Econometric Issues in Forward-Looking Monetary Models

Sophocles Mavroeidis

Nuffield College, Oxford

Trinity Term 2002

Econometric Issues in Forward-Looking Monetary Models

SOPHOCLES MAVROEIDIS

Nuffield College

Thesis submitted for the Degree of Doctor of Philosophy at the Department of Economics,
University of Oxford.
Trinity 2002.

Abstract

Recently, single equation approaches for estimating structural models have become popular in the monetary economics literature. In particular, single-equation Generalized Method Moments estimators have been used for estimating forward-looking models with rational expectations. Two important examples are found in Clarida, Galí, and Gertler (1998) for the estimation of forward-looking Taylor rules and in Galí and Gertler (1999) for the estimation of a forward-looking model for inflation dynamics. In this thesis, we address the issues of identification which have been overlooked due to the incompleteness of the single-equation formulations. We provide extensions to existing results on the properties of GMM estimators and inference under weak identification, pertaining to situations in which only functions of the parameters of interest are identified, and structural residuals exhibit negative autocorrelation. We also characterize the power of the Hansen test to detect mis-specification, and address the issues arising from using too many irrelevant instruments as well as from general corrections for residual autocorrelation, beyond what is implied by the maintained model.

In general, we show that the non-modelled variables cannot be weakly exogenous for the parameters of interest, and that they are informative about the identification and mis-specification of the model. Modelling the reduced form helps identify pathological situations in which the structural parameters are weakly identified and the GMM estimators are inconsistent and biased in the direction of OLS. We also find the OLS bias to be increasing in the number of over-identifying instruments, even when the latter are irrelevant, thus demonstrating the dangers of using too many potentially irrelevant instruments. Finally, with regards to the "New Phillips curve", we conclude that, for the US economy, this model is either un-identified or mis-specified, casting doubts on its utility as a model of inflation dynamics.

Contents

List of Tables

List of Figures

Acknowledgements

I am indebted to David Hendry and Adrian Pagan, whose sharp and stimulating supervision has proven invaluable for the writing of this thesis. Their insightful comments and criticism on earlier drafts, as well as their rigorous scrutiny of my analysis, have been most helpful in shaping my thinking and, therefore, I owe most of this work to them.

Several other people have contributed directly or indirectly to the writing of this thesis. I have benefited greatly from stimulating discussions with my fellow students. In particular, I am grateful to Guillaume Chevillon and George Konaris with whom I often shared my thoughts. They listened attentively, and gave me invaluable comments and suggestions. I would also like to thank Kenny Bhanthumnavin, Michael Massmann, Manos Venardos, Jack Wong, Stefan de Wachter, among others, for their feedback at various points during this project.

In the early stages, I benefited from illuminating discussions with my college supervisor Richard Spady. I also thank Bent Nielsen who often helped me on mathematical issues, Jurgen Doornik and Hans-Martin Krolzig for their help on programming, as well as Steve Bond, John Muellbauer, Neil Shephard and various other seminar or conference participants for their useful feedback during presentations of parts of this thesis.

Financial support from Nuffield College (funded studentship), the UK Economic and Social Research Council (doctoral award No: R00429924298) and the Department of Economics at the University of Oxford, is gratefully acknowledged.

I would also like to thank the secretaries at Nuffield College, Maureen Baker and Glynis Baleham, for their help with my sometimes unreasonable requests throughout this period.

Last but not least, I would like to thank my family and my friends for their patience and support during this very stressful period of my life. I am especially grateful to my parents and to Menia, whose love and encouragement has proven invaluable for the writing of this thesis. Without them this project would have never been completed.

Chapter 1

Introduction

Forward-looking models are commonly used in monetary economics both by academics and practitioners, in order to advise on, or assess the efficacy of, monetary policy. In recent years, small-scale forward-looking macro models have become influential in central banks around the world, especially so relative to the traditional large-scale macro models of the seventies, thus becoming important for monetary policy making.

There are both theoretical and practical reasons for this growing popularity. On theoretical grounds, first, by explicitly incorporating forward-looking components, these models address the Lucas (1976) critique, which reduced-form models do not. Second, because they are usually built on micro-foundations, it is argued that they represent underlying economic structure. Moreover, their so-called 'structural' parameters admit interesting economic interpretations, see the discussion in section 4.1, and thus they are more appealing than the reduced-form models. Third, these models are based on 'rational expectations', which have become an essential feature of most macroeconomic models.

There is a widespread view that economic agents are forward looking in the setting of monetary policy as well as in forming expectations about future inflation, as seen by the following quotes by prominent central bankers:

> "The challenge of monetary policy is to interpret current data on the economy [...] with an eye to anticipating future inflationary forces and to countering them by taking action in advance." Alan Greenspan, Chairman of the Federal Reserve Board, *Humphrey-Hawkins testimony*, 1994.

> "[A]fter just a couple of years of [inflation] targeting, [...] expectations over a 2-year horizon [...] tended to be affected little by what was happening to current inflation rates. This was in marked contrast to earlier periods in Canadian history, in which expectations for the future had been fairly tightly linked to recently observed inflation rates." David Dodge, Governor of the Bank of Canada, *Speech at the AEA annual meeting*, Atlanta 2002.

This prompted researchers to develop models of the form

$$\begin{aligned} \text{Pure:} \quad y_t &= \beta \, \mathrm{E}(y_{t+1}|\mathcal{F}_t) + e_t \\ \text{Hybrid:} \quad y_t &= \beta \, \mathrm{E}(y_{t+1}|\mathcal{F}_t) + \gamma \, y_{t-1} + e_t \end{aligned}$$

The former is a pure forward-looking model, whereas the latter is a hybrid version containing both forward and backward-looking adjustment. These models have been used to address the following questions that are central to the current monetary policy debate: (i) Are agents forward-looking? (Are expectations rational?) (ii) How important is forward-looking behaviour compared to "backwardness"? In what follows, we will show that these models are not suitable to address those questions.

1.1 Motivation

Although some of the results given here have wider applicability, our motivation comes from the monetary economics literature, where single equation approaches for estimating structural models have recently become popular. In particular, single-equation GMM methods have been used for estimating forward-looking models with rational expectations.

One example comes from the recent literature on monetary policy rules, where it has become common practice to estimate Taylor-type rules from historical data, see Taylor (1999) and the papers therein. One approach, popularized by Clarida, Galí, and Gertler (1998), is the estimation of the reaction function parameters from a single equation of the form:

$$r_t = \bar{r} + \alpha \left(\mathrm{E}(\pi_{t+j}|\mathcal{F}_t) - \bar{\pi} \right) + \beta \, \mathrm{E}(y_{t+i}|\mathcal{F}_t) + v_t \tag{1.1}$$

where r, π and y denote the real interest rate, inflation and output gap respectively, $\mathrm{E}(\cdot|\mathcal{F}_t)$ denotes expectations conditional on the available information, and i, j are specified.

Another important example is the influential paper of Galí and Gertler (1999), which uses the same econometric methodology in estimating the "New Phillips curve", a forward-looking model for inflation dynamics:

$$\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} \tag{1.2}$$

where s_t is a measure of unit labour costs. Other examples of forward-looking Phillips curves include the models proposed in Buiter and Jewitt (1985), Fuhrer and Moore (1995), Batini, Jackson, and Nickell (2000) and Galí, Gertler, and López-Salido (2001).

In view of the fact that such equations involve unobservable expectations of variables, researchers proceed as follows. They replace expectations by actual realizations of the variables and derive orthogonality conditions that may be used to estimate the parameters of the model with the Generalized Method of Moments (GMM). These moment conditions are derived based on the assumption of rational (model-consistent) expectations, i.e., that the expectation-induced 'errors in variables' must be orthogonal to the information set available to the agents, denoted $\mathcal{F}_t$ at the time the expectations are formed. The nature of the model (mainly the forward-looking horizons i, j in (1.1), say) and the assumptions about the properties of the 'error' v_t in the reaction function guide the choice of the GMM estimator, i.e., whether corrections should be made for the presence of serial correlation or heteroscedasticity of the residuals. Finally, as a diagnostic check of the 'validity' of this approach, researchers rely on the Hansen-Sargan test of over-identifying restrictions, to verify that the choice of instruments was correct. This is also seen as an indirect test of the rational expectations formulation, i.e., the extent to which expectations are unbiased given the available information.

This approach is popular because it is relatively easy to implement. It apparently obviates the need to model the whole system of variables involved in analysis, and in particular those that are thought of as 'exogenous'; it is known to be robust to a wider range of Data Generating Processes than FIML estimators (Hansen (1982)); and in general, it is expected to work well for the estimation of various types of Euler Equation models under weak conditions.

However, it is easy to see why such an approach invites criticism. First, it is not grounded on prior testing for the lack of feedbacks in the variables. This is a necessary condition for the absence of information loss in the estimation and inference on the parameters of interest. As we show below, the properties of the non-modelled variables are crucial for the identification of the model's parameters, even when the former are thought to be 'exogenously' determined. The non-modelled variables will not be weakly exogenous in the sense of Engle, Hendry, and Richard

(1983) in general, except in special cases, see discussion on p. 132. Thus, modelling the complete system will generally be informative.

Secondly, pathological cases such as 'weak instruments' are common across the spectrum of applied econometrics, and have been shown to impart serious distortions on the distributions of the estimators and test statistics, thus invalidating conventional inference, see for instance Hansen, Heaton, and Yaron (1996) and other papers in that issue of the *Journal of Business and Economic Statistics*, as well as Staiger and Stock (1997). Even when identification is not an issue, the finite-sample properties of GMM estimators and inference therefore can deviate substantially from the assumed asymptotic approximations when instruments are not strongly exogenous, when many irrelevant instruments are used, and when corrections are made for general serial correlation and heteroscedasticity of the residuals. All of these are common features of the models we discuss below.

Thirdly, a test of the over-identifying restrictions, which is the main mis-specification test used in this approach, may not be informative about the validity of the model or the 'rationality' of expectations.

1.2 IV regression with weak instruments

A considerable part of this thesis is devoted to analyzing the implications of the failure of the identification condition for the properties of various GMM estimators and test statistics. To motivate this analysis, we offer a simple exposition of this issue in the context of a univariate linear IV regression with fixed instruments. In this case, the analytics are simple and provide a useful insight into the more general asymptotic theory given in chapter 2 below, as well as a benchmark for interpreting the results of the Monte Carlo experiments in chapter 3.

Consider the IV estimator of a parameter θ in the model (1.3):

$$y = Y\theta + u \tag{1.3}$$

$$Y = Z\Pi + v \tag{1.4}$$

where (y, Y) is a $T \times (1+p)$ matrix of endogenous variables, Z is a non-stochastic $(T \times k)$ matrix of instrumental variables, such that $\lim_{T\to\infty} T^{-1}Z'Z = \Sigma_{ZZ}$, with $\text{rank}(\Sigma_{ZZ}) = \text{rank}(Z'Z) = k$ for all T, and $U = (u, v) \sim N(0,\ \Sigma_{UU} \otimes I_T)$. The quantity $\lambda = \Sigma_{vv}^{-1}\Sigma_{vu}$ measures the 'endogeneity' of Y and determines the bias of the OLS estimator of θ. When $\lambda \neq 0$, Y is clearly not weakly exogenous for θ in (1.3).

The IV estimator of θ is:

$$\widehat{\theta}_{IV} = \left(Y'Z\,(Z'Z)^{-1}\,Z'\,Y\right)^{-1} Y'\,Z\,(Z'\,Z)^{-1}\,Z'\,y = \left(\widehat{\Pi}'\,(Z'\,Z)\,\widehat{\Pi}\right)^{-1} \widehat{\Pi}'Z'\,y.$$

where $\widehat{\Pi}$ is the OLS estimator of Π in the 'first-stage' regression (1.4). When rank(Π) $= p$, the limiting distribution of $\widehat{\theta}_{IV}$ follows from standard asymptotic theory:

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) = \frac{1}{\sqrt{T}}\left(\widehat{\Pi}'\left(\frac{Z\,Z}{T}\right)\widehat{\Pi}\right)^{-1} \widehat{\Pi}'Z'u \xrightarrow{D} \mathsf{N}\left[0, \sigma_u^2\,(\Pi'\Sigma_{ZZ}\Pi)^{-1}\right]. \tag{1.5}$$

However, when rank(Π) $< p$, this approximation breaks down. It is easier to see what happens first in the univariate just-identified case, where $p = k = 1$, with $\Pi = 0$. Defining $e_t = (u_t - v_t'\,\lambda) \perp v_t$, with variance σ_e^2, the IV estimator can be written as:

$$\widehat{\theta}_{IV} = \theta_0 + \frac{Z'u}{Z'v} = \theta_0 + \lambda + \frac{Z'e}{Z'v} \sim (\theta_0 + \lambda) + \sigma\, t_1$$

where $\sigma = \sigma_e/\sigma_v$ and t_1 follows a Student's t-distribution with 1 degree of freedom (also known as the Cauchy distribution). This distributional result holds approximately (for large T), but also exactly (for any T), under the normality assumption for (u, v) and the non-randomness of the instruments. Thus, we see that in the un-identified case, the IV estimator is far from normal and exhibits a 'double' inconsistency: it is $O_p(1)$ (i.e., its variability does not fall with T), and centered on the plim of the OLS estimator, which is $(\theta_0 + \lambda)$. Figure 1.1 illustrates the fat-tailed nature of the distribution of the estimator by means of a Monte Carlo simulation, for a sample size of $T = 500$ using 10^5 replications.

Next, we look at what happens when we add more irrelevant instruments, i.e., $k > 1$ and still $\Pi = 0$. This time, the distribution of the IV estimator becomes:

$$\widehat{\theta}_{IV} - \theta_0 = \lambda + \frac{v'\,P_Z\,e}{v'\,P_Z\,v} \sim \lambda + \frac{\sigma}{\sqrt{k}}\,t_k$$

where $P_Z = Z\,(Z'Z)^{-1}Z'$ and t_k is distributed as Student's t with k degrees of freedom. We notice that the IV estimator now has moments up to the degree of over-identification, $k - 1$, and that its variance is falling linearly with the number of instruments. In other words, adding 'valid' but irrelevant instruments will make the distribution of the estimator resemble the asymptotic normal approximation, as illustrated in the top panel of figure 1.2 (contrast with the just-identified case above). Table 1.1 reports summary statistics for the simulated IV estimators, with $k = 1, \ldots, 5$ for a sample size of 500, and 100000 replications. As expected, the simulated second moments exhibit

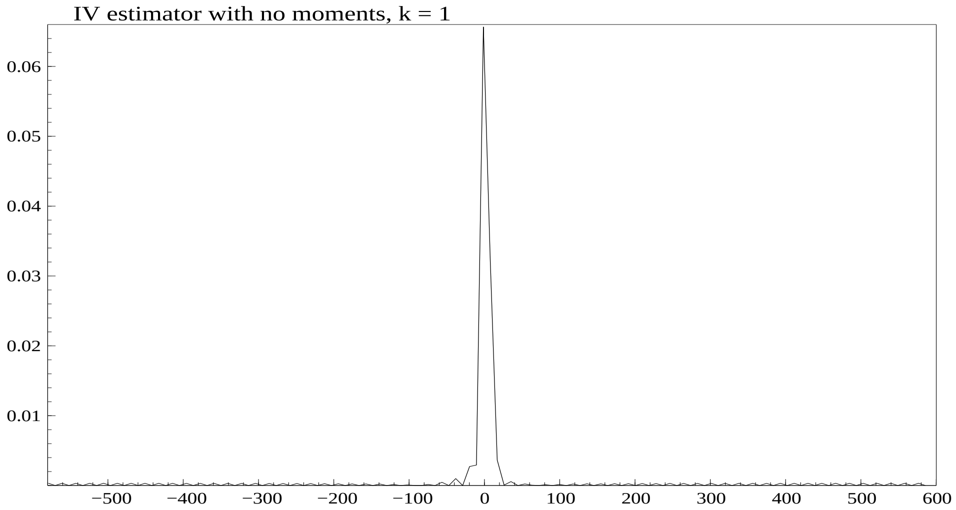


Figure 1.1: Just-identified IV estimator: no moments and wild Monte Carlo behaviour. The data is generated from equations (1.3)-(1.4), with $\theta = 1$, $\Pi = 0$, $\Sigma_{uu} = \Sigma_{vv} = 1$, $\Sigma_{vu} = 0.5$.

erratic behaviour for $k \leq 2$, since the corresponding theoretical moments do not exist. However, for $k > 2$, both moments exist, and the (over-identified) IV estimator exhibits a similar behaviour to that of the OLS estimator given at the bottom, albeit with much lower precision.

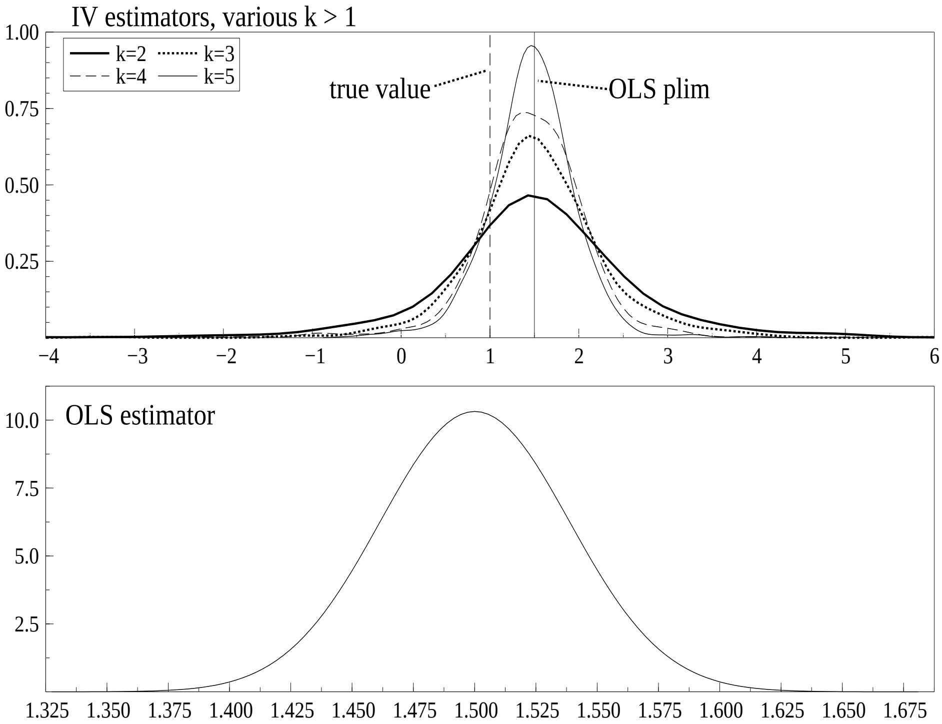
Table 1.1: First two moments and mean bias of IV estimator of θ.

k	Mean	St. Dev.	Mean Bias	RMSE	MCSE
1	1.8802	230.96	0.8802	230.97	0.7304
2	1.5163	2.6310	0.5163	2.6811	0.0083
3	1.5014	0.8620	0.5014	0.9972	0.0027
4	1.4993	0.6093	0.4993	0.7878	0.0019
5	1.5001	0.49831	0.5001	0.7060	0.0016
OLS:	1.5001	0.0387	0.5001	0.50158	0.0001

Simulation parameters: $\theta_0 = 1$, $\Sigma_{uu} = \Sigma_{vv} = 1$, $\lambda = 0.5$, $T = 500$, 10^5 replications.

Moreover, despite the falling variance, the mean bias of the estimator remains unchanged by the addition of instruments. Namely, the IV estimator exhibits the full OLS bias of λ, on average. This is hardly surprising, since, as the number of instruments approaches the samples size T, the IV estimator essentially becomes the OLS estimator, which is $\sqrt{T}$-consistent for $\theta_0 + \lambda$. It is in that sense that the un-identified IV estimator can be viewed as a “noised-up” OLS estimator.

Weak identification Building on the above discussion, we wish to investigate what happens when identification is ‘weak’, i.e., $\Pi \neq 0$ but ‘close’ to zero, in a sense that will be made precise in chapter 2. One approach is to develop higher order asymptotic approximations to the finite-sample



IV estimators, various k > 1

OLS estimator

Figure 1.2: Over-identified IV estimator versus OLS.
The data is generated from equations (1.3)-(1.4), with $\theta = 1$, $\Pi = 0$, $\Sigma_{uu} = \Sigma_{vv} = 1$, $\Sigma_{vu} = 0.5$.

distribution of the estimator, along the lines of Rothenberg (1984). Another approach, proposed by Staiger and Stock (1997), is to derive an alternative first-order asymptotic theory by linking the key parameter Π to the sample size.

Both of these approaches can be motivated by re-writing the IV estimator $\widehat{\theta}$ as a function of some pivotal statistics. This is straightforward in the univariate case, $p = 1$. Define the two independent standard normal variates:

$$z = \begin{pmatrix} z_v \\ z_e \end{pmatrix} = \begin{pmatrix} \sigma_v^{-1}\, v' \\ \sigma_e^{-1}\, e' \end{pmatrix} Z\,\Pi\,(\Pi' Z' Z \Pi)^{-1/2} \sim N(0, I),$$

and the variables S:

$$S = \begin{pmatrix} S_v & s \\ s & S_e \end{pmatrix} = \begin{pmatrix} v' P_Z\, v/\sigma_v^2 & v' P_Z\, e/(\sigma_v \sigma_e) \\ e' P_Z\, v/(\sigma_v \sigma_e) & e' P_Z\, e/\sigma_e^2 \end{pmatrix} \sim \mathcal{W}_2(k, I_2)$$

such that S follows a two-dimensional central Wishart distribution with k degrees of freedom. Also, let $\mu_T = (\Pi' Z' Z \Pi)^{1/2}/\sigma_v$. Then, the IV estimator can be written as:

$$\widehat{\theta}_{IV} - \theta_0 = \frac{(\lambda\, z_v + \sigma\, z_e)\mu_T + \lambda\, S_v + \sigma\, s}{\mu_T^2 + 2\, z_v\, \mu_T + S_v}. \tag{1.6}$$

In the completely un-identified case, $\Pi = 0$ and $\mu = 0$, (1.6) simplifies to the expressions given above. In the 'weakly' identified case, i.e., when $\Pi = O(T^{-1/2})$ (i.e., local to zero), $\mu_T = \mu$, reflecting the fact that the information is not increasing with the sample size. In that case, the estimator is still inconsistent ($O_p(1)$), and its distribution is non-Gaussian, depending also on two more nuisance parameters, μ and σ.

In the well-identified case, $\mu_T = O(T^{1/2})$, we see that it is this parameter that drives the convergence of the distribution to normality. Upon re-scaling by μ_T, we may re-write (1.6) as:

$$\mu_T(\widehat{\theta}_{IV} - \theta_0) = \frac{(\lambda\, z_v + \sigma\, z_e) + \lambda\, S_v/\mu_T + \sigma\, s/\mu_T}{1 + 2\, z_v/\mu_T\ + S_v/\mu_T^2}.$$

Since the terms divided by μ_T are small, we can perform a Nagar expansion to order $O_p(\mu_T^{-1})$:

$$\mu_T(\widehat{\theta}_{IV} - \theta_0) = (\lambda\, z_v + \sigma\, z_e) + \frac{\lambda\, S + \sigma\, s - 2\,(\lambda\, z_v + \sigma\, z_e)\, z_v}{\mu_T} + o_p(\mu_T)$$

where the leading term $\lambda\, z_v + \sigma\, z_e$ follows a $N(0, \sigma_u^2)$, as expected. Rothenberg (1984) explains how this approach can be used to derive higher-order Edgeworth expansions with a view to characterizing the departure of the distribution from normality, and offer a better approximation. The local-to-zero approach of Staiger and Stock (1997), whereby $\Pi = C/\sqrt{T}$ for some constant matrix C, is an alternative to the higher-order approximations of Rothenberg (1984). Stock and Wright (2000) generalized the local-to-zero approach for non-linear GMM models, see chapter 2. This is the approach we adopt below, both for the theoretical analysis and for Monte Carlo experiments.

1.3 Thesis Outline

This thesis provides a critique of the single-equation GMM approach to the estimation of forward-looking models, with emphasis on monetary economics applications. We are not concerned with any economic theoretical considerations arising in models like (1.1) or (1.2), even though they are undoubtedly important. For example, we are not questioning the requirement of model-consistency of expectations (rational expectations), which has been the subject of considerable debate elsewhere, see Hendry (2000). Nor are we going to question the interpretation of the structural parameters or challenge the micro-foundations involved. Rather, we shall focus entirely on the econometric issues involved in the estimation of those models by the GMM.

In particular, we will show why single-equation modelling results in loss of information about the identifiability of the structural model and examine the implications of weak identification for GMM estimation and inference in cases which are shown to be empirically relevant. Also, we analyze the usual Hansen-Sargan test of over-identifying restrictions, and show why it has low power to detect mis-specification in this context.

The following issues will be explored:

1. The properties of GMM estimators and inference in dynamic forward-looking models, when identification is weak.

2. The power of the Hansen-Sargan test of over-identifying restrictions when the model is mis-specified.

3. The implications of mis-specification for GMM estimation and inference on the parameters of interest.

We start by reviewing and extending recent results in the literature on GMM under weak identification in chapter 2. This is done for a generalization of the standard identification condition, which allows for partial or weak identification of functions of the parameters of interest, as well as failure of the orthogonality conditions. Asymptotic approximations to the distributions of GMM estimators and tests are derived as an extension of Stock and Wright (2000), and various inferential procedures that are robust to weak identification are discussed. We also study different ways of accounting for serial correlation in the structural model residuals, which are intrinsic in forward-looking models. Next, we characterize the asymptotic power of the Hansen-Sargan mis-specification test, and uncover the implicit null hypothesis being tested. Finally, we develop tests for identification and excess serial correlation that may be used in forward-looking models.

