Mann, H.B. and Wald,A.(1943) "On the Statistical Treatment of Linear Stochastic Difference Equations", Econometric, 11, pp.173-220.

Marcus, M. and Minc, H. (1964) "A Survey of Matrix Theory and Matrix Inequalities", Allyn and Bacon, Boston.

Mariano, R.S. (1972) "The Existence of Moments of the Ordinary Least Squares and Two Stage Least Squares Estimators", Econometrica, 40,pp.643-52.

Mariano, R.S. (1973) "Approximations to the Distribution Functions of Theils k-class Estimations", Econometrica, 41, pp.715-23.

Mariano, R.S. and Sawa, T. (1972) "The Exact Finite Sample Distribution of the Limited Information Maximum Likelihood Estimator in the Case of Two Included Endogenous Variables", Journal of the American Statistical Association, 67, pp.159-163.

Marquardt, D.W. (1970) "Generalised Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation", Technometrics, 12,pp 591-612.

Meyer, J.R. and Kuh, E. (1957) "How Extraneous are Extraneous Estimates?", Review of Economics and Statistics, 39, pp.380-93.

Meyer-Spasche, R. (1972) "A Constructive Method for Solving the Liapounov Equation for Complex Matrices", Numerische Mathematik, 19, pp.433-38.

Meyer-Spasche, R. (1973) "A Method of Solving the Stability Problem for Complex Matrices", Numerische Mathematik, 20, pp.364-71.

Mitchell, B.M. (1970) "Estimation of Large Econometric Models", Ph.D. Thesis, Massachusetts Institute of Technology.

Mood, A.M., Graybill, F. and Boes, D.C. (1974) "Introduction to the Theory of Statistics", McGraw-Hill, Tokyo.

Morishima, M. (1964) "Equilibrium, Stability and Growth", Oxford University Press, London.

Mosbaek, E. and Wold, H.O.A. (1970) "Interdependent Systems: Structure and Estimation", North-Holland, Amsterdam.

Nagar, A.L. (1959) "The Bias and Moment Matrix of the General k-class Estimators of the Parameters in Simultaneous Equations", Econometrica, 27, pp.575-95.

Nagar, A.L. (1960) "A Monte Carlo Study of Alternative Simultaneous Equation Estimators", Econometrica, 28, pp.428-442.

Nagar, A.L. (1962) "Double k-class Estimators of Parameters in Simultaneous Equations and their Small Sample Properties", International Economic Review, 3, pp.168-88.

Nagar, A.L. (1968) "Stochastic Simulation of the Brookings Econometric Model" in (eds) Dusenberry, J.S., Fromm, G., Klein, L.R., Kuh, E. (1968).

Nerlove, M. (1960) "The Market Demand for Durable Goods: A Comment", Econometrica, 28, pp.132-42.

Oberhofer,W. and Kmenta, J. (1974) "A General Procedure of Obtaining Maximum Likelihood Estimates in Generalised Regression Models", Econometrica, 42, pp.579-90.

Oi, W.Y. (1969) "On the Relationship Among Different Members of the k-class", International Economic Review, 10, pp. 36-46.

Ortega, J.M. and Rheinboldt, W.C. (1970) "Iterative Solution of Nonlinear Equations in Several Variables", Academic Press, New York.

Ostrowski, A. (1960) "Solution of Equations and Systems of Equations", Academic Press, New York.

Ostrowski, A. (1964) "Positive Matrices and Functional Analysis" in Schneider (1964).

Parks, R.W. (1967) "Efficient Estimation of Regression Equations when Disturbances are both Serially and Contemporaneously Correlated", Journal of the American Statistical Association, 62, pp.500-9.

Parzen, E. (1960) "Modern Probability Theory and Its Applications", Wiley, New York.

Peters, G. and Wilkinson, J.H. (1970) "Handbook Series Linear Algebra: Eigenvectors of Real and Complex Matrices by Q.R. and L.R. Transformations", Numerische Mathematik, 16, pp.181-204.