The following chapter then provides Monte Carlo evidence on the asymptotic results given in chapter 2, when instruments include lags of the regressand and possibly other endogenous variables and hence are not strongly exogenous, as well as under serially-correlated structural errors. The quality of the first-order asymptotic approximation is evaluated, and the relative merits of alternative GMM estimators under weak identification are assessed. We also numerically evaluate the utility of the robust inferential methods for dynamic forward-looking macro models. Moreover, we analyze the implications of different GMM estimators for the null rejection probability of the Hansen-Sargan test, and study the size and power of the excess serial correlation and identification tests developed in the previous chapter. A new control variable is introduced and helps reduce the variance of the Monte Carlo experiments.

Chapter 4 then applies these results to the analysis of a particular forward-looking model, the New-Keynesian Phillips Curve of Galí and Gertler (1999). This model has been the subject of criticism elsewhere, see Rudd and Whelan (2001) and Bårdsen, Jansen, and Nymoen (2002). Our results are broadly in line with the conclusions in those papers, casting doubts on the utility of the New-Keynesian Phillips curve as a model of inflation dynamics. However, the main objective of the chapter is not the testing of the above model, but rather more general: we raise methodological issues that appear to be overlooked in the monetary economics literature, and uncover generic weaknesses of single-equation GMM methods in the context of dynamic macro models. Our proposed methodology, which is applicable to any forward-looking model, is based on modelling the reduced form of the whole system under analysis, and the use of the underlying economic theory to assess the identification and mis-specification of proposed forward-looking models. This approach can be more informative than formal identification and mis-specification pre-tests, and we recommend that it is applied prior to embarking on any structural econometric modelling.

Finally, all of the numerical results reported below are derived using Ox version 3.2, see Doornik (2001). The computer code for the analysis of rational expectations models, GMM estimation, Monte Carlo simulation, as well as asymptotic analysis in this thesis was developed by the author and is documented in the appendix at the end.

Chapter 2

GMM regression with weak instruments

The Generalized Method of Moments (GMM) has traditionally been a popular alternative to Full Information Maximum Likelihood (FIML) for the estimation of structural models. There are several reasons for this: it apparently obviates the need to model the whole system of variables involved in analysis, and in particular those that are thought of as 'exogenous'; GMM is known to be robust to a wider range of Data Generating Processes than is FIML (Hansen (1982)); and GMM usually involves relatively small efficiency loss against the latter (e.g., West (1986)). In general, GMM is expected to work well for the estimation of various types of Euler Equation models under weak conditions.

However, this generality comes at a cost. First, when 'too many' instruments are used, the chances of including irrelevant ones is high ('over-instrumenting'), thus increasing the bias of the GMM estimators and reducing the power to detect invalid restrictions. Secondly, the common practice of 'robustifying' inference by means of general serial correlation corrections ('over-correction') is likely to affect specification testing and obscure a model's inadequacies. These possibilities are particularly relevant for dynamic forward-looking models, where the instrument set is potentially infinite and serially-correlated errors are common.

Third, there is a substantive condition that must be satisfied for GMM estimation to work, namely the model must be *identified* on the available information set. It is precisely the failure of that condition that is responsible for many of the documented ills of GMM.

In the last decade, there has been growing research on the properties of GMM estimators of structural models when the correlation between the instruments and the endogenous regressors is low. A number of studies have reported poor performance of GMM estimators in various different contexts, e.g., Hansen, Heaton, and Yaron (1996) for asset pricing models, Burnside and Eichenbaum (1996) for business cycle models, Fuhrer, Moore, and Schuh (1995) for inventory models, amongst others. Alongside these studies, a corresponding analytical literature has emerged. We distinguish two main agendas in this research project. On the one hand, there is an attempt to characterize the implications of weak identification for GMM estimators and inference procedures, and hence propose methods that are robust to identification failure. On the other hand, there is a need to develop diagnostic tools capable of detecting identification problems.

We first give a brief historical overview of the developments in this literature. Exact finite-sample distribution results exist for the IV and general k-class estimators in linear simultaneous equations models under strong identification, see Phillips (1983) for a comprehensive review. These results were extended to the case of partially identified models, in which some of the structural parameters are totally un-identified, by Phillips (1989) and Choi and Phillips (1992). Later, the idea of local-to-zero asymptotic analysis was introduced by Staiger and Stock (1997), who characterized the properties of k-class estimators and associated inference under this type of 'weak' identification. These asymptotic approximations were shown to provide accurate descriptions of finite-sample distributions in some cases, as well as useful prescriptions for empirical work. The idea of local-to-zero asymptotics was further extended by Stock and Wright (2000) to a wider class of GMM estimators of non-linear models. One of the main conclusions of that work was that apart from the failure of standard asymptotic theory, weak identification implied that different types of GMM estimators were no longer asymptotically equivalent, as in the strongly identified case, leading to potentially significant differences in inference.

The objective of this chapter is twofold. On the one hand, we wish to review the latest developments in this area and provide a unified account of the main findings. On the other hand, we wish to make a number of contributions which are of particular relevance to the estimation of dynamic forward-looking models. We provide some extensions of the main asymptotic results pertaining to the estimation of functions of identified and weakly identified parameters, and to serially-correlated structural errors, which are common in dynamic rational expectations models. We discuss some related issues arising in the estimation of the optimal GMM weights by means of Heteroscedasticity and Autocorrelation Consistent (HAC) estimators.

The analysis in this chapter clarifies that the main problems for GMM estimation and inference arise from failures of the identification condition. Hence, there is a need to develop tools that can diagnose such failures, which may be due to either weak identification or mis-specification. So, contributing to ongoing research in this area, we develop a new test for identification that is suitable for forward-looking linear models estimated by GMM. Next we analyze the Hansen-Sargan test of the over-identifying restrictions under the weak identification assumption D on page 18. We characterize the asymptotic distribution of the test statistic under the null and the alternative, and propose a boundedly pivotal version of the test that provides valid but conservative inference under weak instruments. Given the low power of this test, especially when the number of instruments is large, we examine an alternative specification test of excess residual serial correlation, which would also provide evidence on model mis-specification.

The structure of the chapter is as follows. Section 2.1 lists the main assumptions, gives motivating examples and defines some key terms. The following section provides the main asymptotic results on the properties of GMM estimators and inference under weak instruments. Then, we discuss GMM covariance matrix estimation in section 2.3. Identification and mis-specification tests are developed and discussed in sections 2.4 and 2.5, and finally, we summarize the key findings at the end.

2.1 GMM estimation and inference procedures

Consider a random sample of data $\{\mathrm{y}_t,\ t = 1, \ldots, T\}$, and a p-dimensional vector of parameters $\theta \in \boldsymbol{\Theta} \subseteq \Re^p$. Let $h(\mathrm{y}_t,\, \theta)$ be a G-dimensional vector-valued function whose expectation under the distribution of y_t, conditional on some information set $\mathcal{F}_t$, vanishes at the true value of θ, denoted θ_0, i.e.:[1]

$$\mathrm{E}(h(\mathrm{y}_t,\, \theta_0)|\mathcal{F}_t) = 0. \tag{2.1}$$

Also, we let $Z_t \in \mathcal{F}_t$ denote a k-dimensional vector of instruments. For convenience, the notation is summarized in table 2.1.

Let W_T be a weighting matrix of dimension $Gk \times Gk$, which may depend on the data and parameters. Also, let $f_t(\theta) = h_t(\theta) \otimes Z_t$ denote the moment conditions per observation, whose unconditional expectation vanishes at the true parameter point. The GMM estimator $\widehat{\boldsymbol{\theta}}_T$ is defined

---

[1] Strictly speaking, this condition must hold *for all possible true parameter points* in $\boldsymbol{\Theta}$, i.e., $\mathrm{E}_\theta(h(\mathrm{y}_t, \theta)|\mathcal{F}_t) = 0$, $\forall \theta \in \boldsymbol{\Theta}$, see Davidson and MacKinnon (1994).

Table 2.1: Notation.

Mnemonic	Definition	Description
y_t		Endogenous variables/regressors.
Z_t	$\in \mathcal{F}_t \quad (k \times 1)$	Instruments.
θ	$\in \boldsymbol{\Theta} \quad (p \times 1)$	Parameter vector.
Moments		
$h_t(\theta)$	$= h(\mathrm{y}_t,\, \theta) \quad (G \times 1)$	Moment functions (G equations).
$f_t(\theta)$	$= h(\mathrm{y}_t,\, \theta) \otimes Z_t, \qquad (Gk \times 1)$	Moment conditions per observation.
$g_T(\theta)$	$= T^{-1} \sum_{t=1}^{T} f_t(\theta), \qquad (Gk \times 1)$	Sample moments.
$m_T(\theta)$	$= T^{-1} \sum_{t=1}^{T} \mathrm{E}(f_t(\theta) \vert \mathcal{F}_t), \qquad (Gk \times 1)$	Population moments.
$\Psi_T(\theta)$	$= T^{-1/2} \sum_{t=1}^{T} [f_t(\theta) - \mathrm{E}(f_t(\theta) \vert \mathcal{F}_t)]$	Centered & scaled sample moments.
Jacobian of moments		
$d_t(\theta)$	$= \frac{\partial f_t(\theta)}{\partial \theta'}, \quad (Gk \times p)$	Jacobian of moments per observation.
$D_T(\theta)$	$= T^{-1} \sum_{t=1}^{T} d_t(\theta)$	Sample analogue.
$\mathcal{J}(\theta)$	$= \operatorname*{plim}_{T \to \infty} D_T(\theta)$	Probability limit of Jacobian.
$\Xi_T(\theta)$	$= T^{-1/2} \sum_{t=1}^{T} [d_t(\theta) - \mathrm{E}(d_t(\theta) \vert \mathcal{F}_t)]$	Centered & scaled Jacobian.
Variances		
Σ_{ZZ}	$= \operatorname*{plim}_{T \to \infty} T^{-1} \sum_{t=1}^{T} Z_t Z_t', \qquad (k \times k)$	Second moments of instrument vector.
$\widehat{\Sigma}_{ZZ}$	$= T^{-1} \sum_{t=1}^{T} Z_t Z_t'$	Sample analogue.
$V_T(\theta)$	$= \mathrm{var}[\Psi_T(\theta)], \qquad (Gk \times Gk)$	Variance of sample moments at T.
$\Omega(\theta_1, \theta_2)$	$= \mathrm{E}\left[\Psi(\theta_1)\, \Psi(\theta_2)'\right], \quad (Gk \times Gk)$	Limiting variance of sample moments.
$\mathcal{V}_T(\theta)$	$= \mathrm{var}[\mathrm{vec}\,(\Xi_T(\theta))], \qquad (Gkp \times Gkp)$	Variance of the average Mom. Jacob.
$\mathcal{C}_T(\theta)$	$= \mathrm{cov}\left[\mathrm{vec}\left(\Xi_T(\theta)\right), \Psi_T(\theta)\right], \qquad (Gkp \times Gk)$	Covar. of mom. cond. and their Jacob.
GMM criterion		
$W_T(\bar{\theta})$	$W_T(\bar{\theta}_T(\theta)), \quad (Gk \times Gk)$	Weighting matrix.
$W(\theta)$	$\operatorname*{plim}_{T \to \infty} W_T(\theta)$	Probability limit of W_T.
$\mathcal{Q}_T(\theta; \bar{\theta})$	$= T\ g_T(\theta)' W_T(\bar{\theta}) g_T(\theta)$	GMM criterion function.

as the value of $\boldsymbol{\theta}$ that minimizes the criterion function:

$$\mathcal{Q}_T(\theta; \bar{\theta}_T(\theta)) = \left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\theta)\right]' W_T\left(\bar{\theta}_T(\theta)\right)\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t(\theta)\right] \tag{2.2}$$

where $\bar{\theta}_T(\theta)$ is a preliminary estimator of θ, and we follow the notation of Stock and Wright (2000) to allow for various types of GMM estimators, namely one-step (1S), two-step (2S), multi-step (MS) and Continuous Updating Estimators (CUE), e.g., $\bar{\theta}_T(\theta) = \widehat{\theta}_T^{(m-1)}$ for the m-step estimator $\widehat{\theta}_T^{(m)}$, or $\bar{\theta}_T(\theta) = \theta$ for the CUE.

2.1.1 Assumptions

Next, we write down a set of standard regularity conditions that will normally guarantee the consistency, asymptotic normality and asymptotic efficiency of GMM estimators, in conjunction with a global identification condition. This last condition is substantive as opposed to a regularity condition, and it is discussed separately. Some assumptions on the shape of the parameter space and the smoothness of the objective function are given in the appendix and are denoted assumptions A.1 to A.6. In addition, for the asymptotic theory, we need a law of large numbers, a central limit theorem and some structure for the weighting matrix.

Let $g_T(\theta) = T^{-1}\sum_{t=1}^{T} f_t(\theta)$ denote the empirical moment conditions, and define the corresponding population moments by $m_T(\theta) = T^{-1}\sum_{t=1}^{T} \mathrm{E}(f_t(\theta)|\mathcal{F}_t)$. Also, define the centered and scaled moment vector $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}[f_t(\theta) - \mathrm{E}(f_t(\theta)|\mathcal{F}_t)]$.

Assumption B (Central Limit Theory). *The sample moment conditions $g_T(\theta)$ satisfy:*

1. *(WULLN) a Weak Uniform Law of Large Numbers, $g_T(\theta) - m_T(\theta) \xrightarrow{p} 0$ uniformly in $\theta \in \boldsymbol{\Theta}$;*
2. *(FCLT) a Functional Central Limit Theorem, $\Psi_T(\theta) \Rightarrow \Psi(\theta)$ uniformly in $\theta \in \boldsymbol{\Theta}$, where $\Psi(\theta)$ is Gaussian with zero mean and covariance function $\Omega(\theta_1, \theta_2) = \mathrm{E}\left[\Psi(\theta_1)\,\Psi(\theta_2)'\right]$.*[2]

Assumption B.1 is a standard uniform Law of Large Numbers. For some of the results given below, it could be relaxed to the weaker condition $g_T(\theta_0) - m_T(\theta_0) \xrightarrow{p} 0$. Assumption B.2 implies that, for every $\theta \in \boldsymbol{\Theta}$, the sample moments, centered by the respective population moments and scaled by $\sqrt{T}$, have a limiting Normal distribution. Moreover, for any pair of values $\theta_1, \theta_2 \in \boldsymbol{\Theta}$,

---

[2] In line with the general assumptions in the literature, we have taken sufficiently strong assumptions such that the limit of the expectation exists, and equals the probability limit of sample second moments. Technically, there are many conditions under which the two are not equal, e.g. when the expectation doesn't exist. However, this assumption is maintained because interest centers on the finite-sample behaviour where the expectation is the relevant operator.

the corresponding pair $(\Psi_T(\theta_1), \Psi_T(\theta_2))$ is asymptotically jointly Normal with covariance matrix $\Omega(\theta_1, \theta_2)$. An example in the context of linear models is given in chapter 3 below.

A variant of the above assumption is needed for one of the results that follow. This assumption, which matches Kleibergen (2001, Assumption 1), imposes weaker conditions on the convergence of the sample moments than assumption B, but a stronger condition on the derivatives of the moment conditions. Let $d_t(\theta) = \frac{\partial f_t(\theta)}{\partial \theta'}$ denote the Jacobian matrix of $f_t(\theta)$, of dimension $(Gk \times p)$, $D_T(\theta) = T^{-1}\sum_{t=1}^{T} d_t(\theta)$ denote the empirical average moment Jacobian, and $\mathcal{J}(\theta) = \operatorname*{plim}_{T\to\infty} D_T(\theta)$ denote its probability limit. Also, define the scaled and centered Jacobian matrix $\Xi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}[d_t(\theta) - \mathrm{E}(d_t(\theta)|\mathcal{F}_t)]$.

Assumption B′. *The sample moment conditions jointly with the sample Jacobian satisfy the central limit theorem:*

$$\begin{pmatrix} \Psi_T(\theta_0) \\ \operatorname{vec}(\Xi_T(\theta_0)) \end{pmatrix} \xrightarrow{d} N\left(0, \mathbf{V}(\theta_0)\right), \qquad \mathbf{V}(\theta_0) = \begin{pmatrix} V(\theta_0) & \mathcal{C}(\theta_0)' \\ \mathcal{C}(\theta_0) & \mathcal{V}(\theta_0) \end{pmatrix}$$

where the $Gk(p+1)$-square matrix $\mathbf{V}(\theta_0)$ is positive definite, and $V(\theta_0) = \lim_{T\to\infty} V_T(\theta_0) = \Omega(\theta_0, \theta_0)$, $\mathcal{V}(\theta_0) = \lim_{T\to\infty} \mathcal{V}_T(\theta_0)$ and $\mathcal{C}(\theta_0) = \lim_{T\to\infty} \mathcal{C}_T(\theta_0)$.[3]

Assumption C. *$W_T(\theta) \xrightarrow{p} W(\theta)$, uniformly in $\theta \in \mathbf{\Theta}$, where $W(\theta)$ is a non-random symmetric positive definite matrix.*

The assumptions given here are quite high level, and they could be derived from more primitive sufficient conditions, see for example Hansen (1982), or Newey and McFadden (1994).

2.1.2 Identification and Mis-specification

The main substantive condition imposed in a GMM model is that a set of moment conditions vanish only at a *unique* parameter value $\theta_0 \in \mathbf{\Theta}$, usually referred to as the ‘true parameter’.

There are two important ways in which this condition can be violated. First, when there is a *set* of values for which the moment conditions vanish (rather than a unique point), i.e., $m(\theta) = 0, \quad \forall\ \theta \in \mathbf{\Theta}_0 \subset \mathbf{\Theta}$. This is referred to as *partial* or *weak identification* (see formal definition below).[4] Second, when there exists *no* $\theta \in \mathbf{\Theta}$ that satisfies the moment conditions.

---

[3] The assumption that $\mathcal{V}$ is Positive Definite (PD) is not restrictive and it can be relaxed to Positive Semi-Definite (PSD), when there are exogenous variables in the model, at the cost of introducing some extra notation, see Kleibergen (2001, Assumption 1).

[4] There is local and global identification, but in this study we will be concerned primarily with a failure of local identification, i.e., the situation in which $\mathbf{\Theta}_0$ is continuous. For the failure of global (but not local) identification, see Sargan (1983).

This is referred to as *mis-specification.* However, even under mis-specification, there will exist some θ_* that will minimize the limiting objective function, in general. This is referred to as the 'pseudo-true' value of the parameter.[5]

Here, we shall use a slightly generalized definition of identification, to make it consistent with mis-specification.

Definition 2.1 (identification). *The parameters* θ *are (asymptotically) identified if and only if the limiting objective function has a unique minimum, i.e., if there exist* $\theta_* \in \mathbf{\Theta}$ *such that* $\bar{\mathcal{Q}}(\theta_*) < \bar{\mathcal{Q}}(\theta), \forall\ \theta \in \mathbf{\Theta} \setminus \{\theta_*\}$.

Definition 2.2 (mis-specification). *The model is mis-specified if* $\bar{\mathcal{Q}}(\theta) > 0,\ \forall\ \theta \in \mathbf{\Theta}$.

That the objective function has an interior minimum (existence) follows from its continuity and the compactness of $\mathbf{\Theta}$, see assumptions in appendix 2.A.[6] Also, when the limiting objective function is assumed to be twice continuously differentiable (assumption A.3), the uniqueness issue can be investigated by looking at its Hessian at some candidate turning point θ_0. In the case of no mis-specification, the Hessian will simply be $\mathcal{J}(\theta_0)'W(\theta_0)\mathcal{J}(\theta_0)$ and, in view of assumption C, a necessary and sufficient condition for uniqueness is that $\mathcal{J}(\theta_0)$ is of full rank. Moreover, given assumption C, the limiting objective function $\bar{\mathcal{Q}}(\cdot)$ could be seen as 'locally quadratic' in θ, in some neighbourhood of the true parameter $\mathcal{N}(\theta_0)$, whenever θ is both identified and the model is correctly specified. Under mis-specification, however, the resulting Hessian will be more involved, and this condition is no longer sufficient.

It is worth pointing out that identification and mis-specification problems are not necessarily mutually exclusive. They can co-exist when the minimum of the limiting objective function is both bounded above zero for all $\theta \in \mathbf{\Theta}$ (mis-specification), and it is non-unique (lack of identification). Moreover, as will become apparent from the examples below, it is often the case that one model's mis-specification problem is another's identification problem. Namely, a highly restrictive parameterization may be generalized at the cost of losing identification, and vice versa, a poorly identified general model can be rendered identified through restrictions that induce mis-specification.

In order to formalize the identification condition, let us define a generic re-parameterization of θ into α and β, of dimension p_1 and p_2 respectively ($p_1 + p_2 = p$), by an invertible mapping

---

[5] Strictly speaking, we should write $\theta_*(W)$ to reflect the dependence of the pseudo-true value on the limiting weighting matrix W. We drop that dependence for convenience, when it is not necessary.

[6] The latter can be relaxed at the cost of introducing a quasi-concavity assumption on the limiting objective function. This holds trivially for linear models, for example.

$\kappa : \boldsymbol{\Theta} \mapsto A \times B$, and define $\phi(\cdot) = \kappa^{-1}(\cdot)$,

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \kappa(\theta) = \begin{pmatrix} \alpha(\theta) \\ \beta(\theta) \end{pmatrix}$$

The following assumption generalizes the original GMM identification condition of strong asymptotic identifiability, namely that the moment conditions have a unique zero at $\theta = \theta_0$ and that the moment Jacobian is of full rank, see Hansen (1982).

Assumption D (identification). *Assume that $\phi(\cdot)$ is continuous and that*

$$T^{-1}\sum_{t=1}^{T} \mathrm{E}(f_t(\theta)|\mathcal{F}_t) \equiv m_T(\theta) = m_{1T}(\theta)/\sqrt{T} + m_{2T}(\beta),$$

where:

1. $m_{1T}(\theta) \xrightarrow{p} m_1(\theta)$ *uniformly in* $\theta \in \boldsymbol{\Theta}$, $m_1(\theta_0) = 0$ *and* $m_1(\theta)$ *is continuous in* θ *and is bounded in* $\boldsymbol{\Theta}$.
2. $m_{2T}(\beta) \xrightarrow{p} m_2(\beta)$, *uniformly in* $\beta \in B$, *such that* $m_2(\beta_0) = 0$ *and* $m_2(\beta) \neq 0$ *for* $\beta \neq \beta_0$, *where* $\beta_0 = \beta(\theta_0) = \beta(\theta)$ *for* $\theta \in \boldsymbol{\Theta}_0$.
3. $\mathcal{J}_2(\beta) = \partial m_2(\beta)/\partial\beta$ *is continuous, and* $\mathcal{J}_2(\beta_0)$ *has full column rank* p_2.

This assumption matches closely assumption C of Stock and Wright (2000). It concerns the partitioning of the parameter space into identified and ‘weakly’ identified directions. The only generalization is that it allows the identified parameters to be *functions* rather than mere subsets of the original parameter set. This generalization is important in practice because it is rare to encounter an exact partitioning of the parameter space into identified and un-identified subspaces.

Also, note that what is important in assumption D is the function $\beta(\cdot)$. This must be specified completely for this assumption to hold, but the remaining, weakly identified directions are not uniquely determined. In other words, we have some degree of freedom in choosing a basis for α, as it will become apparent in the examples below.[7]

Partial identification is a special case of assumption D. This arises when $m_{1T}(\theta)$ is identically equal to zero for all $\theta \in \boldsymbol{\Theta}$. In that case, the parameters $\alpha = \alpha(\theta)$ are completely un-identified. The other limiting case of complete identification also arises as a special case of this assumption

---

[7] The only other difference with Stock and Wright (2000, Assumption C) is that m_T is defined in terms of conditional rather than unconditional expectations. This distinction is made for clarity and to simplify the ensuing asymptotic analysis, without affecting any of the conclusions. Strictly speaking, this is the assumption on which Stock and Wright base section 3 of their paper.

by setting $\beta(\theta) = \theta$.[8] Neither of these cases can adequately characterize situations in which all of the parameters are identified, but some (possibly) functions of the parameters are better identified than others in a given sample. Assumption D is a way of linking the two limiting cases. Given that $\beta(\theta)$ is well-identified, the distance of $m_1(\phi(\alpha, \beta_0))$ from zero at $\alpha \neq \alpha_0$, in the metric of the limiting weighting matrix $W(\phi(\alpha, \beta_0))$, determines the identifiability of α: for any value $\alpha_1 \neq \alpha_0$, the bigger is $m_1(\phi(\alpha_1, \beta_0))' W(\phi(\alpha_1, \beta_0))\, m_1(\phi(\alpha_1, \beta_0))$ the more concentrated the GMM estimator of α, $\alpha(\widehat{\theta})$ will be around the true value α_0.

It is also useful to distinguish between the terms ‘weak identification’ and ‘weak instruments’. The latter refers to a situation in which the correlation between the instruments Z_t and the endogenous regressors in a linear model is small relative to the sample size. In linear models, this coincides with the ‘weak identification’ in assumption D, see the first two examples in the next subsection. This is due to the fact that the population moments $m_T(\theta)$ are a linear function of the second moments of the data, see (2.6). Hence the measure of identification $m_1(\phi(\alpha, \beta_0))' W(\phi(\alpha, \beta_0))\, m_1(\phi(\alpha, \beta_0))$ is a quadratic form in α whose shape depends only on the correlation between the regressors and the instruments. However, in non-linear models, the above function is no longer a quadratic and the correlation between regressors and instruments is insufficient to characterize its shape, see also the discussion in Stock and Wright (2000, section 2.6).

Finally, we can discuss local and fixed mis-specification using the following modification of assumption D.

Assumption D′ (mis-specification). $T^{-1}\sum_{t=1}^{T} \mathrm{E}(f_t(\theta)|\mathcal{F}_t) = \tilde{m}_T(\theta) + \zeta_T$, *where* $\tilde{m}_T(\theta)$ *has the same form as* $m_T(\theta)$ *in assumption D,* ζ_T *is independent of* θ*, and* $\zeta_T = T^{-1/2}\zeta_l + \zeta_f$.

The parameter ζ_T is introduced to bound the moment conditions (2.1) away from zero for all admissible parameter values $\theta \in \mathbf{\Theta}$, and will in general depend on the second moments of the data, for instance, see the example in appendix 3.B. However, the mis-specification does depend on θ in general, even though ζ_T doesn't.

---

[8] Alternatively, complete identification would arise if $m_{1T}(\theta)$ were of order $O_p(T^{1/2})$.

2.1.3 Examples

To motivate the above assumptions, we provide three representative examples where this framework can be applied.

Single equation linear instrumental variables regression model

The first example is a standard single-equation linear instrumental variables model. It can be either static or dynamic, where the instruments can be thought of as strongly exogenous or fixed, and exact finite-sample distributions results are available (see for example Phillips (1989)). This is an important example, which is used in most of the recent papers to emphasize the connection of the latest contributions with the earlier literature on IV.

Consider the prototype IV regression of y_t on a set of p endogenous variables Y_t^*, using Z_t as instruments:

$$\begin{aligned} y &= Y^* \theta_0 + u && (2.3) \\ Y^* &= Z\,\Pi_T^* + v^* && (2.4) \end{aligned}$$

where $\{y, Y^*, Z\}$ stack the observations for the entire sample, and u_t, v_t^* are orthogonal to Z_t. Suppose that $\Pi_T^* = C^*/\sqrt{T} + B\,d'$ where B and d are full-rank matrices of dimension $k \times p_2$ and $p \times p_2$ respectively. Let the $p \times p_1$ matrix $d_\perp$ denote the orthogonal complement of d, such that $(d, d_\perp)$ is of full rank p, and $d_\perp' d = 0$. Without loss of generality, we may assume that $(d, d_\perp)$ is normalized s.t. $d\,d' + d_\perp\,d_\perp' = I_p$. Then, setting

$$\Phi = \begin{pmatrix} d_\perp & d \end{pmatrix} \quad \text{s.t.} \quad \Phi^{-1} = \begin{pmatrix} d_\perp' \\ d' \end{pmatrix}$$

we see that the re-parameterizing transformation of assumption D can be written as:[9]

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \Phi^{-1}\,\theta \quad \text{and} \quad \phi(\theta) = \Phi \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = d_\perp\,\alpha + d\,\beta \tag{2.5}$$

Using this, and setting $C = (C_\alpha, C_\beta) = C^*\,\Phi$ for clarity, we can transform the regressors and first-stage regression parameters accordingly into:

$$\Pi_T^* = \underbrace{\begin{pmatrix} T^{-1/2}\,C_\alpha, & T^{-1/2}\,C_\beta + B \end{pmatrix}}_{\Pi_T} \begin{pmatrix} d_\perp' \\ d' \end{pmatrix} = \Pi_T\,\Phi^{-1}, \quad Y_t = \Phi' Y_t^*, \quad v_t = \Phi' v_t^*$$

---

[9] Obviously, the identified directions $d'\theta$ are determined up to rotation by any non-singular $p_2 \times p_2$ matrix, so the specification of $\phi(\cdot)$ is non-unique.

so that the system may be re-written as:

$$\begin{aligned} y &= Y\,\Phi^{-1}\theta_0 + u = Y_\alpha\,\alpha_0 + Y_\beta\,\beta_0 + u \\ Y &= Z\,\Pi_T + v = Z\,\left[T^{-1/2}C_\alpha,\;\; T^{-1/2}C_\beta + B\right] + [v_\alpha\;\; v_\beta] \end{aligned}$$

where $U_t = [u_t,\; v_t']'$ and Z_t satisfy $\mathrm{E}[U_t|\; Z_t] = 0$, and where Π and v are partitioned conformably with Y.

In terms of the notation of table 2.1:

$$\begin{aligned} h_t(\theta) &= y_t - \theta'(\Phi^{-1})'Y_t = u_t + (\alpha_0-\alpha)'\,Y_{\alpha,t} + (\beta_0-\beta)'\,Y_{\beta,t} \\ f_t(\theta) &= [u_t + (\theta_0-\theta)'(\Phi^{-1})'Y_t]Z_t \\ \Psi_T(\theta) &= T^{-1/2}\textstyle\sum_{t=1}^{T}\left(\left[1\;\;(\theta_0-\theta)'(\Phi^{-1})'\right]U_t\right)]Z_t \\ \Xi_T(\theta) &= T^{-1/2}\textstyle\sum_{t=1}^{T}\left(Z_tY_t' - Z_tZ_t'\Pi_T\right)\Phi^{-1} \\ m_T(\theta) &= -T^{-1}\textstyle\sum_{t=1}^{T} Z_tZ_t'\left\{\left[C_\alpha(\alpha-\alpha_0) + C_\beta(\beta-\beta_0)\right]/\sqrt{T} + B\,(\beta-\beta_0)\right\} \end{aligned} \tag{2.6}$$

It is easy to see that assumptions B and C hold if we assume that the first two sample moments of U_t and Z_t converge in probability to some positive definite matrices Σ_{UU} and Σ_{ZZ}, respectively, and $T^{-1/2}\sum_{t=1}^{T} U_t\otimes Z_t \overset{d}{\to} N(0,\Sigma_{UU}\otimes\Sigma_{ZZ})$.

Assumption D is then satisfied with $m_1\left(\phi(\alpha,\beta)\right) = -\Sigma_{ZZ}\,\left(C_\alpha(\alpha-\alpha_0) + C_\beta\,(\beta-\beta_0)\right)$ and $m_2(\beta) = -\Sigma_{ZZ}\,B\,(\beta-\beta_0)$. Finally, the limiting moment Jacobian w.r.t. the identified parameters β is simply $\mathcal{J}_2(\beta) = -\Sigma_{ZZ}\,B$.

Linear dynamic rational expectations model

This is a special type of an IV regression model where the endogenous regressors include leads of the dependent or other variables. This type of model is of particular interest for a number of reasons that we discuss in detail in chapter 4. First, the previous assumption of strongly exogenous instruments no longer applies. Second, the IV regression error is expected to exhibit serial correlation by construction. Third, there is a potentially infinite set of admissible instruments, most of which are only mildly correlated with the regressors. Fourthly, the identification of the model depends crucially on the properties of the driving process, even when the latter is seemingly exogenous. And finally, it is an example where the parameter set cannot be partitioned exactly into identified and un-identified directions, thus providing a case for the generalization in assumption D over Stock and Wright (2000, Assumption C).

The example here is a pure version of the “New Phillips curve”, a forward-looking model for inflation dynamics, taken from Galí and Gertler (1999, section 2), see section 4.1 for details.[10]

$$\pi_t = \lambda_0 \mathrm{E}(s_t|\mathcal{F}_t) + \gamma_0 \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \epsilon_t \tag{2.7}$$

where π_t and s_t denote the rate of inflation and the labour share, respectively, and ϵ_t is an innovation process. Suppose that $\mathcal{F}_t$ contains information up to time $t-1$. Equation (2.7) can be cast into a linear IV regression:

$$\pi_t = \lambda s_t + \gamma \pi_{t+1} + u_t, \tag{2.8}$$

which is estimated using instruments $Z_t \in \mathcal{F}_t$.

As we discuss in detail in chapter 4, the rational expectations model (2.7) is an incomplete description of the Data Generating Process for inflation. However, to discuss assumptions B to D we need to determine the covariance of the regressors (π_{t+1}, s_t) with the instruments. This can only be derived if a solution to the rational expectation model for inflation is found, and this in turn depends on the unspecified driving process for s_t. So, suppose s_t follows a stationary AR(2) process:

$$s_t = \rho_1 s_{t-1} + \rho_2 s_{t-2} + v_{1t} \tag{2.9}$$

where v_{1t} is assumed to be orthogonal to ϵ_t, e.g., $(\epsilon_t, v_{1t})' \sim NID(0, I_2)$. This equation can be combined with equation (2.7) to yield a solution to the latter. When $|\gamma_0| < 1$ a unique solution exists:

$$\pi_t = \tilde{a}_1\, s_{t-1} + \tilde{a}_2\, s_{t-2} + e_t \tag{2.10}$$

and the forecasting equation for π_{t+1} is:

$$\pi_{t+1} = a_1 s_{t-1} + a_2 s_{t-2} + v_{2t} \tag{2.11}$$

where v_{2t} is a function of $(\epsilon_{t+1}, v_{1,t+1}, v_{1,t})$ and (a_1, a_2) are functions of the model's parameters $(\lambda_0, \gamma_0, \rho_1, \rho_2)$.[11] The instrument set contains anything in $\mathcal{F}_t$, so consider the over-identifying set $Z_t = (s_{t-1}, s_{t-2}, \pi_{t-1}, \pi_{t-2})$. In terms of the notation in the previous example, $\theta = (\lambda, \gamma)'$, $Y_t = (s_t, \pi_{t+1})'$, $v_t = (v_{1t}, v_{2t})'$.

---

[10] This simple model is chosen over the slightly more complicated hybrid model proposed in that paper (the "New Keynesian Philips curve"), because it is more tractable and simplifies the discussion of the key ideas in this chapter. The hybrid version of the model is analyzed numerically in chapter 4.

[11] It is not difficult to verify that $a_1 = \rho_1 \tilde{a}_1 + \tilde{a}_2$ and $a_2 = \rho_2 \tilde{a}_1$ satisfy (2.7), with

$$\tilde{a}_1 = \frac{\lambda_0(\rho_1 + \gamma_0 \rho_2)}{1 - \rho_1\gamma_0 - \rho_2\gamma_0^2}, \qquad \tilde{a}_2 = \frac{\lambda_0 \rho_2}{1 - \rho_1\gamma_0 - \rho_2\gamma_0^2}$$

The solution to the system, equations (2.9) and (2.10) is a VAR(2) for (π_t, s_t), and hence assumptions B and C will be satisfied under stationarity and ergodicity. The moment conditions are

$$m(\theta) = -\Sigma_{ZZ} \begin{pmatrix} \rho_1 & a_1 \\ \rho_2 & a_2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} (\theta - \theta_0) = M(\rho_2)\,(\theta - \theta_0)$$

where $M(\cdot)$ is a matrix-valued function, whose rank depends on ρ_2 (note that Σ_{ZZ}, a_1, a_2 are all known functions of ρ_2). In particular, when $\rho_2 = 0$, a_2 also vanishes and $M(0)$ is clearly of reduced rank (≤ 1), giving an example of *partial* identification. Evidently, a crucial parameter governing the identification of θ is ρ_2.

To get an example of weak identification, in the sense of assumption D, we employ the device of linking the key parameter ρ_2 to the sample size, i.e., $\rho_2 = c/\sqrt{T}$. Then, taking a power series expansion of $M(\rho_2)$ around 0 to order $T^{-1/2}$ yields

$$m(\theta) = M(0)\,(\theta - \theta_0) + \left(\left.\frac{\partial M(\rho_2)}{\partial \rho_2}\right|_{\rho_2=0} (\theta - \theta_0) \right) c/\sqrt{T},$$

so that

$$m_2(\beta) = M(0)\,(\theta - \theta_0) = -\Sigma_{ZZ}^0 \begin{pmatrix} \rho_1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & a_0 \end{pmatrix} (\theta_0 - \theta)}_{(\beta_0 - \beta)} \tag{2.12}$$

where Σ_{ZZ}^0 is the variance of the instruments when $\rho_2 = 0$, $a_0 = \frac{\lambda_0 \rho_1}{1 - \gamma_0 \rho_1}$, that is, $\tilde{a}_1$ with $\rho_2 = 0$, and the identified *combination* of the parameters is thus $\beta(\theta) = (\lambda + a_0\,\gamma)$. Of course, any function of θ other than $\beta(\theta)$ would be acceptable as a basis for the un-identified direction $\alpha(\theta)$, so we may choose simply $\alpha = \gamma$. Thus, the re-parameterization $\kappa(\cdot)$ from the original parameters θ to α and β is:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \kappa(\theta) = \begin{pmatrix} 0 & 1 \\ 1 & a_0 \end{pmatrix} \begin{pmatrix} \lambda \\ \gamma \end{pmatrix},$$

which is clearly continuous and invertible, satisfying assumption D:

$$\phi(\alpha, \beta) = \kappa^{-1}(\alpha, \beta) = \begin{pmatrix} -a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{2.13}$$

Non-linear model

Finally, we present a variant of the above example to discuss the more general case where the re-parameterizing transformation $\phi(\cdot)$ is non-linear.

Using micro-foundations, Galí and Gertler (1999) express the parameter λ of the above model in terms of a 'deeper' parameter φ, measuring the degree of price inertia in the market, $\lambda = \frac{1}{\varphi}(1-\varphi)(1-\gamma\varphi)$. Setting $\theta = (\varphi, \gamma)$, we see that the analysis in the previous example follows exactly if we substitute $\frac{1}{\varphi}(1-\varphi)(1-\gamma\varphi)$ for λ. The change of parameters does not affect much the limiting expression for the moment conditions $m(\theta)$, but the re-parameterizing transformation into (α, β) and its inverse now become non-linear:

$$\kappa(\theta) = \begin{pmatrix} \gamma \\ \frac{1}{\varphi}(1-\varphi)(1-\gamma\varphi) + a_0\,\gamma \end{pmatrix}, \quad \phi(\alpha,\beta) = \begin{pmatrix} \frac{1+\beta+\alpha(1-a_0)\pm\sqrt{(1+\beta+\alpha-\alpha\,a_0)^2-4\,\alpha}}{2\alpha} \\ \alpha \end{pmatrix}$$

Evidently, without further restrictions, $\phi(\alpha, \beta)$ is not a function, violating the invertibility of $\kappa(\theta)$ condition in Assumption D.[12] To overcome the problem of lack of global identification, we can restrict attention to one of the two solutions for φ in terms of α, β in $\phi(\cdot)$. In particular, we choose $\varphi = \frac{1+\beta+\alpha(1-a_0)-\sqrt{(1+\beta+\alpha-\alpha a_0)^2-4\alpha}}{2\alpha}$, which by l'Hopital's rule is seen to be continuous at $\alpha = 0$ and can also lie in the unit interval, as required by the underlying economic theory for the friction parameter φ. Now, the re-parameterizing transformation becomes both continuous and invertible, satisfying Assumption D.

Finally, figure 2.1 plots the identified hyperplanes for the last two of the above examples, in θ-space. This is the locus of points over which the limiting GMM criterion function will be flat. Any point outside the locus can be ruled out asymptotically, but the points inside are indistinguishable based on the available information. This analysis helps to guide the choice over possible theoretically motivated restrictions. Two examples of affine restrictions are imposed on the linear model (figure 2.1 left). The dotted line represents a restriction that will yield identification, whereas the dashed line shows a mis-specifying restriction that doesn't help identify the parameters. Intuitively, a restriction is identifying if it does *not* restrict only the already identified combinations β. In other words, economic theory must be informative in the directions in which the data is not, if it is to be useful in identifying an otherwise un-identified model. Further examples of this will be discussed in chapter 4, with reference to the New Keynesian Phillips curve.

---

[12] This situation corresponds to a failure of global rather than local identification, i.e., when there exists a discrete number of points (2 in this case) in $\boldsymbol{\Theta}$ at which the moment conditions $m(\theta)$ vanish, see chapter 4 for details.

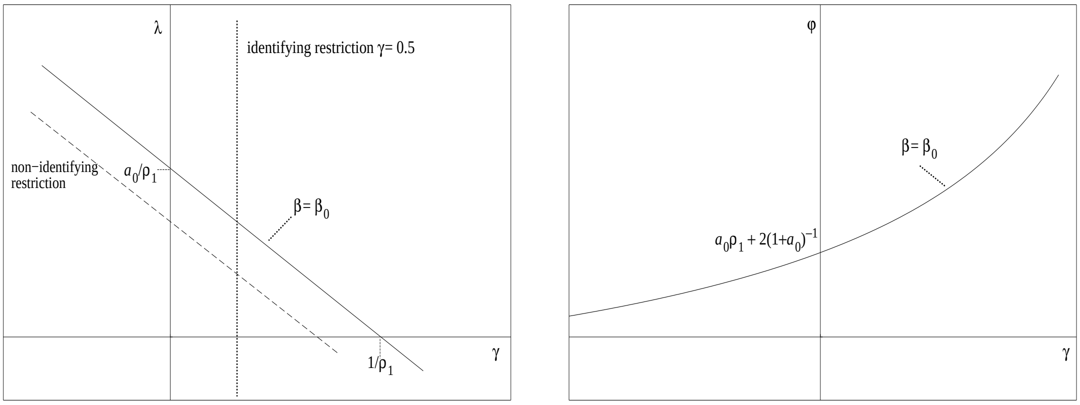


Figure 2.1: Identified hyperplanes for the examples in section 2.1.3.

2.1.4 Some important definitions

The concentration parameter and the degree of under-identification

In the context of linear simultaneous equations models, Sargan (1958) discussed the use of the minimum canonical correlation between the endogenous variables and the instruments as a measure of identifiability and instrument relevance.[13]

In models which are linear in the regressors, as it is the case for all the examples in this chapter, there exists a single small-dimensional summary of the strength of identification, known as the concentration ratio and denoted $\mu_T'\mu_T$ below, which was shown by Anderson (1977) to govern the speed of convergence of the estimators to their asymptotic distributions. In general, this quantity is a square matrix of dimension equal to the number of endogenous parameters. In a single-equation linear model of the form

$$\begin{aligned} y_t &= \theta' Y_t + u_t \\ Y_t &= \Pi' Z_t + v_t, \end{aligned}$$

with no heteroscedasticity or autocorrelation in u_t, and letting $\mathrm{var}(v_t) = \Sigma_{vv}$, the latter is given by:

$$(\mu_T'\mu_T) = T\,\Sigma_{vv}^{-1/2}\Pi'\,\Sigma_{ZZ}\,\Pi\Sigma_{vv}^{-1/2}. \tag{2.14}$$

This is essentially the proportion of the variation in the endogenous variables Y that is forecastable by the instruments Z, relative to the un-forecastable part v. In other words, it is a

---

[13] The use of canonical correlations for the selection of relevant instruments has recently been explored by Hall and Peixe (2001) and Hall and Inoue (2001). This is an issue we return to in section 2.5.2 below.

multivariate signal-noise ratio scaled up by the sample size, measuring the information embodied in the instruments for the parameters of interest. For consistency of the IV estimators of θ, we need this to be unbounded as the sample size grows (increasing information). However, when that matrix is of reduced-rank, we have a situation akin to a reduced-rank information matrix in Maximum Likelihood, which results in partial identification.

The eigenvalues of this matrix are also interpretable, since they relate one-to-one with the squared canonical correlations of the endogenous regressors with the instruments, ρ_i^2, $i = 1, \ldots, p$, which solve

$$|\Sigma'_{ZY}\, \Sigma_{ZZ}^{-1}\, \Sigma_{ZY} - \rho^2\, \Sigma_{YY}| = 0$$

Since $\Sigma_{YY} = \Sigma_{vv} + \Sigma'_{ZY}\, \Sigma_{ZZ}^{-1}\, \Sigma_{ZY}$, and $\Pi = \Sigma_{ZZ}^{-1}\, \Sigma_{ZY}$ we see the above can be written as

$$\left|\Pi'\, \Sigma_{ZZ}\, \Pi - \frac{\rho^2}{1-\rho^2}\, \Sigma_{vv}\right| = \left|\Pi'\, \Sigma_{ZZ}\, \Pi\, \Sigma_{vv}^{-1} - \frac{\rho^2}{1-\rho^2} I_p\right| = 0$$

so that

$$\mu_i^2 \equiv \text{eigen}(\mu'_T \mu_T) = \frac{T\, \rho_i^2}{1-\rho_i^2}$$

Under strong identification, those eigenvalues are $O(T)$, reflecting the increase in the information as the sample grows. However, under the weak instruments assumption D, a number p_1 of those eigenvalues will be $O(1)$, reflecting the weak identification in the directions implied by the corresponding eigenvectors. Also, identification is stronger the higher the canonical correlations of the regressors with the instruments. Thus, intuitively, the presence of k_1 exogenous regressors in the model ($\rho_i = 1$, for some $i = 1, \ldots k_1$) will mean that an equal number of the eigenvalues of the concentration matrix will be 'infinite'.[14]

In our analysis below, we use a generalization of the above definition (2.14) to account for possible serial correlation and heteroscedasticity in the residuals u_t:

$$(\mu'_T \mu_T) = T\, \sigma_u^2\, \Sigma_{vv}^{-1/2} \Pi'\, \Sigma_{ZZ}\, V(\theta_0)^{-1}\, \Sigma_{ZZ}\, \Pi \Sigma_{vv}^{-1/2} \tag{2.15}$$

where the scaling factor σ_u^2 is the unconditional variance of u_t, when it is constant over time, or it can be interpreted as an average variance when u_t is unconditionally heteroscedastic. When there is no heteroscedasticity or serial correlation ($V(\theta_0) = \sigma_u^2\, \Sigma_{ZZ}$), (2.15) simplifies to (2.14).

---

[14] When exogenous regressors are present in the model (Y_t replaced by $\bar{Y}_t = (Y'_t, Z'_{1t})'$, see section 3.1), the above measure cannot be computed, since the variance of the first-stage error is singular. However, the generalized eigenvalues of the matrix pencil $|T\, \Pi'\, \Sigma_{ZZ}\, \Pi - \mu\, \Sigma_{vv}|$ will serve as an alternative, and intuitive measure of the strength of identification, see discussion in appendix 3.B.

Distinction between instrument validity and instrument relevance

The fact that an instrument is 'admissible' in the information set does not necessarily imply that it is informative. Unfortunately, this distinction is not always clear in empirical work, namely oftentimes instrument validity and instrument relevance are confused. As this distinction is at the heart of the weak instruments problem, and is especially relevant in dynamic rational expectations models, it is important to emphasize it.

Definition 2.3 (Instrument validity). *A variable* $Z_{i,t}$ *is a* valid *instrument for the estimation of a structural model if it does not violate its moment conditions* $\mathrm{E}\left[h(\mathrm{y}_t,\, \theta_0) \otimes Z_{i,t}\right] = 0$.

Definition 2.4 (Instrument relevance). *A variable* $Z_{i,t}$ *is a* relevant *instrument for the estimation of a structural model if using it would increase at least one of the eigenvalues of the concentration matrix (equivalently increase at least one of the canonical correlations between the instruments and endogenous regressors). Otherwise,* $Z_{i,t}$ *is irrelevant.*

Also, within the set of relevant instruments, we would like to distinguish between the ones that would help identify an otherwise weakly identified model, denoted 'strongly relevant', and the ones that are informative only about the already identified parameter combinations, called 'weakly relevant'.

Definition 2.5 (Strong versus weak relevance). *For a weakly identified model, where the minimum eigenvalue of the concentration matrix* μ_{min} *is small, that is* $O(1)$*, an instrument is* strongly relevant *if its inclusion would increase at least* μ_{min}*, whereas it is* weakly relevant *if it is relevant, but does not increase* μ_{min}*.*

Similarly, we need to define precisely what we mean by the degree of over-identification. In other words, we need to distinguish between *nominal* and *actual* over-identification.

Definition 2.6 (Nominal and actual over-identification). *The nominal over-identification of a model is the difference between the number of instruments used and the number of estimated (free) parameters,* $Gk - p$*. The* actual *degree of over-identification is nominal over-identification less the number of irrelevant instruments. When this number is negative, it measures the* **degree of under-identification.**[15]

---

[15] A model whose degree of under-identification is greater than 0 is also referred to as partially identified.

2.2 General asymptotic theory

In this section, we summarize the results in the literature on the asymptotic distributions of GMM estimators under weak identification, the behaviour of the limiting objective function, and hence, the distribution of tests based on it.

2.2.1 Estimator distributions, and the limiting objective function

The first result is an immediate extension of Stock and Wright (2000, Theorem 1). Let $\widehat{\beta}(\alpha) = \operatorname{argmin}_{\beta \in B} \mathcal{Q}_T\left(\phi(\alpha,\beta); \bar{\theta}\left(\phi(\alpha,\beta)\right)\right)$, let $\widehat{\alpha} = \operatorname{argmin}_{\alpha \in A} \mathcal{Q}_T\left(\phi(\alpha,\widehat{\beta}(\alpha)); \bar{\theta}\left(\phi(\alpha,\widehat{\beta}(\alpha))\right)\right)$, and let $\widehat{\beta} = \widehat{\beta}(\widehat{\alpha})$. Proofs of all theorems are relegated to the mathematical appendix 2.A at the end of this chapter.