Powell, M.J.D., (1971) "On the Covergence of the Variable Metric Algorithm", Journal of the Institute of Mathematics and its Applications, 7, pp. 21-36.

Powell, M.J.D. (1972) "Quadratic Termination Properties of Minimisation Algorithms: I Statement and Discussion of Results, II Proofs of Theorems", Journal of the Institute of Mathematics and its Applications, 10, pp. 333-58.

Press, S.J. (1972) "Applied Multivariate Analysis", Holt, Rinehart and Winston Inc., New York.

Quandt, R. (1965) "Small Sample Properties of k-class Estimators", International Economic Review, 6, pp.92-104.

Radner, R. (1958) "Minimax Estimation for Linear Regressions", Annals of Mathematical Statistics, 29, pp.1244-51.

Rao, C.R. (1973) "Linear Statistical Inference and its Applications", (2nd Edition) Wiley, New York

Rao, C.R. and Mitra, S.K. (1971) "Generalised Inverse of Matrices and its Applications", Wiley, New York.

Renton, G.A. (ed) (1975) "Modelling the Economy", Heinemann, London.

Richardson, D.H. and Rohr, R.J. (1971) "Distribution of a Structural and Statistic for the Case of Two Included Endogenous Variables", Journal of the American Statistical Association, 66, pp. 375-82.

Richardson, D.W. and Wu,D.M. (1971) "A Note on the Comaprison of Ordinary and Two Stage Least Squares Estimators", Econometrica, 39, pp.973-83.

Rothenburg, T.J. and Leenders, C.T. (1964). "Efficient Estimation of Simultaneous Equation Systems", Econometrica, 32, pp. 57-76.

Salmon, M.H. (1973) Comments in Lyttkens (1973).

Salmon, M.H. and Eaton, J. (1975) "Estimation Problems in Large Econometric Models: An Application of Various Estimation Techniques to the London Business School Model", in G.A.Renton, (1975).

Samuelson, P.A. (1941) "Conditions that the Roots of a Polynomial be Less than Unity in Absolute Value": Annals of Mathematical Statistics, XII, pp.360-4.

Sargan, J.D. (1958) "The Estimation of Economic Relationships using Instrumental Variables", Econometrica, 26, pp. 393-415.

Sargan, J.D. (1964) "Wages and Prices in the United Kingdom: a Study in Econometric Methodology", in Hart, P.E., Mills, G and Whitaker,J.K. (eds), "Econometric Analysis for National Economic Planning", Colston Papers No.16, Butterworths, London.

Sargan, J.D. (1970) "The Finite Sample Distribution of FIML Estimators", 2nd World Congress of the Econometric Society, Cambridge, England. September 1970.

Sargan, J.D. (1973) "The Tails of the FIML Estimates of the Reduced Form Coefficients", Unpublished Paper, Mimeo, London School of Economics.

Sargan, J.D. (1974) "The Validity of Nagar's Expansion for the Moments of Econometric Estimators", Econometrica, 42, pp.169-77.

Sargan, J.D. (1975a) "Asymptotic Theory and Large Models", International Economic Review, 16,pp.75-92.

Sargan, J.D. (1975b) "Gramm-Charlier Approximations Applied to t Ratios of k-class Estimators", Econometrica, 42, pp.327-46.

Sargan, J.D. and Mikail, W. (1971) "A General Approximation to the Distribution of Instrumental VAriables Estimates", Econometrica, 39, 131-169.

Savin, N.E. (1973) "Systems k-class Estimators", Econometrica, 41, pp.1225-36.

Sawa, T. (1969). "The Exact Sampling Distribution of Ordinary Least Squares and Two Stage Least Squares Estimators", Journal of American Statistical Association, 64, pp.923-37.