Theorem 2.1. *Suppose that assumptions A.1 to A.5 and B, C and D hold, and that* $\bar{\theta}_T(\theta) \Rightarrow \bar{\theta}(\theta)$ *uniformly in* θ*. Then:*

$$\begin{aligned}
\mathcal{Q}_T\left(\phi(\alpha,\beta_0+b/\sqrt{T}); \bar{\theta}_T(\phi(\alpha,\beta_0))\right) &\Rightarrow \left[\Psi(\phi(\alpha,\beta_0)) + m_1(\phi(\alpha,\beta_0)) + \mathcal{J}_2(\beta_0)\, b\right]' \\
&\quad \times W\left(\bar{\theta}(\phi(\alpha,\beta_0))\right) \left[\Psi(\phi(\alpha,\beta_0)) + m_1(\phi(\alpha,\beta_0)) + \mathcal{J}_2(\beta_0)\, b\right] \\
&\equiv \bar{\mathcal{Q}}\left(\alpha, b\,;\alpha\right)
\end{aligned}$$

(ii) If $\bar{\mathcal{Q}}\left(\alpha, b; \alpha\right)$ *has a unique minimum over* $(\alpha', b')'$*, then* $\left(\widehat{\alpha},\ \sqrt{T}(\widehat{\beta}-\beta_0)\right) \xrightarrow{d} (\alpha^*,\ b^*)$*, where* $\alpha^* = \operatorname{argmin}_{\alpha \in A} \mathcal{Q}^*\left(\alpha\right)$ *and where*

$$\begin{aligned}
b^* &= -\left[\mathcal{J}_2(\beta_0)'\tilde{W}(\alpha^*)\mathcal{J}_2(\beta_0)\right]^{-1} \mathcal{J}_2(\beta_0)'\tilde{W}(\alpha^*)\left[\Psi(\phi(\alpha^*,\beta_0)) + m_1(\phi(\alpha^*,\beta_0))\right], \\
\mathcal{Q}^*\left(\alpha\right) &= \left[\Psi(\phi(\alpha,\beta_0)) + m_1(\phi(\alpha,\beta_0))\right]' M(\alpha) \left[\Psi(\phi(\alpha,\beta_0)) + m_1(\phi(\alpha,\beta_0))\right], \\
M(\alpha) &= \tilde{W}(\alpha) - \tilde{W}(\alpha)\mathcal{J}_2(\beta_0)\left[\mathcal{J}_2(\beta_0)'\tilde{W}(\alpha)\mathcal{J}_2(\beta_0)\right]^{-1} \mathcal{J}_2(\beta_0)'\tilde{W}(\alpha), \\
\tilde{W}(\alpha) &= W\left(\bar{\theta}(\phi(\alpha,\beta_0))\right)
\end{aligned}$$

(iii) Hence, a first-order asymptotic approximation to the distribution of $\widehat{\theta}$ *is given by* $\phi(\alpha^*,\beta_0)$*.*

Remarks In the special case that $\phi(\cdot)$ is the identity mapping, $\theta = (\alpha,\beta)$ and theorem 2.1 simplifies to Stock and Wright (2000, Theorem 1). The GMM estimators of the identified parameters $\widehat{\beta}$ are $\sqrt{T}$-consistent, whereas the estimators of the weakly identified parameters $\widehat{\alpha}$ are $O_p(1)$. In general, however, all of the parameter estimators will be $O_p(1)$ and therefore inconsistent. Evidently, only when $\theta_i = \beta_i$ for some parameters i, is $\widehat{\theta}_i$ $\sqrt{T}$-consistent for θ_i.

Theorem 2.1 can be used to simulate the asymptotic distribution of various GMM estimators of θ and related test statistics, see chapter 3. It is general enough to cover all of the known distributional results in the IV and GMM literature so far. The results given in Phillips (1983) for the well-identified model, those in Phillips (1989) and Choi and Phillips (1992) for the partially identified model, as well as Staiger and Stock (1997) for the linear weakly identified model, are special cases of this, as emphasized in Stock and Wright (2000).

The asymptotic distribution evidently depends on nuisance parameters. The dependence on the number of moment conditions Gk is apparent from the fact that the limiting distribution involves a quadratic form of (Gk) independent standard normal variates. Other nuisance parameters arise from $m_1(\alpha)'m_1(\alpha)$ or whenever the limit of the weighting matrix does not coincide with the inverse of the variance of Ψ at $\theta^* \neq \theta_0$.

The asymptotic distribution under (local) mis-specification (assumption D$'$) follows straightforwardly from theorem 2.1:

Corollary 2.2. *If we replace assumption D with D$'$, and assume further that there exists a unique $\beta_* \in B$ such that $m_2(\beta_*) + \zeta_f = 0$ and $\mathcal{J}_2(\beta_*)$ has full rank p_2, then the conclusions of theorem 2.1 remain valid if we substitute β_* for β_0 and $[m_1(\cdot, \beta_*) + \zeta_l]$ for $m_1(\cdot, \beta_0)$.*

In other words, if the mis-specification is assumed to be 'local-to-zero', it will manifest itself partly as a bias on the well-identified parameters $(\beta_* - \beta_0)$, which is undetectable, and partly as an additional non-centrality in the limiting objective function.

Note that the above results also characterize the special cases where all the parameters are well-identified $(\beta = \theta)$, or when some of the parameters are completely un-identified $(m_1(\alpha) = 0,\ \forall \alpha \in A)$. In the well-identified case, a further interesting implication of corollary 2.2 is that, even when $\beta_* = \beta_0$, in which case the GMM estimator of β will be $\sqrt{T}$-consistent, as usual, the presence of mis-specification $(\zeta_l \neq 0)$ will give rise to an $O_p(T^{-1/2})$ bias to the estimator. This highlights the fact that not all of the mis-specification will be detectable, a point which formalized in lemma 2.7 below.

It is worth emphasizing that different GMM estimators that would be asymptotically equivalent under strong identification are no longer so in the cases of weak or partial identification. This is true both for estimators using different weighting matrices, e.g., HAC versus non-robust variance estimators, and for different *types* of GMM estimators using the *same* weighting matrix, e.g., 2-step (2S) versus continuously updated (CUE) estimators. The standard asymptotic equivalence argument breaks down when we depart from strong identification simply because the different variance estimators are usually not asymptotically equivalent at any point other than θ_0. Since θ

is no longer confined to a neighbourhood of θ_0, the limiting objective functions corresponding to different estimators will not be the same, in general.

2.2.2 Distribution of conventional test statistics

Standard GMM test statistics include Wald (W), Likelihood Ratio (LR) and Score-type (LM) tests, as well as the Hansen-Sargan J-statistic of over-identifying restrictions.
Letting $\mathbf{S}_T(\theta) = D_T(\theta)' \left[V_T(\theta)\right]^{-1} D_T(\theta)$, we have:[16]

$$\begin{aligned} \text{Wald} &= T\,(\widehat{\theta} - \theta_0)' \left[\mathbf{S}_T(\bar{\theta}(\widehat{\theta}))\right] (\widehat{\theta} - \theta_0) \\ LR &= \mathcal{Q}_T(\theta_0; \bar{\theta}(\theta_0)) - \mathcal{Q}_T\left(\widehat{\theta}; \bar{\theta}(\widehat{\theta})\right) \qquad (2.16) \\ LM &= T\, g_T(\theta_0)' V_T^{-1} D_T \left[\mathbf{S}_T(\theta_0)\right]^{-1} D_T' V_T^{-1} g_T(\theta_0) \qquad (2.17) \end{aligned}$$

Under strong identification, these tests are asymptotically χ^2 with degrees of freedom equal to the number of parameters. However, in the more general case, it is now well-known that these statistics are not asymptotically pivotal, and hence their use to draw inferences can be highly misleading, see Dufour (1997), Staiger and Stock (1997), Wang and Zivot (1998). All statistics except for the score test are functions of the GMM estimator $\widehat{\theta}$, whose distribution under weak instruments depends on unknown nuisance parameters, see theorem 2.1.

From that theorem, it is straightforward to derive the distribution of the above statistics, as is done for example in Stock and Wright (2000, Corollary 4). However, as Wang and Zivot (1998) also point out, the resulting asymptotic approximations are not of much use as they stand, because they require more assumptions about the underlying data generating process than researchers are willing to make, and these are specific to every given application (e.g., our example 2 above, where we have specified the source of weak identification as being the local-to-zero second order dynamics in the exogenous process s_t).

Some useful intuition about the non-standard distributions of the test statistics is offered by Pagan and Robertson (1997), who observe that the problem arises mainly due to these statistics involving “division” by a matrix that is random asymptotically, unlike the the well-identified case where it is fixed. This is true of all statistics but the score test.

Moreover, Dufour (1997) shows that any Wald-based confidence sets are hopeless in the case of weak identification, because they have zero coverage probabilities. To see this, consider a completely unidentified model. In that case, the distribution of any GMM estimator $\widehat{\theta}$ is independent

---

[16] The criterion function $\mathcal{Q}_T$ is already scaled by T in our notation, see table 2.1.

of the true parameter value θ_0, say, reflecting the fact that there is no information in the data about θ_0. A correct α-level Wald test does not exist, because we can always find a value for $\theta_0 \in \mathbf{\Theta}$ such that the test rejects with arbitrarily high probability. (The level of a test is the maximal rejection probability of the test over all possible null hypotheses $\theta_0 \in \mathbf{\Theta}$.) A confidence set based upon inverting a Wald test will therefore have zero coverage.

The above discussion is not applicable to the score (LM) test, since it does not involve any function of $\widehat{\theta}$. That does not mean, however, that it is free from problems. One apparent problem is that it uses the inverse of $\mathbf{S}_T$, which gets very large asymptotically when some parameters are weakly identified. This is because the average moment Jacobian D_T converges to a reduced rank matrix $\mathcal{J}$ under the weak identification assumption D, and therefore $\mathbf{S}_T$ also tends to a reduced rank matrix, causing its inverse to become unbounded. However, a closer look reveals that $D_T(\theta_0)\left[D_T(\theta_0)'V(\theta_0)^{-1}D_T(\theta_0)\right]^{-1}D_T(\theta_0)'$ will not blow up as T increases, and hence the reduced rank of the limiting moment Jacobian cannot be the source of the problem.[17]

The problem mainly occurs because when the limiting Jacobian is of reduced rank, the distribution of the vector $D_T(\theta_0)'V(\theta_0)^{-1}g_T(\theta_0)$ is not Gaussian with zero mean as in the full rank case, but rather a non-central mixed Gaussian (i.e., random mixture of normals), with both location and scale parameters being random, see lemma 2.8 in appendix 2.A. Central limiting mixed Gaussian distributions are common in econometrics (see e.g., Johansen (1995) for tests on the cointegrating space) and they do not pose any problem for inference, because provided we can condition on the random variance parameter, the resulting quadratic form is still asymptotically χ^2. However, a random non-centrality will suffice to render the resulting quadratic form non-standard, see e.g., the simple univariate example in Staiger and Stock (1997, section 2).[18]

2.2.3 Similar tests and robust inference

The above discussion shows that the distributions of the test statistics are non-standard and non-pivotal w.r.t. nuisance parameters, hence rendering them practically useless for empirical work. There are three ways to deal with this problem. One approach is to look for alternative

---

[17] Using the singular value decomposition of the limiting Jacobian, $\mathcal{J} = S\Lambda U'$, where Λ is a diagonal matrix holding the p_2 non-zero singular values of $\mathcal{J}$, and S, U are $Gk \times p_2$ and $p \times p_2$ dimensional semi-orthogonal matrices respectively, we see that:

$$\mathcal{J}\left(\mathcal{J}'V^{-1}\mathcal{J}\right)^{-1}\mathcal{J} = S\Lambda U'\left(U\Lambda S'V^{-1}S\Lambda U'\right)^{-1}U\Lambda S' = S\left(S'V^{-1}S\right)^{-1}S'$$

which is bounded. (We used the generalized inverse to derive this result.)

[18] Note also that the quadratic form for the LM test is not using the correct asymptotic variance found in lemma 2.8, i.e., $\mathbf{S} \neq v(\tilde{\Xi})$, which will complicate the resulting asymptotic distribution further.

statistics which are asymptotically pivotal. Another is to look for particular versions of the conventional statistics which are boundedly pivotal, upon which similar but conservative tests can be based. A third approach is to modify the critical values of non-pivotal statistics in order to derive asymptotically similar tests.

The first approach has lead researchers to propose statistics of the type of Anderson and Rubin (1949). Such inferential procedures derive valid confidence sets based on inverting the GMM objective function at various parameter points. Moreira (2001) provides a characterization of the family of pivotal tests in the context of linear IV regression.

An important result is Stock and Wright (2000, Theorem 2), which states that the limiting distribution of the objective function for the continuously updated GMM estimator (CUE), when evaluated at the true parameter point θ_0 is asymptotically χ^2, with degrees of freedom equal to the total number of instruments, Gk.

Theorem 2.3 (Stock and Wright). *If* $\Psi_T(\theta_0) \xrightarrow{d} N(0, V(\theta_0))$, $W_T(\theta_0) \xrightarrow{p} V(\theta_0)^{-1}$ *and the model is correctly specified* $m(\theta_0) = 0$, *then:*

$$\mathcal{Q}_T(\theta_0; \theta_0) \xrightarrow{d} \chi^2(Gk)$$

This result provides the basis for constructing valid confidence intervals for θ. In particular, Stock and Wright (2000) proposed a confidence set called "S-set", which has correct coverage rate asymptotically under the null hypothesis, irrespective of the failure of identification:

Definition 2.7 (S-set). *A confidence set of* $1-\alpha$ *coverage rate for* θ *is given by:*

$$S_\alpha = \left\{\theta \in \boldsymbol{\Theta} \mid \mathcal{Q}_T(\theta; \theta) \leq \chi^2_\alpha(Gk)\right\}$$

In order to illustrate the implications of weak identification in situations when only functions of the parameters θ are well-identified, we computed a 90% confidence S-set for the parameters $\theta = (\lambda, \gamma)'$ of the pure forward-looking New Phillips Curve, example 2 in section 2.1.3. The S-set is plotted in θ-space in figure 2.2, compare with left panel of figure 2.1. The graph suggests that this set is potentially unbounded, which is consistent with the weak identification. Upon comparison with figure 2.1 above, we see that the confidence set comprises entirely the well-identified hyperplane – which, for the parameters reported in Galí and Gertler (1999, Table 1), is the line $\lambda = 0.27 - 0.25\,\gamma$, and highlights the relationship between the θ-space and the (α, β)-space.

Another asymptotically pivotal test has been recently proposed by Kleibergen (2001). This

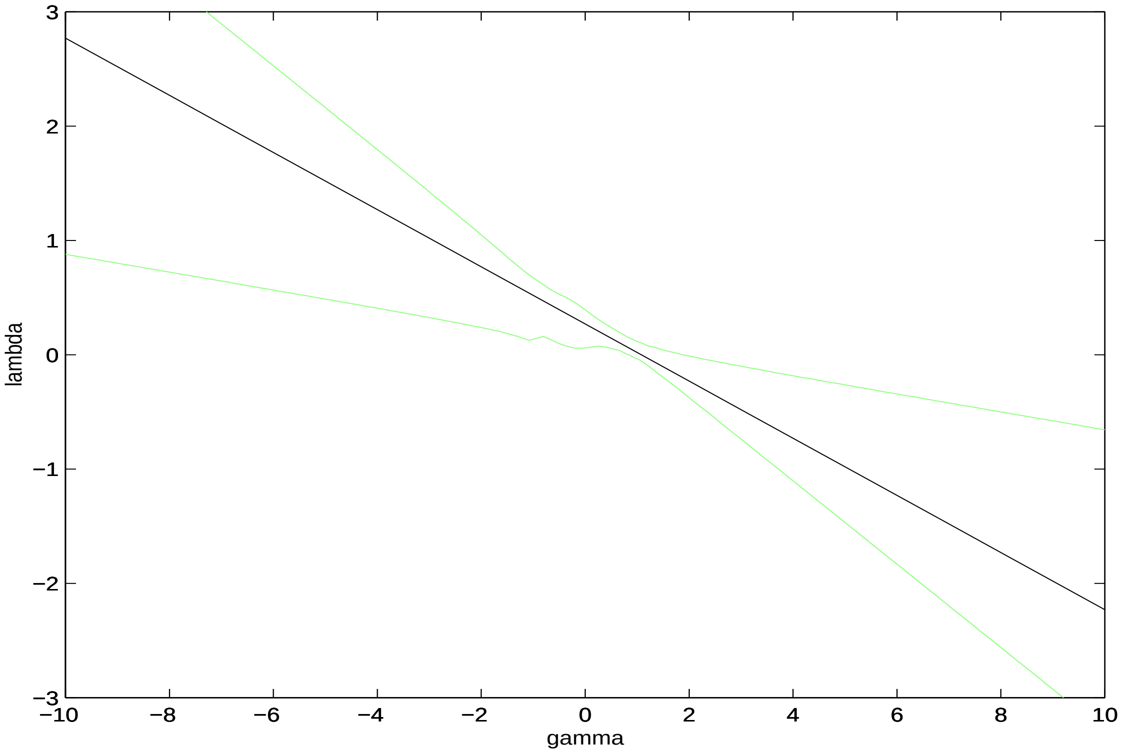


Figure 2.2: 90% S-set for the parameters (λ, γ) of the New Keynesian Phillips curve, example 2 in section 2.1.3 (based on simulated data).

score-like statistic is based on the full derivative of the GMM objective function:

$$\frac{1}{2}\frac{\partial \mathcal{Q}(\theta_0)}{\partial \theta'} = g_T(\theta_0)'V_T(\theta_0)^{-1}D_T(\theta_0) - \frac{1}{2}\left(g_T(\theta_0)'V_T(\theta_0)^{-1} \otimes g_T(\theta_0)'V_T(\theta_0)^{-1}\right)\frac{\partial \text{vec}\left(V_T(\theta_0)\right)}{\partial \theta'}$$

Kleibergen observed that in the general case, when $\mathcal{J}$ is possibly of reduced rank, the second term does not vanish asymptotically, and that, upon replacing V_T by its limit and dropping the dependence on θ_0, the above expression can be re-written as:

$$g_T'V^{-1}\underbrace{\left[D_T - \left(\mathcal{C}_1 V^{-1} g_T, \ldots, \mathcal{C}_p V^{-1} g_T\right)\right]}_{D_T^\perp}$$

where $\mathcal{C}_i = \text{cov}\left[g_T(\theta_0), \frac{\partial g_T(\theta_0)}{\partial \theta_i}\right]$ and thus $D_T^\perp$ is the residual of the projection of the moment Jacobian D_T on the moment conditions g_T. Therefore, it is evident that, by construction, $D_T^\perp$ is asymptotically independent of g_T, so that the asymptotic distribution of the above expression scaled by $\sqrt{T}$, conditional on $D_T^\perp$, is p-dimensional Gaussian with mean 0 and variance $(D_T^\perp)'V^{-1}D_T^\perp$, see Kleibergen (2001, Theorem 1). Exploiting this limiting conditional (central) Normality, and replacing V_T and $D_T^\perp$ with some consistent estimators, we can compute a quadratic form that will be asymptotically χ^2 with degrees of freedom equal to the number of parameters p:

$$K_T(\theta_0) = T\ g_T(\theta_0)'\widehat{V}_T(\theta_0)^{-1}\widehat{D}_T^{\perp}(\theta_0)\left[\widehat{D}_T^{\perp}(\theta_0)'\widehat{V}_T(\theta_0)^{-1}\widehat{D}_T^{\perp}(\theta_0)\right]^{-1}\widehat{D}_T^{\perp}(\theta_0)'\widehat{V}_T(\theta_0)^{-1}g_T(\theta_0) \tag{2.18}$$

Straightforwardly, we may then define the following asymptotically pivotal confidence set:

Definition 2.8 (K-set). *A confidence set of* $1-\alpha$ *coverage rate for* θ *is given by:*

$$K_\alpha = \left\{\theta \in \boldsymbol{\Theta} \mid K_T(\theta) \leq \chi^2_\alpha(p)\right\}$$

Next, we give a result in the spirit of the second approach of looking for versions of the classical tests which turn out to be asymptotically (boundedly) pivotal. Wang and Zivot (1998) observed that, in the context of a linear IV model, the LR statistic based on the LIML estimator, and the LM statistic given above are boundedly pivotal, that is, bounded by some pivotal random variable (whose distribution is similar w.r.t. nuisance parameters). The following result generalizes Wang and Zivot (1998, Theorem 1) for GMM:

Theorem 2.4. *Let* $\widehat{\theta}_{CUE}$ *be the continuously updated GMM estimator defined as:*

$$\widehat{\theta}_{CUE} = \underset{\theta\in\boldsymbol{\Theta}}{\operatorname{argmin}}\ \mathcal{Q}_T(\theta;\theta),$$

and denote the LR test evaluated at $\widehat{\theta}_{CUE}$ *by* $LR_{CUE} = \mathcal{Q}_T(\theta_0;\theta_0) - \mathcal{Q}_T(\widehat{\theta}_{CUE};\widehat{\theta}_{CUE})$. *Under the assumptions of theorem 2.1,*

1. *If* $W_T(\theta_0) \xrightarrow{p} V(\theta_0)^{-1}$*:*

   $$LM \xrightarrow{d} \int_{\tilde{\Xi}} z(\tilde{\Xi})'z(\tilde{\Xi})\ \mathrm{d}F(\tilde{\Xi}),$$

   where $z(\cdot)$ *conditional on* $\tilde{\Xi}$ *is* p*-dimensional Gaussian with mean* $m(\tilde{\Xi})$ *and variance* $v(\tilde{\Xi})$*, as defined in lemma 2.8 in the appendix 2.A.*

2. *If* $W_T(\theta) \xrightarrow{p} V(\theta)^{-1}$ *uniformly in* $\theta \in \boldsymbol{\Theta}$*:*

   $$LR_{CUE} \Rightarrow \tilde{\mathcal{Q}}(\alpha_0, 0\,;\alpha_0) - \tilde{\mathcal{Q}}(\alpha^*, b^*;\alpha^*)$$

   where $\bar{\mathcal{Q}}(\cdot,\,\cdot\,;\,\cdot)$*,* α^* *and* b^* *are given in theorem 2.1*

3. *When* $Gk = p$*, the above distributions reduce to* $\chi^2(p)$*; when* $Gk > p$ *(over-identification), the above distributions are bounded from above by a* $\chi^2(Gk)$*.*

Part 1 of the theorem says that provided a consistent estimator of the limiting variance of the moment conditions is used, the asymptotic distribution of the LM statistic will be a weighted sum of p non-central $\chi^2(1)$ r.v.s, where both the weights and the non-centrality parameters are random functions involving the nuisance parameters. Under a somewhat stronger condition on the weighting matrix, we derive the distribution of the LR statistic when the CUE is used to estimate θ. In the special case of a linear IV model, the CUE reduces to the LIML estimator, and Wang and Zivot (1998, Theorem 1) is a special case of theorem 2.4.

The last part of the theorem shows that both the LM and the LR_{CUE} statistics will be boundedly pivotal, so they could be used to derive valid confidence sets for the parameters. However, they will be conservative in general, i.e., the resulting confidence intervals will have a coverage rate which is bigger than the nominal one, and will be less powerful than both the S-sets and the K-sets, discussed above.

Note that the conclusions of theorem 2.4 part 3 do not hold for LR_{2SLS}, thus rendering the latter inappropriate under weak instruments. To see this, observe that in the definition of the LR statistic, equation (2.16), $\bar{\theta}(\theta_0) = \widehat{\theta}^{(1)}$, the first-step parameter estimate. It is precisely the dependence on $\widehat{\theta}^{(1)}$ that renders the statistic non-pivotal.

Finally, we comment briefly on the latest development in the spirit of the third approach of basing inference on non-pivotal statistics, by deriving (asymptotically) correct critical values. The only contribution in this direction, to date, is by Moreira (2002), who proposed that inference be based on a conditional Likelihood Ratio statistic for linear IV models. The key idea is that, despite the fact that this statistic is non-pivotal under weak instruments, it can be decomposed into two asymptotically independent quantities $\mathcal{S}$ and $\mathcal{T}$, only the latter depending on the nuisance parameters. Hence, the distribution of the test statistic $\tau(\mathcal{S}, \mathcal{T})$, conditional on $\mathcal{T}$, is similar w.r.t. any nuisance parameters. Correct critical values can then be derived by solving $\Pr(\tau(\mathcal{S}, \widehat{\mathcal{T}}) < c_\alpha) = 1 - \alpha$, given the distribution of $\mathcal{S}$ and the value of $\mathcal{T}$ observed in the sample. Since the above problem doesn't admit an analytical solution, in general, the critical values have to be determined by Monte Carlo simulation.

Moreira (2002) showed that his conditional likelihood ratio test compares favourably to the Anderson-Rubin and the K-statistic (score test), and its power is close to the unfeasible power envelop for similar tests that he derived in Moreira (2001). Unfortunately, his results are rather limited in scope, since they apply only to the very special case of linear models with no heteroscedasticity or autocorrelation. Moreover, unlike the other two similar tests, it is not clear how this test can be extended to GMM.[19]

Remarks Despite their appeal, the above procedures have a number of drawbacks. First of all, when used to test the null hypothesis $\mathrm{H}_0 : \theta = \theta_0$, none of the above statistics is consistent against any fixed alternative on the un-identified parameters under weak instrument asymptotics. Consider, for instance, the generic alternative $\mathrm{H}_1 : \alpha = \alpha_1, \beta = \beta_0 + O(T^{-1/2})$. It is easy to see that the above asymptotic distributions will become (bounded by) non-central χ^2. But the non-centrality parameter will not be of order T under such fixed alternatives, but will be finite in general, e.g., $m_1(\theta_1)'V(\theta_1)^{-1}m_1(\theta_1)$ for the Anderson-Rubin statistic (S-set), where $\theta_1 = \phi(\alpha_1, \beta_0)$ is the parameter value under the alternative H_1.[20]

Secondly, all of the above results are valid conditional on the model being correctly specified, in the sense of assumption D. Under mis-specification, assumption D$'$, the above distributions will become non-central again, in general causing the resulting confidence sets to be tighter than their nominal size. In the case of the S-sets, it is possible that they become empty, and that could safely be interpreted as evidence of mis-specification. However, if that does not happen to be the case, tight confidence sets may be incorrectly interpreted as evidence of more accurate estimation. Thus, in the absence of any valid test of the over-identifying restrictions, the utility of the above procedures is seriously undermined.

2.3 GMM covariance matrix estimation

In order to write down an efficient GMM criterion (2.2) for the estimation of θ, we need an estimator of the asymptotic variance matrix of the moment conditions, defined by:

$$V(\theta) = \lim_{T\to\infty} \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(f_t(\theta) - \mathrm{E}(f_t(\theta)|\mathcal{F}_t)\right)\left(f_s(\theta) - \mathrm{E}(f_s(\theta)|\mathcal{F}_s)\right)'\right] \tag{2.19}$$

Estimators of (2.19) that allow for possible heteroscedasticity and/or autocorrelation in the residuals are known as Heteroscedasticity and/or Autocorrelation Consistent covariance matrix estimators, abbreviated by HC, AC or HAC. A number of different estimators have been proposed in the econometrics literature, see White (1982), Newey and West (1987), Andrews (1991), Andrews and Monahan (1992), den Haan and Levin (1997) and West (1997). These estimators

---

[19] See also Stock, Wright, and Yogo (2002) for more discussion of this and related approaches to inference in GMM with weak instruments.

[20] This result could be formalized in line with Staiger and Stock (1997, Theorem 5b).

can be divided in two groups: non-parametric kernel-based procedures that allow for autocorrelation of unknown form; and parametric ones, that make prior assumptions about the structure of autocorrelation in the moment conditions f_t. Examples of non-parametric estimators are the ones proposed in Newey and West (1987), Andrews (1991) and Andrews and Monahan (1992), whereas examples of parametric estimators are those in den Haan and Levin (1997) and the MA-l estimator proposed by West (1997). The latter is particularly suitable for the dynamic rational expectations models that we consider here.

The following alternative definitions of $V(\theta)$ are used for estimation:

$$V_i(\theta) = \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \left[f_t(\theta) - C_t^i\right] \left[f_s(\theta) - C_s^i\right]' \tag{2.20}$$

where C_t^i is a centering that corrects for non-orthogonality, in particular:

$$\begin{array}{ll} C_t^n = 0 & \text{No centering} \\ C_t^u = \mathrm{E}\left(f_t(\theta)\right) & \text{Unconditional centering} \\ C_t^c = \mathrm{E}(f_t(\theta)|\mathcal{F}_t) & \text{Conditional centering} \end{array}$$

These are clearly identical at $\theta = \theta_0$ under the null of correct specification, since $\mathrm{E}(f_t(\theta_0)|\mathcal{F}_t) = \mathrm{E}f_t(\theta_0) = 0$, and they correspond to $V(\theta)$ when $f_t(\theta)$ is *ergodic* for its second moment. However, this equivalence breaks down under mis-specification or at any other $\theta \neq \theta_0$.[21] In that case, only V_c corresponds to the desired quantity V. Let $\bar{f}_t(\theta) = f_t(\theta) - \mathrm{E}(f_t(\theta)|\mathcal{F}_t)$ and $\bar{h}_t(\theta) = h_t(\theta) - \mathrm{E}(h_t(\theta)|\mathcal{F}_t)$. Then:

$$\begin{aligned} V_c(\theta) &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \bar{f}_t(\theta)\bar{f}_s(\theta)' \\ &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left[\bar{f}_t(\theta)\bar{f}_t(\theta)'\right] + \operatorname*{plim}_{T\to\infty} \sum_{j=1}^{T-1} \frac{1}{T} \sum_{t=j+1}^{T} \left[\bar{f}_t(\theta)\bar{f}_{t-j}(\theta)' + \bar{f}_{t-j}(\theta)\bar{f}_t(\theta)'\right] \\ &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left[\bar{f}_t(\theta)\bar{f}_t(\theta)'\right] + \lim_{T\to\infty} \sum_{j=1}^{T-1} \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=j+1}^{T} \left[\bar{f}_t(\theta)\bar{f}_{t-j}(\theta)' + \bar{f}_{t-j}(\theta)\bar{f}_t(\theta)'\right] \\ &= \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left[\bar{h}_t(\theta)\bar{h}_t(\theta)' \otimes Z_t Z_t'\right] + \\ &\quad + \lim_{T\to\infty} \sum_{j=1}^{T-1} \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=j+1}^{T} \left[\bar{h}_t(\theta)\bar{h}_{t-j}(\theta)' \otimes Z_t Z_{t-j}' + \bar{h}_{t-j}(\theta)\bar{h}_t(\theta)' \otimes Z_{t-j} Z_t'\right] \end{aligned} \tag{2.21}$$

A comment on the above three versions is in order. Even though they are equivalent under the null of correct specification and $\theta = \theta_0$, the different versions can have important implications for the power of the test of over-identifying restrictions, i.e., under the alternative that the model is mis-specified. Recently, Hall (2000) has shown that the use of the uncorrected covariance will lead

---

[21] This is relevant when, for instance, θ is not consistently estimable.

to a version of the Hansen-Sargan test that is asymptotically dominated by the version in which unconditional centering was used, in the sense that the former grows at a slower rate than the latter under a fixed alternative. The intuition for this result is that when the moment conditions are violated, V_n would unambiguously overestimate their true variance compared with V_u (or even V_c, but Hall didn't analyze this), and therefore the Hansen-Sargan test which uses their inverse will tend to be smaller for the uncorrected version, as the sample gets large. This conclusion holds only if the model is identified in the sense of definition 2.1, in which case $\widehat{\theta} \xrightarrow{p} \theta_*$, some pseudo true value.

A similar argument for conditional versus unconditional centering is not easy to make. To see this, consider how V_n and V_u related to the correct version V_c given in (2.21). Assuming stationarity, such that $\mathrm{E} f_t = \mathrm{E} f_s$:

$$\begin{aligned} V_n(\theta) &= V_u(\theta) + \underbrace{\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{E} f_t(\theta)\, \mathrm{E} f_t(\theta)'}_{PSD} \\ V_u(\theta) &= V_c(\theta) + \underbrace{\operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} d_t\, d_s'}_{PSD} \\ &\quad + \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \bar{f}_t\, d_s' + \operatorname*{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \bar{f}_s\, d_t' \end{aligned}$$

where $d_t = \mathrm{E}(f_t|\mathcal{F}_t) - \mathrm{E} f_t$.[22] Even though the uncorrected version is undoubtedly bigger than the un-conditionally corrected one, confirming our intuition, the difference $V_u - V_c$ is not unambiguously Positive Semi-Definite. The term involving the infinite sum of cross products $d_t\, d_s'$ will be PSD by construction since it equals $\operatorname*{plim}_T T^{-1} \sum d_t \sum d_t'$, but the remaining two terms are not sign-definite. In the special case where instruments are strongly exogenous (sometimes referred to as "strictly" exogenous), the last two terms disappear and we expect the conditionally corrected variance to be smaller, thus giving rise to more powerful tests. However, this is never the case for dynamic R.E. models estimated with lagged values of the regressors used as instruments, in which case we have no reason to expect any of the two centering methods to yield a more powerful test.

The unconditionally centered estimator is used in Stock and Wright (2000), whereas Kleibergen (2001) prefers the conditionally centered one. It would be interesting to work out numerically which one (if any) is preferable in dynamic models, but care should be taken over the resulting

---

[22] Proof: The first expression follows obviously from the definition of V_u and stationarity of f_t. The second can be derived by expanding $V_u = T^{-1} \sum_t \sum_s \left(\bar{f}_t + d_t\right) \left(\bar{f}_t + d_t\right)'$ noting that $d_t \in \mathcal{F}_t$ but $\bar{f}_t \in \mathcal{F}_{t+1} \backslash \mathcal{F}_t$, so that $\mathrm{E}\left(\bar{f}_t d_s'\right) = 0,\ \forall\, t \leq s$.

sizes of the tests using different versions.

Finally, the above formula (2.21) can be simplified considerably if we can assume conditional homoscedasticity of h_t and stationarity.

Conditional homoscedasticity Assume that $h_t(\theta)$ is conditionally homoscedastic, namely $\mathrm{E}(\bar{h}_t(\theta)\bar{h}_s(\theta)'|\mathcal{F}_t) = \mathrm{E}\bar{h}_t(\theta)\bar{h}_s(\theta)'$. Let $\Gamma_{x,j}(t)$ denote the jth autocovariance of r.v. x at time t. Then, dropping the dependence of $\Gamma_{\bar{h},j}(t)$ on θ:

$$\begin{aligned} V_c(\theta) &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left[\Gamma_{\bar{h},0}(t)\otimes\Gamma_{Z,0}(t)\right]+ \\ &\quad +\lim_{T\to\infty}\sum_{j=1}^{T-1}\lim_{T\to\infty}\frac{1}{T}\sum_{t=j+1}^{T}\left[\Gamma_{\bar{h},j}(t)\otimes\Gamma_{Z,j}(t)+\Gamma_{\bar{h},j}(t)'\otimes\Gamma_{Z,j}(t)'\right] \end{aligned}$$

Despite the conditional homoscedasticity restriction, when the data is not stationary, the variances are potentially heteroscedastic insofar as they depend on t. To render this simplification useful for estimation, the dependence of the autocovariance functions on time must be ascertained.

Stationarity If we further assume that h_t and Z_t are weakly stationary, this becomes:

$$\begin{aligned} V_c(\theta) &= \Gamma_{\bar{h},0}\otimes\Gamma_{Z,0}+\sum_{j=1}^{\infty}\left[\Gamma_{\bar{h},j}\otimes\Gamma_{Z,j}+\Gamma'_{\bar{h},j}\otimes\Gamma'_{Z,j}\right] \\ &= \sum_{j=-\infty}^{\infty}\Gamma_{\bar{h},j}\otimes\Gamma_{Z,j} \qquad (2.22) \end{aligned}$$

Ergodicity requires declining autocovariances, but estimating the above expression by substituting sample autocovariances is not appropriate, see section 2.3.2.

Lag truncation The infinite sum of autocovariances can be truncated if there is a finite number b, such that:

$$b = \max_j\left\{j=0,1,\ldots \quad \text{s.t. } \Gamma_{\bar{h},j}\neq 0 \text{ and } \Gamma_{Z,j}\neq 0\right\} \qquad (2.23)$$

In other words, the infinite sum of autocovariances is limited by the process with the smallest memory, either $\bar{h}_t$ or Z_t.[23]

---

[23] This is a consequence of *both* stationarity *and* conditional homoscedasticity. To see why the latter is necessary, observe that, even if the instruments had no memory, i.e., $\Gamma_{Z,j}=0$, $\forall j>0$, conditional heteroscedasticity of the autocovariance of $\bar{h}_t$ would imply that the lagged terms in the sum (2.21) would involve higher moments of the Z_t process, thus causing them to be non-zero, in general.

2.3.1 Parametric HAC covariance estimators

First, consider the case of no serial correlation and conditional homoscedasticity. Then, V can be consistently estimated by:

$$\widehat{V}_c(\theta) = \widehat{\Sigma}_{hh}(\theta) \otimes \widehat{\Sigma}_{ZZ}$$

where, dropping θ for simplicity,

$$\widehat{\Sigma}_{hh} = \frac{1}{T}\sum_{t=1}^{T}\left[h_t - \left(\frac{1}{T}\sum_{s=1}^{T} h_s Z_s' \widehat{\Sigma}_{ZZ}^{-1}\right) Z_t\right]\left[h_t - \left(\frac{1}{T}\sum_{s=1}^{T} h_s Z_s' \widehat{\Sigma}_{ZZ}^{-1}\right) Z_t\right]'$$

with $\widehat{\Sigma}_{ZZ} = \frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$. When the p-dimensional parameter vector θ is estimated, a small sample correction is applied upon re-scaling by $\frac{T}{T-p}$. Versions V_n and V_u are more often used, with the only difference coming in the estimation of $\widehat{\Sigma}_{hh}$, namely $\widehat{\Sigma}_{hh} = \frac{1}{T}\sum_{t=1}^{T}\left(h_t - \widehat{C}_t^i\right)\left(h_t - \widehat{C}_t^i\right)'$, with $\widehat{C}_t^n = 0$ for the uncentered version and $\widehat{C}_t^u = \frac{1}{T}\sum_{t=1}^{T} h_t$, for unconditional centering.

When heteroscedasticity of unknown form is assumed, the Heteroscedasticity consistent (HC) estimator of White (1982) can be used:

$$\widehat{V}_i(\theta) = \frac{1}{T}\sum_{t=1}^{T}(f_t - \widehat{C}_t^i)\,(f_t - \widehat{C}_t^i)'$$

with

$$\widehat{C}_t^i = \begin{cases} 0 & \text{, no centering} \\ T^{-1}\sum_{t=1}^{T} f_t & \text{, unconditional} \\ \left(T^{-1}\sum_{s=1}^{T} h_s Z_s \widehat{\Sigma}_{ZZ}^{-1} Z_t\right) \otimes Z_t & \text{, conditional} \end{cases}$$

When a particular structure of serial correlation is assumed for the structural residual, $h_t \sim MA(q)$, we can use West's MA-l estimator.[24] Suppose $h_t(\theta_0) = \boldsymbol{\epsilon}_t + \sum_{j=1}^{q} \Phi_j\, \boldsymbol{\epsilon}_{t-j}$, and define $d_{t+q} = (Z_t \otimes I_G + Z_{t+1} \otimes \Phi_1 + \ldots + Z_{t+q} \otimes \Phi_q)\boldsymbol{\epsilon}_t$. Then, as explained in West (1997):

$$V = \operatorname*{plim}_{T\to\infty} \frac{1}{T-q}\sum_{t=1}^{T-q} d_{t+q}\, d_{t+q}' \tag{2.24}$$

A feasible estimator based on this can be derived by estimating $\{\widehat{\Phi}_j\}$ and $\{\widehat{\boldsymbol{\epsilon}}_t\}$ through fitting a G-dimensional MA(q) process on $h_t(\theta_0)$ and then substituting $\widehat{d}_{t+q} = (Z_t \otimes I + \ldots + Z_{t+q} \otimes \widehat{\Phi}_q)\widehat{\boldsymbol{\epsilon}}_t$ for d_{t+q}. When θ is consistently estimable, consistent estimates of $\{\widehat{\Phi}_j\}$ and $\{\widehat{\boldsymbol{\epsilon}}_t\}$ can derived using $h_t(\widehat{\theta})$.

The appealing features of this estimator are that it is always PSD by construction, unlike the truncated kernel estimator we mention below; and, when the assumption on the autocorrelation structure is correct, it dominates asymptotically the non-parametric estimators proposed

---

[24] In West's notation, l denotes the number of estimated equations, i.e., the dimension of h, which here we denote by G. It is not to be confused with the order of the assumed autocorrelation.

by Andrews and Monahan (1992) and Newey and West (1994), in that it converges at rate $T^{1/2}$ as opposed to at most $T^{1/3}$ for the non-parametric estimators. Moreover, this estimator is appropriate for the forward-looking monetary models that we discuss here, where the structure of autocorrelation is known when the model is correctly specified.

2.3.2 Nonparametric HAC covariance estimators

Nonparametric estimators allow for an unknown structure of the autocovariances in (2.19). However, in order to estimate V consistently, declining weights should be given to higher-order autocovariances to reflect the fact that they are more imprecisely estimated. These kernel weights must decline at an appropriate rate to guarantee consistency and positive definiteness of the resulting estimator, see Andrews (1991).

$$\widehat{V}_T = \sum_{j=-T+1}^{T-1} k(j/b_T)\widehat{\Gamma}_T(j) \tag{2.25}$$

where

$$\widehat{\Gamma}_T(j) = \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T}\left(f_t - \widehat{C}_t^i\right)\left(f_{t-j} - \widehat{C}_{t-j}^i\right)' & \text{for } j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T}\left(f_{t+j} - \widehat{C}_{t+j}^i\right)\left(f_t - \widehat{C}_t^i\right)' & \text{for } j < 0 \end{cases}$$

C^i denotes the type of centering, as defined above, $k(\cdot)$ is a real-valued kernel, and $b_T > 0$ is called the bandwidth parameter, which for most kernels is tantamount to a lag truncation parameter, see (2.23) above, with the only difference that it can be made to depend on the data and the sample size. Finally, when θ is estimated, a finite-sample degrees of freedom correction is applied upon re-scaling by $T/(T-p)$ as before.

Not all kernels will guarantee consistency or positive semi-definiteness of the resulting estimators with probability one.[25] Intuitively, the condition for consistency is that the weights given to non-zero autocovariances approach 1 as the sample size grows. For the MA(q) process, which is relevant for forward-looking models with a forecasting horizon of length q, the so-called *truncated kernel* would be consistent:

$$k(j/q) = \begin{cases} 1 & \text{for } |j| \leq q \\ 0 & \text{otherwise.} \end{cases}$$

Yet, this kernel cannot guarantee a positive semi-definite estimator, and the Monte Carlo reported in West (1997) suggests it will give non-PSD estimates more than 70% of the time in situations very relevant for the forward-looking models, i.e., when the errors exhibit strong negative autocorrelation, see chapter 4.

---

[25] See Andrews (1991, p. 821) for the definition of the class of kernels, $\mathcal{K}_1$ and examples of members of this class.

Two other kernels have gained prominence for nonparametric HAC estimators, the Quadratic Spectral (QS) kernel proposed by Andrews and Monahan (1992) and the Bartlett kernel proposed by Newey and West (1994). In both cases, the selection of the bandwidth parameter is an issue, and automatic data-based procedures have been proposed. Following the Monte Carlo evidence reported in Newey and West (1994) where it is shown than in cases similar to the ones considered here, none of these kernels dominates the other, we will focus on the Bartlett kernel.[26] Pre-whitening and recoloring, which has been found to improve finite-sample behaviour of nonparametric estimators (Andrews and Monahan (1992)) will also be considered.[27]

2.4 Testing identifiability

As a result of the growing concerns about weak instrument problems in the estimation of structural models by GMM, a number of formal and informal methods of detecting identification problems have been recommended. Recognizing the importance of the identification assumption for the consistency and asymptotic normality of IV estimators, Basmann (1960) recommended early on the use of an F-test of joint significance of the instruments in a first-stage regression as a test of identification. Also, it is typically suggested that researchers look at the R^2 or partial R^2 in the auxiliary regression, a low value being indicative of identification problems. However, this informal check, as well as formal F-tests for the joint significance of instruments in the auxiliary regression, are only appropriate as tests for *complete* lack of identification, and are insufficient to detect partial under-identification, i.e., only a subset of structural parameters being un-identified.[28]

Recognizing those shortcomings, Cragg and Donald (1993) proposed a test of the null of partial identification in IV, based on the idea of testing for reduced rank in the coefficients of the first-stage regression, similar to the reduced rank regression (Anderson and Rubin (1949)) approach. Cragg and Donald (1993, theorem 1) is general enough to cover the possibility of serial correlation in the first-stage regression residuals, but implementation in that case was not offered, nor numerically evaluated. Moreover, even though their proposed test can be adapted to our more general setting,

---

[26] For the automatic bandwidth selection, we used the procedure implemented by Hall (2000), as the emphasis is on the test of over-identifying restrictions.

[27] This involves applying a low order VAR(b) filter to the moment conditions f_t to get the residuals f_t^*, compute the nonparametric HAC estimator using f_t^*, $\widehat{V}^*$, and then 're-color' using $\widehat{B}\,\widehat{V}^*\widehat{B}'$, where $\widehat{B} = \left(I_{Gk} - \sum_{i=1}^{b} \widehat{A}_i\right)^{-1}$, and $\widehat{A}_i$ are the coefficients of the auxiliary VAR, see Andrews and Monahan (1992, pp. 954-5) for details. In our simulations later on we use $b = 1$.

[28] As recognized by Nelson and Startz (1990) and Shea (1997), this test will reject with probability one even when the model is partially identified in the sense of Phillips (1989). In terms of the discussion of the previous chapter, this would correspond to the case of the concentration matrix being of reduced but non-zero rank.

it has power against both identification and mis-specification of the moment conditions. So, it is not a pure identification test.

Thus, in this section we develop a similar test that can be seen as an extension to a naive reduced rank test and which is both robust to serial correlation and heteroscedasticity of the moment Jacobian d_t. What is more, this statistic tests only the null hypothesis of partial identification, against an alternative of complete identification, thus being a pure identification test. This test is thus suitable for use in dynamic forward-looking models, where the forecast error in the auxiliary regression is autocorrelated by construction, as explained in section 4, which renders the naive reduced rank tests invalid.[29]

2.4.1 Simple Reduced Rank Regression test

The methodology is based on a combination of Reduced Rank Regression (RRR) and GMM applied to the 'first-stage' (forecasting) regression of an IV procedure. Given that the coefficients of this auxiliary regression are not of interest, we restrict attention to testing for reduced rank.

Consider again the prototype IV regression model with no exogenous variables (see below for the more general case), where Y_t is a p-dimensional vector of endogenous regressors and Z_t is a k-dimensional vector of valid instruments:

$$\begin{aligned} \mathrm{y}_t &= \theta' Y_t + u_t \\ Y_t &= \Pi' Z_t + v_t \end{aligned} \tag{2.26}$$

Only the top equation is of interest, the bottom one being the auxiliary or first-stage regression. As we discussed above, the identification of θ depends on the rank of Π, (which is the same as the rank of the concentration parameter). Letting $\rho(\Pi) = r$ denote the rank of Π, we would like to test different hypotheses for r, allowing for possible serial correlation and heteroscedasticity in v_t. The difference $p - r$ denotes the degree of under-identification of the model, which is also of

---

[29] A completely different approach has been recently proposed by Hahn and Hausman (2002). These authors propose a new specification test that takes a similar approach to that of Hausman (1978), and estimates the structural parameters in two different ways, forward and backward 2SLS. The discrepancy between the two alternative estimators is shown to be informative about the presence of identification problems in the model. This test is particularly interesting because it tests the null hypothesis of strong identification, and, as a result, it has power in exactly the opposite direction to our proposed test. It is therefore interesting to compare its power relative to our test in the context of a particular forward-looking structural model.

interest. So, we wish to test the following sequence of hypotheses:

$$\begin{aligned} &\mathrm{H}_0 : \ \mathrm{rank}(\Pi) = 0 \\ &\mathrm{H}_1 : \ \mathrm{rank}(\Pi) = 1 \\ &\vdots \\ &\mathrm{H}_p : \ \mathrm{rank}(\Pi) = p \end{aligned} \tag{2.27}$$

where the implied restrictions are nested as

$$\mathrm{H}_0 \subset \ldots \subset \mathrm{H}_r \subset \ldots \subset \mathrm{H}_p$$

For each of those hypotheses, we can find a $p \times r$ vector d and a $k \times r$ vector B such that $\Pi = B\, d'$ of rank r. Under H_p, Π is of full rank and it can be estimated consistently by OLS, though this will not be efficient when v_t is heteroscedastic or serially-correlated. Under any of the other hypotheses, the RRR must be applied.

The standard likelihood ratio test statistic for reduced rank (Anderson and Rubin (1949)) is the well-known trace statistic:[30]

$$LR(r) = -T \sum_{i=1}^{p-r} (1 - \widehat{\lambda}_i) \approx T \sum_{i=1}^{p-r} \widehat{\lambda}_i \tag{2.28}$$

where $\widehat{\lambda}_i$, $i = 1, \ldots, p$ are the solutions to the generalized eigenproblem:

$$|\lambda\, \widehat{\Sigma}_{vv} - \widehat{\Sigma}_{YZ} \widehat{\Sigma}_{ZZ}^{-1} \widehat{\Sigma}_{ZY}| = 0$$

The eigenvalues are sorted in ascending order, $(\widehat{\lambda}_1 < \ldots < \widehat{\lambda}_p)$. Notably, this is an estimate of the sum of the $p - r$ smallest eigenvalues of the concentration matrix $\Sigma_{vv}^{-1/2} \Pi' \Sigma_{ZZ} \Pi \Sigma_{vv}^{-1/2}$, see equation (2.15).

A special case of this test has also been recommended by Stock and Yogo (2001), who consider only the test of the hypothesis $\mathrm{H}_{p-1} : \rho(\Pi) = p - 1$, which amounts to testing the minimum eigenvalue of the concentration parameter matrix.

Finally, the resulting estimates of the $p \times r$ and $k \times r$ vectors d and B admit useful interpretations. The former spans the subspace of *identified* linear combinations of the structural parameters θ, i.e., the directions in which the instruments are informative, whereas the latter provides an estimate of the optimal instruments, $B' Z_t$.

The above procedure is, however, invalid when the first-stage regression residuals v_t exhibit serial correlation and/or heteroscedasticity. We therefore propose a test based on one-step GMM

---

[30] The reader can consult Anderson (1984), or Johansen (1995, chapter 6) for details.

estimation, which is potentially inefficient, but yields a procedure that is valid and simple to implement.

2.4.2 Minimum Distance Reduced Rank test

First, we note that the Reduced Rank Regression can be seen as a quasi-MLE procedure which yields consistent estimates of the restricted reduced rank matrix $\Pi = B\,d'$. We shall use the notation Π_r to denote the rank-restricted parameter, with $\widehat{\Pi}_p$ denoting the unrestricted OLS estimator.

We can view the quasi-MLE as a GMM estimator with moment functions:

$$\bar{d}_t(\Pi) = \mathrm{vec}\,(d_t - \mathrm{E}(d_t|\mathcal{F}_t)) = \mathrm{vec}\,(Z_t\,(Y_t - \Pi' Z_t)') = v_t \otimes Z_t$$

where d_t is the Jacobian of the moment conditions of the equation of interest, in line with the notation of table 2.1. The above equation vanishes in expectation at the true parameter point under the null of rank restriction. The corresponding empirical orthogonality restrictions are:

$$\bar{D}_T(\Pi) = \frac{1}{T}\sum_{t=1}^{T} \bar{d}_t(\Pi) \tag{2.29}$$

At the unrestricted MLE (OLS), the above moment conditions are identically equal to zero, but this is not the case under the reduced-rank restriction. Thus, we may write down a GMM criterion function using the inverse of the limiting variance of $\bar{D}_T(\Pi)$ as weighting matrix:

$$\mathcal{Q}_T(\Pi;\tilde{\Pi}) = T\,\bar{D}_T(\Pi)'\,\mathcal{V}(\tilde{\Pi})^{-1}\,\bar{D}_T(\Pi) \tag{2.30}$$

where $\tilde{\Pi}$ is some preliminary estimator, e.g., the unrestricted OLS estimator $\widehat{\Pi}_p$ or the Reduced Rank Regression estimator $\widehat{\Pi}_r = \widehat{B}\,\widehat{d}'$ (QMLE). It is possible to obtain an efficient estimate of Π by minimizing this objective function under the reduced rank restriction. This could be used to test the reduced rank restriction correctly, provided a consistent estimator of the variance of the moment conditions (2.29) can be found, $\widehat{\mathcal{V}}_T \xrightarrow{p} \mathcal{V}$. However, a simpler test that avoids re-estimation can be based on the one-step GMM estimator $\widehat{\Pi}_r$ using the following result.