Sawa,T. (1972) "Finite Sample Properties of the k-class Estimators", Econometrica, 40, pp.653-80.

Scharf, W. (1976) "K-matrix-class Estimators and the Full Information Maximum Likelihood Estimators as a Special Case", Journal of Econometrics, 4, pp. 41-51.

Schneider, H. (1964) "Recent Advances in Matrix Theory", Mathematics Research Centre Publication No. 12, University of Wisconsin Press, University of Wisconsin.

Stein, P. (1952) "Some General Theorems on Iterants", Journal Reserve National Bureau of Standards, 48, pp. 82-3.

Steward, D.V. (1962) "On an Approach to Techniques for the Analysis of the Structure of Large Systems", SIAM Review, 4, pp.321-42.

Suits, D.B. (1962) "Forecasting and Analysis with an Econometric Model", American Economic Review, pp.104.32.

Summers, R. (1965) "A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators", Econometrica, 33, pp.1-41.

Swamy, P.A.V.B. and Holmes, J. (1971) "The Use of Undersized Samples in the Estimation of Simultaneous Equation Systems". Econometrica, 39, pp.455-8.

Taussky, O. (1964a) "One the Variation of the Characteristic Roots of a Finite Matrix Under Various Changes in the Elements", in Schneider (1964).

Taussky, O. (1964b) "Matrices C with $C^n \to 0$", Journal of Algebra, 1, pp. 5-10.

Theil, H. (1971) "Principles of Econometrics", Wiley, New York.

Theil, H. (1973) "A Simple Modification to the Two Stage Least Squares Procedure for Undersized Samples" In Goldberger & Duncan (1973).

Theil, H. and Goldberger, A.S. (1961) "On Pure and Mixed Statistical Estimation in Economics", International Economic Review, 2, pp.65-78.

Telser, L.G. (1964) "Iterative Estimation of a Set of Linear Regression Equations", Journal of the American Statistical Association, 59, pp. 845-62.

Tinbergen, J. (1952) "On the Theory of Economic Policy", North-Holland, Amsterdam.

Varga, R.S. (1962) "Matrix Iterative Analysis", Prentice-Hall, Englewood Cliffs, N.J.

Verdoon, R.J. and Post,J.J. (1964) "Capacity and Short Term Multipliers", in Hart, P.E., Mills, G. and Whitaker J.K. (eds), "Econometric Analysis for National Economic Planning", Colston Papers No. 16, Butterworths, London.

Wald, A. (1949) "Note on the Consistency of the Maximum Likelihood Estimator", Annals of Mathematical Statistics, 20, pp.595-601.

Walsh, G.R. (1975) "Methods of Optimisation", Wiley, New York.

Westlake, J.R. (1968) "A Handbook of Numerical Matrix Inversion and Solution of Linear Equations", Wiley, New York.

Wilkinson, J.H. (1965) "The Algebraic Eigenvalue Problem", Oxford University Press, London.

Wilks, .S. (1962) "Mathematical Statistics", Wiley, New York.

Wold, H.O.A. (1938) "A Study in the Analysis of Stationary Time Series", Almquist and Wiksell, Stockholm.

Wold, H.O.A. (1963) "A Theorem of the Consistency of Least Squares Regression", Sankya 25A, 2, pp. 211-15.

Wold, H.O.A. (1965) "A Fix-point Theorem with Econometric Background, I-II", Arkiv für Mathematik, 6, pp.209-40.

Wold, H.O.A. (1974) "Causal Flows with Latent Variables: Partings of the Ways in the Light of NIPALS Modelling", European Economic Review, 5, pp. 67-86.

Young, D.M. (1971) "Iterative Solution of Large Linear Systems", Academic Press, New York.

Zangwill, W. (1967) "Minimising a Function Without Calculating Derivatives", Computer Journal, 10, pp. 293-6.

Zellner, A. (1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", Journal of the American Statistical Association, 57, pp. 348-68.