Lemma 2.5. *Assume that* $T^{-1}\sum_{t=1}^{T} Y_t Y_t' \xrightarrow{p} \Sigma_{YY}$, $T^{-1}\sum_{t=1}^{T} Z_t Z_t' \xrightarrow{p} \Sigma_{ZZ}$, *where* Σ_{YY} *and* Σ_{ZZ} *are positive definite,* $T^{-1}\sum_{t=1}^{T} Y_t Z_t' \xrightarrow{p} \Sigma_{YZ}$ *of rank* $r < p$, *and* $\sqrt{T}\bar{D}_T(\Pi_0) \xrightarrow{d} N(0,\mathcal{V})$, *where* $\Pi_0 = \Sigma_{ZZ}^{-1}\,\Sigma_{YZ}'$. *Then,*

$$T\,\bar{D}_T(\widehat{\Pi}_r)'\left[S\,\left(Q'\,\mathcal{V}\,Q\right)^{-1}\,S'\right]\bar{D}_T(\widehat{\Pi}_r) \xrightarrow{d} \chi^2_l, \quad l = (k-r)(p-r).$$

where S, Q *are full rank* $kp \times m$ *matrices,* $m = kp - l$*, s.t.* $S\,Q' = (I_p - \mathcal{J}_{\bar{D}}\left(\mathcal{J}_{\bar{D}}'\mathcal{W}\mathcal{J}_{\bar{D}}\right)^{-1}\mathcal{J}_{\bar{D}}'\mathcal{W})$*, and* $\mathcal{W}$ *and* $\mathcal{J}_{\bar{D}}$ *are defined in the appendix 2.A.*

We note that the objective function (2.30) cannot be used directly at an inefficient estimate of Π_r, and hence the need to replace $\mathcal{V}^{-1}$ with $\left[S\,(Q'\,\mathcal{V}\,Q)^{-1}\,S'\right]$.

In order to render the above result operational, we need to substitute consistent estimators for $\mathcal{J}_{\bar{D}}$, $\mathcal{W}$ (hence S and Q) and $\mathcal{V}$. The former are easily found upon replacing sample moments for population moments in their definition. Based on the discussion of section 2.3, $\mathcal{V}$ can also be estimated consistently by a HAC estimator, under the conditions of the lemma. Of course, the choice of the HAC estimator is likely to make a difference to the finite-sample size and power properties of the test.

In fact, since the auxiliary regression is not meant to be an adequate model of the dynamic evolution of Y, it is possible that potentially infinite-order serial correlation is left out in the residuals. This would be the case if Y_t have a rich dynamic structure, but only a few instruments are used for the estimation of θ, and these may not be sufficient to capture all of the dynamics in Y_t. Thus, it seems difficult to justify the use of a parametric estimator without further assumption, and in our implementation, we shall use the Newey-West nonparametric estimator of the form (2.25).[31] So, the test statistic is:

Definition 2.9 (Quasi Minimum Distance Reduced Rank statistic).

$$QMD(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}\left(Z_t\,(Y_t' - Z_t'\,\widehat{\Pi}_r)\right)'\left[\widehat{S}\,\left(\widehat{Q}'\,\widehat{\mathcal{V}}(\widehat{\Pi}_r)\,\widehat{Q}\right)^{-1}\,\widehat{S}'\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}\left(Z_t\,(Y_t' - Z_t'\widehat{\Pi}_r)\right) \tag{2.31}$$

Exogenous regressors The above procedure need only be slightly modified in the case of exogenous regressors. When some of the Z_t are included in the model (2.26) as exogenous regressors, say Z_{1t}, with Z_{2t} used as instruments, the model becomes

$$\begin{aligned} y_t &= \theta_1'\,Y_t + \theta_2' Z_{1t} + u_t \\ Y_t &= \Pi_1' Z_{1t} + \Pi_2' Z_{2t} + v_t \end{aligned} \tag{2.32}$$

The identification of $\theta = (\theta_1', \theta_2')'$ is now governed by the rank of Π_2. The $QMD(r)$ will be computed as before but using a reduced rank estimate of Π_2 after correcting for Z_{1t} in the first

---

[31] $\mathcal{V}$ could also be estimated at any other consistent estimate of Π, e.g., the unrestricted OLS estimator $\widehat{\Pi}_p$. However, Monte Carlo evidence suggests that using the more efficient estimate $\widehat{\Pi}_r$ improves the size of the resulting test statistic.

stage regression, i.e., by reduced rank regression on:

$$R_{1t} = \Pi_2' R_{2t} + e_t$$

where $R_{1t} = M_{Z_2} Y_t$ and $R_{2t} = M_{Z_2} Z_t$, where M_{Z_2} is the orthogonal projector onto $\mathrm{Col}^{\perp}(Z_2)$.

Why not an efficient GLS test? It is well-known that GLS-type estimators are inconsistent in this context – Hayashi and Sims (1983), Cumby, Huizinga, and Obstfeld (1983). Because of the structure of serial correlation, which is a moving average with respect to the future, any re-scaling of data to render the distribution of the error v_t serially uncorrelated will induce a correlation between the transformed regressors and the new residuals. In contrast, OLS is consistent but inappropriate for inference, and hence the need for the correction for serial correlation, see Brown and Maital (1981) for a similar problem in a different context.

Of course, a more efficient way to estimate the reduced rank matrix Π_r would be by 2-step or multi-step GMM (the CUE would be unstable given the large number of parameters). Yet, it is not clear that the resulting test would be either better sized or more powerful in finite samples, given that it is asymptotically equivalent to the 1-step GMM-based test. It remains an open question to check how this simple test compares to a more efficient version in finite samples.

Extension to GMM Recently, Wright (2000) proposed a test for partial identification that is applicable to non-linear GMM models. The test is essentially a reduced rank test for the Jacobian of the moment conditions D_T, see table 2.1. The test that Wright proposed is based on Cragg and Donald (1997) and tests the null hypothesis that the rank of the limiting Jacobian $\mathcal{J}(\theta)$ is $r < p$, using a Bonferroni argument to circumvent the dependence on the nuisance parameter θ since it is not consistently estimable under the null hypothesis of partial identification. Wright showed his proposed test to be asymptotically conservative, but useful in detecting lack of identification.

A few remarks about this test are in order. It is our impression that an asymptotically similar, and therefore correctly sized statistic could be derived even under the null of under-identification. This follows from the observation that the Jacobian will be the same for all $\theta \in \mathbf{\Theta}_0$ defined in assumption D, i.e., over the set of points at which the GMM criterion function is flat. It suffices then to impose just enough (p_1) identifying restrictions to estimate θ_R consistently, and provided the restrictions are chosen carefully, this parameter will lie in $\mathbf{\Theta}_0$. Wright's statistic will then be asymptotically similar w.r.t. θ, as would be any other nonparametric test of reduced-rank based on Cragg and Donald (1997).[32]

2.5 Mis-specification testing

In view of the discussion in section 2.2, the importance of a mis-specification test cannot be emphasized more. It is important to point out that our focus here is on mis-specification of the type specified by definition 2.2, so we abstract from parameter instability or any other form of non-stationarity.

2.5.1 The Hansen-Sargan test

The main specification test in GMM is a test of the over-identifying restrictions and the most widely used such test is the Hansen-Sargan test, also known as the J-statistic, defined as the value of the minimized objective function, namely:[33]

$$J = \mathcal{Q}_T(\widehat{\theta};\widehat{\theta}) \tag{2.33}$$

It is well-known that under weak identification, assumption D, the distribution of the J-test is non-standard and depends on nuisance parameters, see Stock and Wright (2000, Corollary 4). However, based on Stock and Wright (2000, Theorem 3), we can find a boundedly pivotal version of the test. This intuition is formalized in the following result.

Theorem 2.6. *Under the assumptions of theorem 2.1,*

1. *If $W_T(\theta) \xrightarrow{p} V(\theta)^{-1}$ uniformly in $\theta \in \mathbf{\Theta}$:*

   $$J_{CUE} = \mathcal{Q}_T(\widehat{\theta}_c;\widehat{\theta}_c) \Rightarrow \bar{\mathcal{Q}}(\alpha^*, b^*;\alpha^*)$$

   where $\widehat{\theta}_c$ is the CUE, and $\bar{\mathcal{Q}}(\alpha^*, b^*;\alpha^*)$ is as defined in theorem 2.1, with $W(\bar{\theta}(\phi(\alpha,\beta_0))) = V(\phi(\alpha,\beta_0))^{-1}$.

2. *Under the null hypothesis of correct specification (assumption D), the above distribution is bounded from above by a $\chi^2(Gk - p_2)$.*

---

[32] Another, nonparametric test that could be used to test rank restrictions in the Jacobian is Robin and Smith (2000).

[33] This is not to be confused with the J-test for non-nested hypotheses proposed by Davidson and MacKinnon (1981).

This result shows that the J-test will not have the conventional χ^2 distribution with degrees of freedom equal to the degree of over-identification. Its asymptotic distribution is not similar w.r.t. nuisance parameters in $\mu(\alpha) = V\left(\phi(\alpha, \beta_0)\right)^{-1/2} m_1\left(\phi(\alpha, \beta)\right)$;[34] nor can a correction to this distribution be derived, because the nuisance parameters are not consistently estimable in this framework. For instance, in a linear IV model, the distribution simplifies to a non-central χ^2 indexed by the inconsistently estimable parameter α, see also the discussion in Stock and Wright (2000, Section 2.5).

The second part of the theorem is more interesting because it provides a bound to the non-pivotal distribution of the J statistic, thus showing that an asymptotically *boundedly* pivotal test can be derived by evaluating the objective function at the CUE. This would not be the case if the J statistic were computed using any multi-step GMM estimator, since the inequality (2.A.2) in the proof need not hold, see appendix 2.A. It is noteworthy that the resulting inference will be asymptotically conservative, i.e., the test will under-reject under the null. Furthermore, we have no indication of the extent of the size distortion, which clearly depends on the nuisance parameters.

For the analysis of mis-specification in chapter 4, it is also useful to investigate the power of the J-test when the model is well-identified. The following result establishes the distribution of the statistic under a local alternative.

Lemma 2.7. *When all the parameters are well-identified, $\theta \equiv \beta$ and $p \equiv p_2$ in assumption D', and under a local alternative ($\zeta_f = 0$), the J-statistic is asymptotically distributed as $\chi^2(Gk - p, \nu^2)$, with non-centrality parameter*

$$\nu^2 = \zeta_l' \left[V(\beta_0)^{-1} - V(\beta_0)^{-1} \mathcal{J}(\beta_0) \left(\mathcal{J}(\beta_0)' V(\beta_0)^{-1} \mathcal{J}(\beta_0) \right)^{-1} \mathcal{J}(\beta_0)' V(\beta_0)^{-1} \right] \zeta_l \qquad (2.34)$$

Hereafter, ν will be referred to as the *mis-specification parameter*, which governs the degree to which mis-specification is *detectable*, similarly to the concentration parameter, governing the degree of identification. The definition of ν is intuitive, since it makes clear that the extent to which any mis-specification is detectable only depends on the distance of the *over-identifying* restrictions from orthogonality, in the metric of the variance of the moment conditions.

---

[34] The quantity $\mu(\alpha)'\mu(\alpha)$ normalized by the range of α can be thought of as a generalization of the concentration parameter for GMM, see discussion in Stock and Wright (2000, section 2.6).

Power

On top of the above weaknesses, the J-test is affected by the common problems of over-instrumenting and over-correction for serial correlation. It is well-known that the former trivially reduces power in the well-identified case. The main driver of power is the non-centrality parameter. When the model is well-identified, this parameter is fixed, and the distribution of the test under the alternative is simply χ^2 with degrees of freedom equal to the number of over-identifying restrictions. It can be shown that $\chi^2(k+j,\delta) > \chi^2(k,\delta)$ for all $j > 0$, as illustrated in figure 2.3. Since adding irrelevant instruments increases the degrees of freedom without affecting the non-centrality parameter, over-instrumenting reduces power.[35]

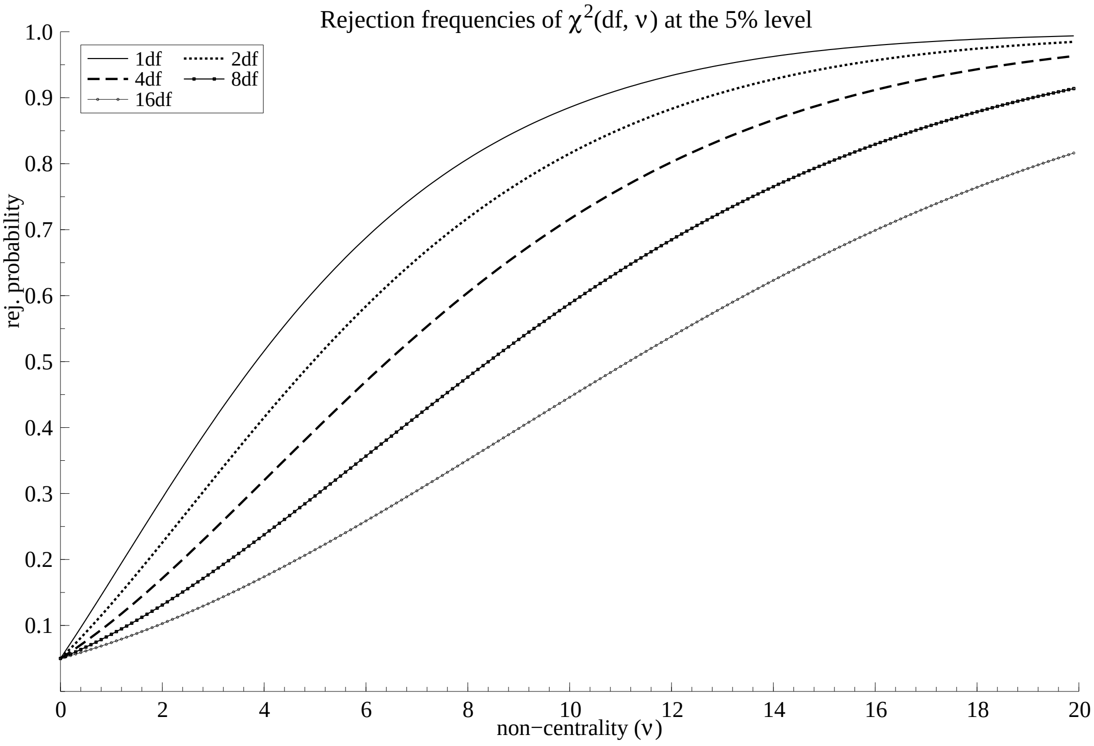


Figure 2.3: Tail probabilities of the non-central χ^2 distribution.

The impact of serial correlation corrections on the power is less obvious. The various HAC estimators proposed have been found to have poor finite-sample properties. Monte Carlo evidence reported in Andrews (1991), Newey and West (1994) and West (1997) shows considerable finite-sample size distortions when their proposed HAC estimators are used to derive a t-test in a regression. The issue here is how the HAC estimators affect the size of the Hansen-Sargan test.

---

[35] In the weakly-identified case, a similar argument is more difficult to establish.

This is investigated by Monte Carlo in chapter 3, and also for the specific example of the New Phillips Curve in chapter 4. According to those results, size distortions differ using different HAC estimators, and tend to be more dramatic the more general is the assumed serial correlation. This also helps explain the results reported by Clarida, Galí, and Gertler (1998), who estimated the same forward-looking Taylor rule for six industrialized countries and reported p-values for the Hansen test of 0.999 in all cases, which is indicative of size distortion.

2.5.2 Parsimonious selection of instruments

The problem of over-instrumenting is well-known to applied econometricians, with some notable exceptions e.g., Clarida, Galí, and Gertler (1998) and Galí and Gertler (1999) who use more than 10 instruments per endogenous parameters. To address this issue, it is commonly recommended that applied researchers carefully select a subset out of a potentially large admissible instrument set, an approach also known as "parsimonious selection of instruments". This can be done either by sequential t- or F-tests on the first-stage regression, or by appropriate information criteria, such as the Canonical Correlations Information Criterion (CCIC) recently proposed by Hall and Peixe (2001) for IV and Hall and Inoue (2001) for GMM. The CCIC, as they call it, is shown to consistently select the subset of relevant instruments, within a given instrument set. However, its utility in a dynamic context, where the instrument set is potentially very large, has not yet been investigated. Nor has it been used in conjunction with a test of over-identifying restrictions to improve power (e.g., pre-testing to reduce number of instruments on which to base the *J*-test).[36]

Despite its obvious utility, we would like to raise an issue here that appears to have been overlooked, namely whether this selection procedure is likely to bias or distort the size of the over-identifying restrictions test on the smaller instrument set. In other words, we want to ask the question of whether the instruments that are being excluded are the ones that are likely to correlate less (or more) with the resulting structural residuals than the included ones, thus biasing the test statistic upwards (downwards).

---

[36] Hall and Peixe (2001) have done some Monte Carlo studies to assess the utility of their Canonical Correlations Information Criterion (CCIC) in improving the finite-sample properties of t-tests, with encouraging results, but have not used it for the purpose we discuss here. Moreover, the consistency of this criterion in selecting relevant instruments relies on the model being well-identified, thus rendering it inappropriate under weak instruments.

Impact on the size of tests

We formalize the problem as follows. Let $\{Z_i\},\ i = 1, \ldots, k$ denote the set of admissible instruments, J be the Hansen-Sargan statistic when all instruments are used, and let J_{-i} be the statistic using all but the ith instrument Z_i. In other words, there is a set of k different J-statistics that can be formed. Under the null of correct specification, the asymptotic critical value for each one of them is c_{k-1}, the α-quantile of the $\chi^2(k-1)$ distribution. Also denote by NRP_i the corresponding (asymptotic)[37] null rejection probability given a particular Instrument Selection Rule, ISR, namely

$$\mathrm{NRP}_i(\mathrm{ISR}) = \frac{\Pr(J_{-i} \text{ rejects } \cap\ J_{-i} \text{ selected})}{\Pr(J_{-i} \text{ selected})}. \tag{2.35}$$

where the event $\{J_{-i} \text{ selected}\}$ is the same as the event $\{Z_i \text{ removed}\}$. Hence, the null rejection probability of this parsimonious test, given the selection rule, is simply:

$$\mathrm{NRP}(\mathrm{ISR}) = \mathrm{E}\left(\mathrm{NRP}_i\, |\mathrm{ISR}\right) = \sum_{i=1}^{k} \Pr(J_{-i} \text{ rejects } \cap\ J_{-i} \text{ selected}) \tag{2.36}$$

It is obvious that if the instrument selection rule is independent from the data, then the parsimonious test will be asymptotically similar with respect to it, i.e., $\mathrm{NRP}_i = \Pr(J_{-i} \text{ rejects}) = \alpha$ in (2.35). An important exception occurs when the model is effectively just identified, and the random selection rule removes a relevant instrument with positive probability. But this is unlikely to happen when the ISR is data-dependent, and this is the case of interest.

As an example of a rule causing over-rejection, consider the following: "select the subset that maximizes the resulting Hansen-Sargan test", i.e., $\max_i(J_{-i})$. In this case,

$$\Pr(J_{-i} \text{ rejects } \cap\ J_{-i} \text{ selected}) \geq \Pr(J_{-i} \text{ rejects }) \Pr(J_{-i} \text{ selected})$$

since, in general, $\Pr(\max_i(X_i) > c) \geq \Pr(X_i > c)$ for any sample $\{X_i\}\ i = 1, \ldots k$. Suppose that Z_i are iid, in which case $\Pr(J_{-i} = \max_i(J_i)) = 1/k$, the same for all of them. Then, clearly,

$$\mathrm{NRP}(\mathrm{ISR}) = k \Pr(J_{-i} \text{ rejects } \cap\ J_{-i} \text{ selected}) \geq k \Pr(J_{-i} \text{ rejects }) \Pr(J_{-i} \text{ selected}) = \alpha$$

For any realistic ISR it is rather difficult, if not impossible, to compute NRP(ISR) analytically, so one has to resort to numerical methods. An interesting question is whether the size distortion is less when an irrelevant instrument is removed, as opposed to a relevant one. If that conjecture were verified, it would call for the use of selection rules which are known to select the relevant

---

[37] The J-statistic has been shown to be asymptotically pivotal when the model is well-identified, but it is *exactly* pivotal when the instruments are strongly exogenous, provided the latter have second moments, see also the discussion in chapter 3. When this is not the case, NRP_i is understood to be $\lim_{T\to\infty} \Pr_T(J_{-i} > c_{k-1}\,|\,\mathrm{ISR})$.

instruments with very high probability, e.g., the automatic model selection procedure advocated by Krolzig and Hendry (2001) and implemented by PcGets. However, we will not pursue this issue here, as it is outside the scope of the present study.

Impact on power of tests

Another point worth mentioning is that parsimonious instrument selection will not unambiguously increase power under the alternative of misspecification, assumption D$'$. An example of this is when an instrument Z_i is uninformative for the endogenous regressors ('insignificant' in the first stage regression), but it actually correlates with the regressand, and hence contributes to the non-centrality parameter of the Hansen-Sargan test, when included in the instrument set.[38] If this instrument is removed, both the degrees of freedom and the non-centrality of the test will be reduced, see lemma 2.7. In the extreme case where the only non-centrality arises from that instrument, the parsimonious test may actually lose power completely, since a data-driven instrument selection rule will remove that instrument with very high probability.

Some authors have proposed instrument selection procedures that remove both invalid as well as irrelevant instruments (Andrews (1999)). However, this cannot be justified in a dynamic rational expectations model because a testable implication of 'rationality' is precisely the orthogonality of the error to everything in the information set.

To avoid this problem, two suggestions can be made. First, after parsimoniously selecting a set of instruments in the first stage regression, one can also test for any correlation in the model's residuals against any excluded instruments. It appears that this can be done by including those variables as additional regressors in the model, and performing an F-test of joint significance, but whether an F-test would be valid, given the instrument selection rule remains an issue. Namely, the size of the test could be distorted by the rule, as in the case of the parsimonious J-test above. Alternatively, the parsimonious selection procedure should be based on the reduced form of the entire system, not just the first stage regression, and an instrument should be removed only if it is uninformative for both Y_t *and* y_t.

2.5.3 Testing for excess serial correlation

Another fruitful approach to testing mis-specification, which is relevant in a dynamic context, is testing for excess serial correlation (SC) in the structural residuals. This test could potentially

---

[38] It also biases the resulting estimators, as we prove in corollary 2.2 and lemma 2.7.

be more informative than the usual J-test to detect mis-specification in certain cases, because it addresses the issue of over-instrumenting by drastically reducing the number of degrees of freedom. In the case of a linear model, where lags of the endogenous variables are used as instruments, we can offer some straightforward intuition for this claim. Instead of testing the orthogonality of the residuals against all of the over-identifying instruments, the serial-correlation test only tests against a small linear combination of the instruments, the past residuals.[39] This in turn implies that the SC statistic cannot exceed numerically the J-statistic, but it will have many fewer degrees of freedom than the latter, in general, and hence it could be more powerful.

This is particularly relevant for the dynamic forward looking models we consider here. When a structural model is dynamically mis-specified, that mis-specification could potentially show up as extra serial correlation in the model's residuals, beyond what would have been implied by the model if it were correctly specified. To exemplify, consider the Galí-Gertler example in section 2.1.3 above. The IV regression residual in model (2.8) is seen to exhibit first-order autocorrelation even under correct specification, which arises due to the forward-looking nature of the model. However, dynamic mis-specification would correspond to lags of π_t, s_t being incorrectly omitted from the model, and if these were used as instruments, they would be invalid. Moreover, it is very likely that the omitted variables, which are themselves autocorrelated, will induce extra serial correlation in the model's residuals. It is, of course, likely that instruments may be non-exogenous in other ways that do not induce any detectable excess serial correlation. But the important point is that any detected excess serial correlation is sufficient (albeit not necessary) to imply instrument non-exogeneity, i.e., mis-specification.

Here, we discuss a simple test of excess serial correlation based on Cumby and Huizinga (1992), and suggest ways in which it can be adapted to the case of weak identification.

The test

Consider the structural disturbances $\epsilon_t \equiv h_t(\theta)$. The hypothesis to be tested is that the disturbances are a moving average process of known order q, against the alternative that autocorrelations at lags greater than q are non-zero. Letting ρ_j denote the jth autocorrelation of ϵ_t, the hypothesis

---

[39] Provided of course that enough lags of the regressors/regressant are used in Z to be able to express lagged residuals as a combination of Z.

can be formulated as:

$$\begin{aligned} \mathrm{H}_0: &\quad \rho_j = 0, \quad \forall\, j = q+1, \dots q+s \\ \mathrm{H}_1: &\quad \rho_j \neq 0, \quad \text{for some } \ j = q+1, \dots q+s \end{aligned} \tag{2.37}$$

We propose testing the above hypothesis in a GMM framework, by testing the validity of the following moment conditions:

$$\mathrm{E}(\epsilon_t\, \epsilon_{t-q-j}) = 0, \quad j = 1, \dots, s$$

Define the $s \times 1$ vector $U_t = (\epsilon_{t-q-1}, \dots, \epsilon_{t-q-s})'$, the $s \times 1$ vector $w_t(\theta) = U_t(\theta)\epsilon_t(\theta)$, and the $T \times s$ data matrix $U = (U_1, \dots, U_T)'$ with $\epsilon_t = 0$, for $t < 0$, and $\epsilon = (\epsilon_1, \dots, \epsilon_T)'$. The corresponding empirical moments can then be written as:

$$\frac{1}{T} U' \epsilon = \frac{1}{T} \sum_{t=1}^{T} U_t\, \epsilon_t = \frac{1}{T} \sum_{t=1}^{T} w_t(\theta) \tag{2.38}$$

The notation $w_t(\theta)$ emphasizes the dependence of these moments on the original model parameters θ, which arise through the structural errors $\epsilon_t = h_t(\theta)$.

We distinguish three cases: (i) The true disturbances are known, i.e., $\theta = \theta_0$; (ii) θ_0 is consistently estimable; and (iii) θ_0 is not consistently estimable. The reason is apparent if we perform the following asymptotic expansion of $w_t(\theta)$ over θ:

$$w_t(\widehat{\theta}) = w_t(\theta_0) + \left. \frac{\partial\, w_t}{\partial\, \theta'} \right|_{\theta = \widehat{\theta}} (\widehat{\theta} - \theta_0) + O_p(\|\widehat{\theta} - \theta_0\|^2) \tag{2.39}$$

Under conventional stationarity and ergodicity assumptions for the disturbance ϵ_t, $T^{-1} \sum_{t=1}^{T} w_t(\theta_0)$ is $O_p(T^{-1/2})$ and asympotically normally distributed, and $T^{-1} \sum_{t=1}^{T} \frac{\partial\, w_t(\widehat{\theta})}{\partial\, \theta'} = O_p(1)$. Yet, even when the disturbances w_t can be consistently estimated by $\widehat{w}_t = w_t(\widehat{\theta})$, there will be a $O_p(T^{-1/2})$ discrepancy which has to be accounted for when using $T^{-1} \sum_{t=1}^{T} \widehat{w}_t$ rather than $T^{-1} \sum_{t=1}^{T} w_t$ to form a test.

Under the cases of partial or weak identification, $\widehat{\theta} - \theta_0 = O_p(1)$, and higher-order terms in (2.39) are no longer negligible.

When the estimated residuals are used in forming the moment conditions (2.38), the impact of the estimation of θ has to be taken into account in estimating the correct variance, $V_{\widehat{w}}$ as opposed to V_w. Averaging out equation (2.39), and using the fact that $\sqrt{T}(\widehat{\theta} - \theta_0) = \sqrt{T} H_T(\theta_0)\, g_T(\theta_0) + o_p(1)$

when the model is well-identified, we have:

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widehat{w}_t &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} w_t(\theta_0) + \frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial\, w_t}{\partial\, \theta'}\right|_{\theta_0} H_T(\theta_0)\sqrt{T}g_T(\theta_0) + o_p(1) \\ &= \begin{pmatrix} B_T\, H_T & I_s \end{pmatrix} \underbrace{\begin{pmatrix} T^{-1/2}\, Z'\epsilon \\ T^{-1/2}\, U'\epsilon \end{pmatrix}}_{\sqrt{T}g_T^+} + o_p(1) \end{aligned} \tag{2.40}$$

where $B_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial\, w_t}{\partial\, \theta'}$ is $(s \times p)$, and $H_T(\theta_0) = \left(D_T(\theta_0)'V_T(\theta_0)^{-1}D_T(\theta_0)\right)^{-1} D_T(\theta_0)V_T(\theta_0)^{-1}$ is a $(p \times k)$ matrix, see variable definitions in table 2.1, p. 14 ($G = 1$ here). The second line shows how the estimated autocovariances $\sum \widehat{w}_t$ can be expressed as a linear combination of an extended set of orthogonality conditions, where U_t has been appended to the list of instruments Z_t. Let $g_T^+(\theta_0)$ denote the extended average moment conditions, see (2.40). Hence, the desired (asymptotic) variance $V_{\widehat{w}}$ is a linear combination of the (asymptotic) variance matrix of those new moment conditions, which is

$$V_{g^+} = A\text{var}(\sqrt{T}\, g_T^+) = \begin{pmatrix} V & C' \\ C & V_w \end{pmatrix} \tag{2.41}$$

where V is the variance of the original moment conditions $g_T = Z'\epsilon/\sqrt{T}$ and C is the covariance between the two sets of moment conditions. As before, this quantity can be consistently estimated using a parametric or nonparametric HAC, but in view of the discussion in the previous subsection, West's MA-l estimator seems to be the most appropriate choice: An MA(q) process would be fitted to the estimated residuals $\widehat{\epsilon}_t = \widehat{v}_t + \widehat{\phi}_1\, \widehat{v}_{t-1} + \ldots + \widehat{\phi}_q\, \widehat{v}_{t-q}$, the $(k + s \times 1)$ auxiliary process $\widehat{d}_{t+q}$ would be formed,

$$\widehat{d}_{t+q} = \left[\begin{pmatrix} Z_t \\ \widehat{U}_t \end{pmatrix} + \ldots + \begin{pmatrix} Z_{t+q} \\ \widehat{U}_{t+q} \end{pmatrix}\widehat{\phi}_q\right]\widehat{v}_t,$$

and the MA-l estimator $\widehat{V}_{g^+}$ would then be given by (2.24).

To complete the estimation problem, we need to determine B_T. This is also a Jacobian of the new moment conditions $U'\epsilon$ with respect to θ, with the only apparent complication that the new 'instruments' U_t are also functions of θ:

$$B_T = \frac{1}{T}\sum_{t=1}^{T}\underset{s\times 1}{U_t}\,\underset{1\times p}{\frac{\partial\, h_t}{\partial\, \theta'}} + \frac{1}{T}\sum_{t=1}^{T}\underset{s\times p}{\frac{\partial\, U_t}{\partial\, \theta'}}\,\epsilon_t$$

Since U_t are past disturbances, the second term involves cross-products of ϵ_t with variables in the information set $\mathcal{F}_t$. Hence, that term is $o_p(1)$ in dynamic rational expectations models, since the disturbance is assumed to be an innovation with respect to the instruments as well as its own history. Thus, B_T can be consistently estimated by:

$$\widehat{B}_T = \frac{1}{T}\sum_{t=1}^{T} \widehat{U}_t \frac{\partial h_t}{\partial \theta'}\bigg|_{\widehat{\theta}}$$

Finally, a consistent estimator of $V_{\widehat{w}}$ is given by:

$$\widehat{V}_{\widehat{w}} = \begin{pmatrix} \widehat{B}_T \, \widehat{H}_T & I_s \end{pmatrix} \widehat{V}_{g^+} \begin{pmatrix} \widehat{H}_T' \, \widehat{B}_T' \\ I_s \end{pmatrix}$$

where $\widehat{H}_T = H_T(\widehat{\theta})$.

The test statistic using the estimated residuals would be:

$$\mathrm{SC}(q, s; \widehat{\theta}) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \widehat{w}_t\right) \widehat{V}_{\widehat{w}}^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \widehat{w}_t\right) \xrightarrow{d} \chi^2(s) \tag{2.42}$$

Finally, this test is valid when all of the parameters are well-identified. In the case of weak identification, the above test is no longer valid, since the expansion (2.40) does not hold.[40]

2.6 Summary of the key findings

In this chapter, we reviewed the recent developments in the literature on GMM regression with weak identification, and provided some extensions that are relevant for the estimation of forward-looking rational expectations models. The analysis was based on the approach of Stock and Wright (2000), with a slight generalization of their identification condition to allow for functions rather than subsets of the original parameters being well-identified.

The main contributions were the following.

1. We generalized the identification assumption of Stock and Wright (2000), by defining a re-parameterization of the original parameters θ, into weakly-identified and well-identified subsets, α and β respectively. Except in special cases where some elements of β correspond exactly to elements of θ, we found the GMM estimators of all parameters θ to be of order $O_p(1)$, and derived their asymptotic distribution. This result generalizes similar results in Choi and Phillips (1992) and Stock and Wright (2000).

---

[40] In that case, a boundedly pivotal version of the above test can be derived using a Bonferroni argument over the weakly identified parameters. However, this test will be conservative, and its power could be very low. The details of this are not relevant for the present study, and are therefore dealt with in a separate paper.

2. We provided some useful definitions of instrument relevance as well as nominal versus actual over-identification and under-identification. We also suggested a generalization of the concentration parameter matrix in linear models with serially correlated errors, as a measure of the strength of identification. Since this parameter is generally not consistently estimable under weak identification, it is useful mainly for Monte Carlo experiments in which the DGP is known, as in chapters 3 and 4 below.

3. We discussed the implications of weak identification for conventional GMM-based test statistics, such as the Wald, LR, LM and J statistics. It had already been established in the literature that these statistics are non-pivotal under weak identification and inference based on them can be highly misleading. Some slight contribution was offered by extending the results of Wang and Zivot (1998) to GMM, showing that LR and LM tests are non-pivotal, but can be bounded from above by a χ^2 distribution, leading to valid but conservative inference. More useful are the tests proposed by Kleibergen (2002) and Moreira (2002), which are correctly sized irrespective of the strength of identification and have been shown to have good power properties.

4. Given the need to account for residual serial correlation, which is intrinsic in forward-looking models, we discuss different choices of HAC estimators for the variance of the empirical moment conditions. We show that different estimators will not be asymptotically equivalent at $\theta \neq \theta_0$, and hence the choice of HAC estimator can have implications for the power of the resulting tests.

5. In order to diagnose identification problems, we developed a test of partial identification which is applicable to forward-looking rational expectations models, denoted QMD (Quasi Minimum Distance). This is a test of reduced rank in the coefficient of the first-stage regression of the endogenous regressors on the instruments, correcting for serial correlation and heteroscedasticity in the residuals. It is simpler to compute than the test proposed by Cragg and Donald (1993) and, unlike the latter, it is a pure identification test.

6. Testing the validity of the moment conditions is an important part of the evaluation of a structural model. We showed that the power of the Hansen-Sargan test is driven by some measure of the distance of the moment conditions from zero, which we denoted as the 'mis-specification parameter', ν. Thus, we uncovered the implicit null hypothesis being

tested, namely that $\nu = 0$. Under local mis-specification, this parameter is, of course, not consistently estimable, but it is useful in Monte Carlo settings. In the case of weak identification, the Hansen-Sargan test is known to be non-pivotal, leading to unreliable inference (Staiger and Stock 1997 and Stock and Wright 2000). We derived a boundedly pivotal version of the test which can be used to draw valid but conservative inference.

7. In rational expectations models, serial correlation in the structural residuals will be symptomatic of mis-specification. Hence, mis-specification can be tested indirectly via residual serial correlation tests. However, in forward-looking models, some form of residual serial correlation arises by construction and so we need to test for 'excess' serial correlation (beyond what is implied by the model). We developed such an excess serial correlation test similar to the one proposed in Cumby and Huizinga (1992).

2.A Mathematical appendix

The following regularity conditions are needed:

Assumption A.1. $\mathbf{\Theta}$ *is compact.*

Assumption A.2. *The moment function* $h : \mathcal{Y} \times \mathbf{\Theta} \mapsto \Re^G$ *is measurable on the Borel field of* $\mathcal{Y}$ $\mathcal{B}(\mathcal{Y})$ *for each* $\theta \in \mathbf{\Theta}$*, and it is continuous on* $\mathbf{\Theta}$ *with probability 1.*

Assumption A.3. $h(\mathrm{y}, \theta)$ *is twice continuously differentiable w.r.t.* θ *on* $\mathrm{int}\,(\mathbf{\Theta})$*, and its first and second derivatives are measurable on* $\mathcal{B}(\mathcal{Y})$ *for each* $\theta \in \mathrm{int}\,(\mathbf{\Theta})$*.*

Assumption A.4. *The minimum (minima) of the limiting objective function* $\bar{\mathcal{Q}}(\theta;\theta)$ *is (are) in the interior of* $\mathbf{\Theta}$*.*

Assumption A.5. *The limiting objective function* $\bar{\mathcal{Q}}(\theta;\theta)$ *is twice continuously differentiable on* $\mathbf{\Theta}$*.*

Assumption A.6. $\partial g_T(\theta)/\partial\theta' \equiv D_T(\theta) \xrightarrow{p} \mathcal{J}(\theta)$ *uniformly in* $\theta \in \mathbf{\Theta}$*.*

Proof of theorem 2.1. The proof is very similar to Stock and Wright (2000, Theorem 1). The only point to note is that $\phi(\alpha, \beta + b/\sqrt{T}) \to \phi(\alpha, \beta)$ follows from the continuity of $\phi(\cdot)$, assumption D. The consistency of $\widehat{\beta}$ is proved in Stock and Wright (2000, Lemma A1) and the rest of the proof is identical to the proof of Theorem 1 in that paper. Part (iii) of the theorem follows from the continuity of $\phi(\alpha, \beta_0)$ in α. □

Proof of corollary 2.2. The assumption that there exist $\beta_* \in B$ such that $m_2(\beta_*) + \zeta_f = 0$ and $\mathcal{J}_2(\beta_*)$ has full rank p_2 implies both the global and local identification of β_*. Thus, adapting Stock and Wright (2000, Lemma A.1) accordingly, we establish that $(\widehat{\beta} - \beta_*) = O_p(T^{-1/2})$ and the remaining results follow immediately as for theorem 2.1. □

Lemma 2.8. *Under assumptions A.3, A.5, A.6, B′ with $\mathcal{V}(\theta_0)$ being non-singular, and D with $m_1(\theta)$ also being differentiable at θ_0, with $\mathcal{J}_1 = \frac{\partial}{\partial \theta'} m_1(\theta_0)$:*

$$\begin{gathered}\left(A_T D_T(\theta_0)' V_T(\theta_0)^{-1} D_T(\theta_0) A_T\right)^{-1/2} A_T D_T(\theta_0)' V_T(\theta_0)^{-1} \sqrt{T} g_T(\theta_0) \xrightarrow{d} \\ \xrightarrow{d} \int N\left[\, m(\tilde{\Xi}),\ v(\tilde{\Xi})\right]\, \mathrm{d}F(\tilde{\Xi}),\end{gathered} \tag{2.A.1}$$

where

$$\begin{aligned}
m(\tilde{\Xi}) &= G(\tilde{\Xi})\, \tilde{\mathcal{C}}_1' \tilde{\mathcal{V}}_1^{-1} \left(\mathrm{vec}\left(\tilde{\Xi}\right) - \mathrm{vec}\left(\mathcal{J}_1 \bar{\mathcal{B}}_\perp'\right)\right)\\
v(\tilde{\Xi}) &= G(\tilde{\Xi}) \left(V - \tilde{\mathcal{C}}' \tilde{\mathcal{V}}^{-1} \tilde{\mathcal{C}}\right) G(\tilde{\Xi})'\\
G(\tilde{\Xi}) &= \begin{pmatrix} \tilde{\Xi}' V^{-1} \tilde{\Xi} & \tilde{\Xi}' V^{-1} \mathcal{J}_2 \\ \mathcal{J}_2' V^{-1} \tilde{\Xi} & \mathcal{J}_2' V^{-1} \mathcal{J}_2 \end{pmatrix}^{-1/2} \begin{pmatrix} \tilde{\Xi}' V^{-1} \\ \mathcal{J}_2' V^{-1} \end{pmatrix}\\
\tilde{\mathcal{V}} &= \left(\bar{\mathcal{B}}_\perp' \otimes I_{Gk}\right) \mathcal{V} \left(\bar{\mathcal{B}}_\perp \otimes I_{Gk}\right)\\
\tilde{\mathcal{C}} &= \left(\bar{\mathcal{B}}_\perp' \otimes I_{Gk}\right) \mathcal{C}\\
A_T &= \begin{pmatrix} \sqrt{T}\, \underset{p \times p_1}{\bar{\mathcal{B}}_\perp} & \underset{p \times p_2}{\bar{\mathcal{B}}} \end{pmatrix}\\
\mathcal{B}' &= \tfrac{\partial \beta(\theta_0)}{\partial \theta'}.
\end{aligned}$$

Additionally, $\bar{\mathcal{B}} = \mathcal{B}(\mathcal{B}'\mathcal{B})^{-1}$, $\mathcal{B}_\perp$ is such that $\mathcal{B}_\perp' \mathcal{B} = 0$, and $\bar{\mathcal{B}}_\perp = \mathcal{B}_\perp (\mathcal{B}_\perp' \mathcal{B}_\perp)^{-1}$. The random $Gk \times p_1$ matrix $\tilde{\Xi}$ is Gaussian with mean $\mathcal{J}_1 \bar{\mathcal{B}}_\perp$ and variance matrix

$$\left(\bar{\mathcal{B}}_\perp' \otimes I_{Gk}\right) \mathcal{V}(\theta_0) \left(\bar{\mathcal{B}}_\perp \otimes I_{Gk}\right)$$

Proof. The main idea behind the proof is to express the statistic of interest in terms of two asymptotically independent normal random vectors, η and $\boldsymbol{\xi}_1$, say. Then, conditional on $\boldsymbol{\xi}_1$, the distribution of the statistic is normal, and hence the unconditional distribution is a location-scale mixture of normals, with the mixing occurring over $\boldsymbol{\xi}_1$. This is what Phillips (1989) termed a Limiting Mixed Gaussian (LMG) distribution.[41]

For clarity, we adopt the notation $\nabla_x F(x_0) = \left.\frac{\partial F(x)}{\partial x'}\right|_{x=x_0}$ for the partial derivative of a function F w.r.t. a vector x, evaluated at x_0. We drop the dependence of various quantities on θ_0 for brevity. By assumptions B′, D, and the differentiability of m_1, we have

$$D_T = \mathcal{J} + T^{-1/2} \mathcal{J}_1 + T^{-1/2}\, \Xi + o_p(T^{-1/2})$$

---

[41] The Student's t-distribution is a classic example belonging to this class. See for instance the un-identified IVE example in the introduction.

where $\operatorname{vec}(\Xi) \sim N(0, \mathcal{V}(\theta_0))$ and

$$\mathcal{J} = \underbrace{\nabla_\beta\, m_2(\beta_0)}_{Gk\times p_2}\ \underbrace{\nabla_\theta\, \beta(\theta_0)}_{p_2\times p} = \mathcal{J}_2\, \mathcal{B}.$$

Post-multiplying by A_T, we may re-write the above as:

$$D_T A_T = \begin{pmatrix} \underbrace{(\Xi + \mathcal{J}_1)\bar{\mathcal{B}}_\perp}_{\tilde{\Xi}} & \mathcal{J}_2 \end{pmatrix} + o_p(1).$$

Thus, we see that the matrix that multiplies g_T in equation (2.A.1) satisfies:

$$\left(A_T' D_T' V_T^{-1} D_T\, A_T\right)^{-1/2} A_T' D_T' V_T^{-1} = \begin{pmatrix} \tilde{\Xi}' V^{-1} \tilde{\Xi} & \tilde{\Xi}' V^{-1} \mathcal{J}_2 \\ \mathcal{J}_2' V^{-1} \tilde{\Xi} & \mathcal{J}_2' V^{-1} \mathcal{J}_2 \end{pmatrix}^{-1/2} \begin{pmatrix} \tilde{\Xi}' V^{-1} \\ \mathcal{J}_2' V^{-1} \end{pmatrix} + o_p(1).$$

Next, consider the joint limiting distribution of g_T and $\operatorname{vec}\left(D_T\, \bar{\mathcal{B}}_\perp\right)$ derived from assumption B$'$:

$$\begin{pmatrix} \sqrt{T}\, g_T \\ \sqrt{T}\operatorname{vec}\left(D_T \bar{\mathcal{B}}_\perp'\right) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Psi \\ \boldsymbol{\xi}_1 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \operatorname{vec}\left(\mathcal{J}_1 \bar{\mathcal{B}}_\perp'\right) \end{pmatrix}, \begin{pmatrix} V & \mathcal{C}'\left(\bar{\mathcal{B}}_\perp \otimes I_{Gk}\right) \\ \left(\bar{\mathcal{B}}_\perp' \otimes I_{Gk}\right)\mathcal{C} & \left(\bar{\mathcal{B}}_\perp' \otimes I_{Gk}\right)\mathcal{V}\left(\bar{\mathcal{B}}_\perp \otimes I_{Gk}\right) \end{pmatrix}\right].$$

Next, we orthogonalize Ψ on $\operatorname{vec}\left(\tilde{\Xi}\right)$, and denote the residual:

$$\eta = \Psi - \tilde{\mathcal{C}}' \tilde{\mathcal{V}}^{-1}\left(\operatorname{vec}\left(\tilde{\Xi}\right) - \operatorname{vec}\left(\mathcal{J}\bar{\mathcal{B}}_\perp'\right)\right).$$

By assumption B$'$, η is distributed as normal with zero mean and variance $V - \tilde{\mathcal{C}}'\tilde{\mathcal{V}}^{-1}\tilde{\mathcal{C}}$. Substituting

$$\eta + \tilde{\mathcal{C}}' \tilde{\mathcal{V}}^{-1}\left(\operatorname{vec}\left(\tilde{\Xi}\right) - \operatorname{vec}\left(\mathcal{J}\bar{\mathcal{B}}_\perp'\right)\right).$$

for $\sqrt{T} g_T$ in equation (2.A.1), the result follows. □

Proof of theorem 2.4.

1. Follows from lemma 2.8, by constructing a quadratic form involving the quantity in the lemma and observing that A_T cancels.
2. Follows from theorem 2.1, or by extension of Stock and Wright (2000, Corollary 4(h)).
3. Dropping the dependence on θ_0, the LM statistic can be written as:

$$\begin{aligned} LM &= T\, g_T' V_T^{-1} D_T \left[D_T' V_T^{-1} D_T\right]^{-1} D_T' V_T^{-1} g_T \\ &= T\, g_T' V_T^{-1} g_T - T\, g_T' \left(V_T^{-1} - V_T^{-1} D_T \left[D_T' V_T^{-1} D_T\right]^{-1} D_T' V_T^{-1}\right) g_T \\ &= \underbrace{\mathcal{Q}_T(\theta_0;\theta_0)}_{\chi^2(Gk)} - T\, g_T' V_T^{-1/2} M_{V_T^{-1/2} D_T} V_T^{-1/2} g_T \end{aligned}$$

where $M_{V_T^{-1/2} D_T}$ is the orthogonal projector onto the nullspace of $(V_T^{-1/2} D_T)$. $\mathcal{Q}_T(\theta_0;\theta_0) \xrightarrow{p} \chi^2(Gk)$ follows from theorem 2.3. The second term is non-negative w.pr. 1, and it is zero when $Gk = p$, i.e., when the nullspace of $(V_T^{-1/2} D_T)$ is empty.

Similarly for the LR_{CUE} we see that it consists of the asymptotically χ^2 term $\mathcal{Q}_T(\theta_0; \theta_0)$ minus $\tilde{\mathcal{Q}}(\alpha^*; \alpha^*)$, which is non-negative w.pr. 1, and zero when $Gk = p$.

□

Proof of lemma 2.5. The restricted matrix Π_r can be written as $B\, d'$, but B and d are not identified without further restrictions. One such normalization is (re-ordering the columns of d if necessary):

$$\overset{r\times p}{d'} = (I_r, \overset{r\times p-r}{d_2}).$$

The number of free parameters is kr from B and $r(p-r)$ from d, so we can stack them into the $m = r(k+p-r)$ dimensional vector $\mu = \left(\text{vec}\,(B)', \text{vec}\,(d_2)'\right)'$, such that $\Pi_r = \Pi(\mu)$. Hence $\widehat{\Pi}_r = \Pi(\widehat{\mu})$, where $\widehat{\mu}$ (the QMLE) is asymptotically equivalent to the minimizer of the objective function:

$$\bar{D}_T(\Pi(\mu))' \mathcal{W} \bar{D}_T(\Pi(\mu))$$

where $\mathcal{W} = (\Sigma_{vv} \otimes \Sigma_{ZZ})^{-1} = \operatorname*{plim}_{T\to\infty} \left(\widehat{\Sigma}_{vv} \otimes \widehat{\Sigma}_{ZZ}\right)^{-1}$ by assumption. By standard results (c.f. Newey and McFadden (1994)), and letting $\mathcal{V}_0 = \mathcal{V}(\Pi(\mu_0))$

$$\sqrt{T}\,(\widehat{\mu} - \mu) \xrightarrow{d} N\left[0, \mathcal{J}_{\bar{D}}\,(\mathcal{J}_{\bar{D}}\,\mathcal{W}\,\mathcal{J}_{\bar{D}})^{-1}\,\mathcal{J}_{\bar{D}}'\mathcal{W}\,\mathcal{V}_0\,\mathcal{W}\,\mathcal{J}_{\bar{D}}\,(\mathcal{J}_{\bar{D}}\,\mathcal{W}\,\mathcal{J}_{\bar{D}})^{-1}\,\mathcal{J}_{\bar{D}}'\right]$$

where

$$\begin{aligned}
\mathcal{J}_{\bar{D}} &= \operatorname*{plim}_{T\to\infty} \left.\frac{\partial \bar{D}_T(\Pi(\mu))}{\partial \mu'}\right|_{\mu_0} \\
&= \operatorname*{plim}_{T\to\infty} \left.\frac{\partial \bar{D}_T(\Pi(\mu))}{\partial \text{vec}\,(\Pi)'}\,\frac{\partial \text{vec}\,(\Pi)}{\partial \mu'}\right|_{\mu_0} \\
&= -\,(\Sigma_{ZZ} \otimes I_p)\left(d \otimes I_k \quad \begin{pmatrix} 0_{kr\times r(p-r)} \\ (B \otimes I_{p-r})\,K_{r,p-r} \end{pmatrix}\right),
\end{aligned}$$

where $K_{r,p-r}$ is the commutation matrix.[42] The third line follows by applying the result $\frac{\partial \text{vec}(AXB)}{\partial \text{vec}(X)'} = (B' \otimes A)$.

Next, we establish the limiting distribution of $\bar{D}_T(\widehat{\Pi}_r)$, by performing a Taylor expansion over μ to order $T^{-1/2}$:

$$\begin{aligned}
\sqrt{T}\,\bar{D}_T(\widehat{\Pi}_r) &= \sqrt{T}\,\bar{D}_T(\Pi_0) + \mathcal{J}_{\bar{D}}\,\sqrt{T}\,(\widehat{\mu} - \mu) + o_p(1) \\
&\xrightarrow{d} N\left[0, M_{\mathcal{J}_{\bar{D}}}\,\mathcal{V}_0\,M_{\mathcal{J}_{\bar{D}}}'\right]
\end{aligned}$$

---

[42] Let A be a $m \times n$ matrix. Then $K_{m,n}$ is the unique $mn \times mn$ matrix that transforms vec (A) into vec (A'), i.e., $K_{m,n}\text{vec}\,(A) = \text{vec}\,(A')$.

where $M_{\mathcal{J}_{\bar{D}}} = I_{kp} - \mathcal{J}_{\bar{D}} \left(\mathcal{J}_{\bar{D}} \mathcal{W} \mathcal{J}_{\bar{D}}\right)^{-1} \mathcal{J}_{\bar{D}}' \mathcal{W}$ is an idempotent matrix (a skewed projector onto $\mathrm{Col}^{\perp}\{\mathcal{J}_{\bar{D}}\}$), whose rank is equal to its trace, $kp - m = (k-r)(p-r)$. Hence, we can find orthogonal matrices S, Q of dimension $kp \times m$ such that $M_{\mathcal{J}_{\bar{D}}} = S\,Q'$. Then,

$$\sqrt{T}\, S' \bar{D}_T(\widehat{\Pi}_r) \xrightarrow{d} N\left[0, Q' \mathcal{V}_0\, Q\right]$$

and the results follows. □

Proof of theorem 2.6. 1. Application of theorem 2.1.

2. Using the fact that $m_1(\phi(\alpha_0, \beta_0)) = 0$, by assumption D, we have from part 1:

$$\begin{aligned} \mathcal{Q}_T(\phi(\alpha_0, \widehat{\beta}_c); \phi(\alpha_0, \widehat{\beta}_c)) &\xrightarrow{d} [\Psi(\theta_0) + \mathcal{J}_2(\beta_0)\, b^*]'\, V(\theta_0)^{-1} [\Psi(\theta_0) + \mathcal{J}_2(\beta_0)\, b^*] \\ &= \left[V(\theta_0)^{-1/2} \Psi(\theta_0)\right]' M_{V(\theta_0)^{-1/2} \mathcal{J}_2(\beta_0)} \left[V(\theta_0)^{-1/2} \Psi(\theta_0)\right] \\ &\sim \chi^2(Gk - p_2) \end{aligned}$$

where $M_X = I - X(X'X)^{-1}X'$ is an orthogonal projector onto $\mathrm{Col}^{\perp}(X)$, and the last equality follows from the fact that $\mathcal{J}_2(\beta_0)$ is of rank p_2, by assumption D, and $V(\theta_0)^{-1/2}\Psi(\theta_0)$ is Gk-dimensional standard normal, by assumption B. Since

$$\mathcal{Q}_T(\widehat{\theta}_c; \widehat{\theta}_c) = \mathcal{Q}_T(\phi(\widehat{\alpha}, \widehat{\beta}); \phi(\widehat{\alpha}, \widehat{\beta})) \leq \mathcal{Q}_T(\phi(\alpha_0, \widehat{\beta}); \phi(\alpha_0, \widehat{\beta})) \tag{2.A.2}$$

the result follows.

□

Proof of lemma 2.7. First, we note that under the assumptions of the lemma, $(\widehat{\beta} - \beta_0) = O_p(T^{-1/2})$, by standard results, Hansen (1982). By the continuity of the argmin operator, and the uniform convergence assumptions, the limiting distribution of $\widehat{\beta}$ can be derived from the limiting objective function. First, we perform a Taylor expansion of $g_T(\beta_0 + T^{-1/2}\, b)$

$$\begin{aligned} \sqrt{T}\, g_T(\beta_0 + T^{-1/2}\, b) &\equiv \Psi_T(\beta_0) + \zeta_l + \sqrt{T}\, \tilde{m}_T(\beta_0 + T^{-1/2}\, b) \\ &= \Psi(\beta_0) + \zeta_l + \mathcal{J}(\beta_0)\, b + o_p(1) \end{aligned}$$

since $\tilde{m}(\beta_0) = 0$ and $\nabla_\beta\, m_T(\beta_0) \xrightarrow{p} \mathcal{J}(\beta_0)$ by assumption. Hence, the limiting objective function $\bar{\mathcal{Q}}(b)$ is:

$$\begin{aligned} \mathcal{Q}_T(\beta_0 + b/\sqrt{T}; \beta_0) &= T\, g_T\left(\beta_0 + b/\sqrt{T}\right)' \widehat{V}_T(\beta_0)^{-1} g_T\left(\beta_0 + b/\sqrt{T}\right) \\ &\Rightarrow [\Psi(\beta_0) + \zeta_l + \mathcal{J}(\beta_0)\, b]'\, V(\beta_0)^{-1}\, [\Psi(\beta_0) + \zeta_l + \mathcal{J}(\beta_0)\, b] \equiv \bar{\mathcal{Q}}(b) \end{aligned}$$

By van der Vaart and Wellner (1996, Lemma 3.2.1), $\sqrt{T}(\widehat{\beta} - \beta_0) \xrightarrow{d} \operatorname{argmin}_b \bar{\mathcal{Q}}(b)$. Dropping the dependence on β_0 for compactness:

$$\sqrt{T}(\widehat{\beta} - \beta_0) = -\left(\mathcal{J}'\, V^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}'\, V^{-1}\, (\Psi + \zeta_l)$$

Hence, we may derive the limiting distribution of $g_T(\widehat{\beta})$, by means of a Taylor expansion to order $O_p(T^{-1/2})$

$$\begin{aligned} \sqrt{T}\, g_T(\widehat{\beta}) &= \sqrt{T}\, g_T(\beta_0) + \mathcal{J}(\beta_0)\, \sqrt{T}(\widehat{\beta} - \beta_0) + o_p(1) \\ &= \left[I_{Gk} - \mathcal{J}\left(\mathcal{J}'\, V(\beta_0)^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}'\, V(\beta_0)^{-1}\right] (\Psi + \zeta_l) + o_p(1) \\ &\xrightarrow{d} N\left[M_1\, \zeta_l,\ M_1\, V\, M_1'\right] \end{aligned}$$

where $M_1 = M_{\mathcal{J} \perp V^{-1}\mathcal{J}} = I_{Gk} - \mathcal{J}\left(\mathcal{J}'\, V^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}'\, V^{-1}$ is a skewed projector onto $\mathrm{Col}^{\perp}\{\mathcal{J}\}$. Now, observe that

$$M_1\, V\, M_1' = V - \mathcal{J}\left(\mathcal{J}'\, V^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}' = V^{1/2}\, M_{V^{-1/2}\mathcal{J}}\, V^{1/2} = V^{1/2}\, M_2\, V^{1/2}$$

where $M_2 = M_{V^{-1/2}\mathcal{J}} = I_{Gk} - V^{-1/2}\, \mathcal{J}\left(\mathcal{J}'\, V^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}'\, V^{-1/2}$ is an orthogonal projector onto $\mathrm{Col}^{\perp}\{V^{-1/2}\mathcal{J}\}$, i.e., idempotent, of rank $Gk - p$. Since its eigenvalues are 1 and 0, M_2 can be decomposed into $P\,P'$, such that $P'P = I_{Gk-p}$. Also, it is easily verified that $V^{-1/2}\, M_1 = M_2\, V^{-1/2}$. Then,

$$\sqrt{T}\, g_T(\widehat{\beta}) \overset{a}{=} V^{1/2} P\, z, \quad z \sim N(\mu, I_{Gk-p}), \quad \mu = P'\, V^{-1/2}\, \zeta_l$$

which establishes that

$$\mathcal{Q}_T(\widehat{\beta}; \widehat{\beta}) = T\, g_T(\widehat{\beta})' \widehat{V}_T(\widehat{\beta})^{-1}\, g_T(\widehat{\beta}) \xrightarrow{d} z' P' P\, z = z'z \sim \chi^2(Gk - p, \nu^2)$$

where

$$\nu^2 = \zeta_l'\, V^{-1/2} P\, P' V^{-1/2}\, \zeta_l = \zeta_l'\, V^{-1/2} M_2 V^{-1/2}\, \zeta_l = \zeta_l' \left[V^{-1} - V^{-1}\mathcal{J}\left(\mathcal{J}'\, V^{-1}\, \mathcal{J}\right)^{-1} \mathcal{J}'\, V^{-1}\right] \zeta_l$$

as required. □

Chapter 3

Monte Carlo simulation

In this chapter we offer some simulation evidence on the finite-sample properties of the asymptotic approximations and inference procedures discussed in the previous chapter when: (a) only some functions of the parameters are well-identified; (b) instruments are not strongly exogenous; and (c) the residuals exhibit finite-order serial correlation.[1]

First, we wish to assess the quality of the local-to-zero approximations. This has been investigated in various contexts, see e.g., Staiger and Stock (1997) for IV and LIML estimators in static linear models, and Stock and Wright (2000) for non-linear Consumption CAPM models. Here, we wish to extend this framework to allow for situations in which only some functions of the original parameters are identified, and also evaluate the implications of using lagged endogenous variables as instruments, as well as various HAC covariance estimators for the quality of these asymptotic approximations, and the finite-sample coverage probabilities of robust confidence sets, such as the S-set.

Then, we assess the size of the Quasi Minimum Distance (QMD) test for identification, and we compare it with the uncorrected likelihood ratio Reduced Rank (RR) test. We find that, albeit far from perfect, this test is much better sized than the naive RR test, and it is quite powerful (in a sense 'too powerful') under local alternatives.

Next, we simulate the Hansen-Sargan test, assessing the finite-sample size of the boundedly pivotal version we propose in theorem 2.6 under weak identification, and the implications of over-instrumenting and various HAC estimators for its size in general.

---

[1] Simulation studies of this setup have also been performed elsewhere, see West and Wilcox (1996) and West (1997), but the focus was not on the statistics we are interested in here.

Finally, we assess the finite-sample size of various versions of the excess serial correlation statistic $\mathrm{SC}(q, s)$ and study its power against a relevant alternative.

The structure of the chapter is as follows. Section 3.1 provides the details of the Monte Carlo design. Section 3.2 assesses the quality of the asymptotic approximations provided by theorem 2.1. The following section evaluates the finite-sample coverage rates of the S-set, which is robust to weak identification. Next, section 3.4 analyzes the finite-sample size of the Hansen-Sargan test, and the following two sections analyze the size and power of the QMD and excess serial correlation tests. Finally, section 3.7 summarizes the key findings. Details of the derivation of the asymptotic approximations are given in appendix 3.B, and a new result on variance reduction, which is applied here, is given in appendix 3.A.

3.1 Simulation Experiment Design

For the Monte Carlo experiments of this chapter, we need to specify the following:

1. The Data Generating Process (DGP).
2. The structural model, and the set of instruments used to estimate it.
3. The estimation method: m-step or continuously updated (CUE) GMM.
4. The weighting matrix: parametric or nonparametric HAC covariance matrix estimator.
5. The parameters of interest and the nuisance parameters.
6. The statistics of interest, to be simulated.

The DGP

The DGP will consist of two sets of variables, y and $\mathbf{z}$, where the latter are *not* Granger-caused by the former, thus serving as a pool of strongly exogenous instruments for the parameters of interest, which can be kept fixed in repeated samples. The DGP is parameterized as a VAR(1).

$$\begin{aligned} \mathrm{y}_t &= A_1 \mathrm{y}_{t-1} + A_z \mathbf{z}_t + \mathbf{e}_t \\ \mathbf{z}_t &= C\,\mathbf{z}_{t-1} + \mathbf{v}_t \end{aligned} \tag{3.1}$$

where y_t is a vector of variables to be used as endogenous in the models, $\mathbf{z}_t$ is a $k \times 1$ vector of 'exogenous' variables that can be used as instruments, and

$$\begin{pmatrix} \mathbf{e}_t \\ \mathbf{v}_t \end{pmatrix} \sim \text{NID}\left[0, \begin{pmatrix} \Sigma_{\mathbf{ee}} & 0 \\ 0 & I_k \end{pmatrix}\right]$$

The normalization of $\Sigma_{\mathbf{vv}} = I_k$ is innocuous, see the discussion on invariances of our statistics of interest to certain nuisance parameters in appendix 3.A. The normality assumption is maintained here because our focus is on the implications of using non-strongly exogenous instruments and serial correlation corrections.

This DGP includes Design I of Staiger and Stock (1997) and the New Keynesian Philips curve as special cases.[2] It provides a link between the two (i.e., from a static to a dynamic DGP) as well as allows us to investigate the case of more than one endogenous variable coefficients, a function of which may be poorly identified.

The Model

The model will be a linear structural equation that contains three types of variables: the regressand, the regressors, and the instruments. The notation will be as follows:

Mnemonic	Description	Coefficient
y_t	regressand	-
Y_t	endogenous regressors	θ_1
Z_{1t}	exogenous regressors	θ_2
Z_{2t}	additional instruments	-

Note the difference between the exogenously generated variables $\mathbf{z}_t$ and the notation for instruments Z_t. The latter denotes the *status* of a variable in the model, i.e. what is used instruments, which is in line with the notational convention in the literature. For instance, the instrument set could contain lags of y_t, or additional variables from $\mathbf{z}_t$, the difference being that the latter can be kept fixed in repeated samples (strongly exogenous), whereas the former cannot. Similarly, y_t and Y_t denote the status of the regressand and endogenous regressors in the model. For instance, the first variable in y_t could be the regressand ($\mathrm{y}_{1t} \to y_t$), and the second and third variables could be regressors $(\mathrm{y}_{2t}, \mathrm{y}_{3t})' \to Y_t$; or, in a forward-looking model we would have $(\mathrm{y}_{2t}, \mathrm{y}_{1,t+1})' \to Y_t$, say.

---

[2] Staiger and Stock (1997, Design I) is essentially our Design I, see below, with the only difference in the distribution of the structural errors.

It is also useful to introduce some notation for all of the regressors, both endogenous and exogenous, i.e. everything that appears on the right hand side of the model, $\bar{Y}_t = (Y_t', Z_{1t}')'$. Thus, the generic structural model can be written as:

$$y_t = \theta' \bar{Y}_t + u_t = \theta_1' Y_t + \theta_2' Z_{1t} + u_t \tag{3.2}$$

where $\theta = (\theta_1', \theta_2')'$ has been partitioned accordingly. The moment functions are $h_t(\theta) = y_t - \theta_1' Y_t - \theta_2' Z_{1t}$ and the corresponding orthogonality conditions are:

$$\mathrm{E}\,(h_t(\theta)\, Z_t) = 0. \tag{3.3}$$

Whether these conditions hold depends of course on the parameters of the DGP. When they hold at some unique parameter value θ_0, we will call $u_t = h_t(\theta_0)$ the 'structural error'.

The 'first-stage' or 'forecasting' regression for the endogenous variables is

$$Y_t = \Pi' Z_t + v_t = \Pi_1' Z_{1t} + \Pi_2' Z_{2t} + v_t \tag{3.4}$$

with $\mathrm{E}\,(Z_t\, v_t') = 0$. However, in order to treat all of the parameters jointly, and emphasize the implication for identification, the following equation will be more useful:

$$\bar{Y}_t = \bar{\Pi}' Z_t + \bar{v}_t = \begin{pmatrix} \Pi_1' & \Pi_2' \\ I & 0 \end{pmatrix} \begin{pmatrix} Z_{1t} \\ Z_{2t} \end{pmatrix} + \begin{pmatrix} v_t \\ 0 \end{pmatrix} \tag{3.5}$$

Estimation method and weighting matrix

The model will be estimated by efficient GMM, using a weighting matrix $W_T(\bar{\theta}_T) = \widehat{V}_T(\bar{\theta}_T)^{-1}$, i.e., the inverse of a HAC estimator of the covariance of the moment conditions (3.3), evaluated at $\bar{\theta}_T(\theta)$. The latter will denote different types of estimators, e.g., $\bar{\theta}_T = \widehat{\theta}_T^{(m-1)}$ with $m = 2, 3, \ldots$ is the m-step estimator, whereas $\bar{\theta}_T = \theta$ is the CUE:

$$\widehat{V}_T(\bar{\theta}_T) = \begin{cases} \widehat{\sigma}_u^2\, \widehat{\Sigma}_{ZZ} & \text{1-step estimator.} \\ \widehat{V}_T\left(\widehat{\theta}_T^{(m-1)}\right), & m\text{-step estimators, } m = 2, 3, \ldots \\ \widehat{V}_T(\theta), & \text{continuously updated estimator.} \end{cases}$$

where $\widehat{\sigma}_u^2$ is the sample variance of the structural error, which is a function of $\widehat{\theta}$.[3] The most common estimation methods are 2-Step GMM and the CUE, and these are the ones that we shall focus on in the ensuing experiments. In a linear model with no heteroscedasticity or autocorrelation correction, these are the 2SLS and LIML estimators respectively.

---

[3] This clearly does not affect the estimation of θ, except for the CUE, but is necessary for correct inference based on the objective function.

To complete the estimation settings, particular HAC estimators must be specified. In view of the discussion in section 2.3, we shall consider the following alternatives:

Notation	Description
HOMO	conditional homoscedasticity and no autocorrelation for u_t, $\widehat{V}(\theta) = \widehat{\Sigma}_{ZZ}\,\widehat{\sigma}^2_u(\theta)$
HC	Heteroscedasticity consistent estimator (White)
NW	Heterosc. and autocorr. consistent estimator (Newey-West)
MA-l	parametric HAC estimator (West)

Newey-West's NW estimator uses Bartlett weights and automatic bandwidth selection, unless otherwise specified, see section 2.3.2 for details.

The parameters

The parameters $\theta, \sigma^2_u, \Pi, \Sigma_{vv}, \Sigma_{uv}$, as well as the autocovariance function $\Gamma_{x,j}$ of $x = u_t, v_t$ and Z_t are functions of the original parameters of the DGP (3.1), and they can be computed as explained in appendix 3.B.

These parameters can be partitioned into two sets. The *parameters of interest* are (θ, σ^2_u), whereas the remaining are *nuisance parameters*. In view of the asymptotic theory of chapter 2, we expect only certain functions of the nuisance parameters to matter. These are: the concentration parameter matrix or its eigenvalues μ^2_i, relative to the number of instruments k, which govern the bias of the GMM estimators and the speed of convergence of their distributions to normality; the endogeneity parameter λ, which affects the shape of the distributions and determines the bias of Least Squares estimators;[4] and the variance matrix of the moment conditions, $V(\theta)$. In our experiments, we will allow for first-order serial correlation in the structural errors u_t and forecast errors v_t, so that the respective first-order autocovariances $\Gamma_{u,1}$ and $\Gamma_{v,1}$ will also be important.

The statistics

Our simulations will focus on: (i) the 2-step and CUE estimators or θ, (ii) the LR and J tests, (iii) the coverage rate of S-sets, and (iv) the size and power of the QMD and excess SC test.

The impact of using non-strongly exogenous instruments, such as lags of the dependent variables, and of the different HAC estimators on the finite-sample coverage rates of robust confidence sets, and the power function of the J-test, will be investigated using various designs.

---

[4] The endogeneity parameter effectively determines the implied probability limits of the respective OLS estimators and hence their bias relative to the true value.

The different designs

Four different designs will be used to make the transition from the prototype one-parameter static model with strongly exogenous instruments, Staiger and Stock (1997, Design I), to the two-parameter dynamic model with serially-correlated errors and only one *combination* of the parameters being well-identified, akin to a partial adjustment mechanism.

The various designs will differ either in the estimated models (one or two parameters), in the type of instruments used (strongly exogenous or lagged regressors), and in the autocorrelation of the structural error u_t (zeroth or first-order), see table 3.1. The details of the different designs are given in appendix 3.C, where the reader can find the exact specification of all of the parameters of the DGP (3.1). This information should be sufficient to replicate our analysis.

Table 3.1: The four Monte Carlo Designs, see appendix 3.C for details.

Design	Model	Instruments	DGP	SC
I	$y_t = \theta\, Y_t + u_t$	$Z_t = (\mathbf{z}_{1t}, \ldots, \mathbf{z}_{kt})$	Static	NO
II	$y_t = \theta\, Y_t + u_t$	$Z_t = (y_{t-1}, \ldots, y_{t-k})$	Dynamic	NO
III	$y_t = \theta_1\, Y_t + \theta_2\, Z_{1t} + u_t$	$Z_{1t} = \mathbf{z}_{1t}$, $Z_{2t} = (\mathbf{z}_{2t}, \ldots, \mathbf{z}_{kt})$	Static	NO
IV	$y_t = \theta_1 y_{t+1} + \theta_2\, y_{t-1} + u_t$	lagged y_t, or $\mathbf{z}_t$	Dynamic	YES

The following will be common in all designs. The number of instruments, k, will vary between the number of estimated parameters (1 for designs I and II and 2 for designs III and IV), 4 and 8. For the two-parameter models (designs III and IV) the degree of under-identification will be at most 1, meaning that one linear combination of the parameters will always be well-identified. For all designs, the minimum eigenvalue of the concentration parameter matrix per instrument, μ^2_{min}/k will vary between 0, 1 and 10 (0, low and high). The endogeneity parameter will be set to 0.5 (or close to it), and the true value of the weakly identified parameter combination α_0 will be set to zero unless otherwise specified. The random numbers used in each experiments will be common, in order to limit variation across experiments.

The benchmark case will be a DGP with fixed and strongly exogenous instruments, where the finite-sample distributions coincide exactly with the asymptotic ones. This case will also be used to derive control variables in order to reduce Monte Carlo uncertainty, see appendix 3.A. All the relevant code for the computations in these experiments was developed by the author, see appendix A.

3.2 Quality of local-to-zero approximation

In this section, we assess the quality of the first-order asymptotic approximation to the distributions of estimators and test statistics. Tables 3.2 through 3.5 report the maximum absolute difference between the finite-sample and asymptotic cumulative distribution functions of 2-step and CUE parameter estimators and various test statistics for all four designs described above. These are also known as the Kolmogorov-Smirnov statistics testing the equality of the two distributions.

Panel (a) of table 3.2 represents a minor departure from the benchmark case, where the local-to-zero approximations correspond to the exact finite-sampling distributions.[5] This is verified by the KS statistics, all lying within the acceptance region for the two distributions being equal. Panels (b) and (c) investigate the impact of using various heteroscedasticity and autocorrelation consistent estimators (which are unnecessary here). As expected, the quality of the approximations deteriorates slightly, but not substantially. Also, the impact of the HC and HAC estimators falls with the sample size.

Next, the impact of using non-strongly exogenous instruments can be seen by comparing tables 3.2 and 3.3. Looking at the HOMO version in both designs (panel (a)), we see that the only noticeable difference arises at a sample size of 100 in the un-identified case ($\mu^2/k = 0$) when 4 or 8 instruments are used, e.g., the KS statistic for $\widehat{\theta}_{2S}$ jumps to 0.026 from 0.006, still being quite small. This effect clearly disappears as the sample size grows and is not evident in panels (b) and (c) either.[6] This evidence suggests that the local-to-zero asymptotic distributions provide good approximation to the finite-sampling distributions of estimators and tests at moderately large sample sizes, even when the instruments are not strongly exogenous.

Let us now turn to the distribution of GMM estimators when a *combination* of the estimated parameters is well-identified. We assess the quality of the first-order approximations in the static case with strongly exogenous instruments by comparing panel (a) in tables 3.2 and 3.4. The KS statistics for the two-parameter model of Design III look very similar to the benchmark case, verifying the conclusions of theorem 2.1. The only notable exceptions arise at sample size 100 and $\mu^2/k = 10$, where the KS statistic for $\widehat{\theta}_2^{2S}$ is 0.085, well above the 1% critical value of 0.011. However, this deviation seems to be contributed by higher-order terms, since it decreases with the

---

[5] The only difference is that instruments are stochastic, rather than fixed.

[6] Unreported evidence shows that the effect of using lagged regressors as instruments is more noticeable at very small sample sizes, such as $T = 10$ or $T = 20$. However, we do not find this of practical import for macroeconomic time series.

Table 3.2: Maximum Absolute Difference between Finite-Sample CDF and Asymptotic CDF for Design I.

		$T = 100$						$T = 200$						$T = 500$					
μ^2/k	k	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}
(a) HOMO																			
0	1	0.006	0.007	-	0.006	0.007	-	0.004	0.007	-	0.004	0.007	-	0.010	0.008	-	0.010	0.008	-
0	4	0.004	0.010	0.015	0.004	0.007	0.017	0.005	0.007	0.006	0.006	0.006	0.008	0.007	0.005	0.004	0.006	0.003	0.007
0	8	0.006	0.012	0.033	0.007	0.015	0.033	0.006	0.011	0.017	0.006	0.013	0.017	0.010	0.008	0.012	0.009	0.005	0.008
1	1	0.006	0.007	-	0.006	0.007	-	0.006	0.007	-	0.006	0.007	-	0.005	0.008	-	0.005	0.008	-
1	4	0.006	0.006	0.010	0.008	0.007	0.013	0.004	0.007	0.006	0.008	0.005	0.006	0.006	0.006	0.007	0.009	0.004	0.007
1	8	0.004	0.008	0.027	0.006	0.012	0.027	0.004	0.012	0.011	0.007	0.009	0.014	0.009	0.006	0.015	0.007	0.005	0.017
10	1	0.006	0.007	-	0.006	0.007	-	0.006	0.007	-	0.006	0.007	-	0.007	0.008	-	0.007	0.008	-
10	4	0.006	0.006	0.012	0.007	0.007	0.014	0.005	0.005	0.004	0.008	0.006	0.004	0.009	0.006	0.007	0.013	0.005	0.010
10	8	0.007	0.013	0.027	0.005	0.017	0.027	0.007	0.009	0.010	0.005	0.011	0.012	0.011	0.006	0.016	0.008	0.007	0.016
(b) HC																			
0	4	0.008	0.013	0.029	0.018	0.010	0.030	0.005	0.010	0.011	0.013	0.008	0.012	0.008	0.009	0.006	0.007	0.007	0.010
0	8	0.014	0.024	0.053	0.025	0.024	0.057	0.008	0.012	0.032	0.022	0.013	0.027	0.012	0.010	0.014	0.013	0.007	0.011
1	4	0.008	0.034	0.022	0.017	0.012	0.021	0.007	0.023	0.009	0.013	0.008	0.010	0.008	0.012	0.010	0.014	0.007	0.009
1	8	0.014	0.059	0.042	0.041	0.027	0.041	0.011	0.045	0.021	0.022	0.014	0.023	0.009	0.015	0.018	0.013	0.009	0.017
10	4	0.010	0.031	0.022	0.014	0.011	0.023	0.007	0.027	0.008	0.010	0.005	0.009	0.009	0.020	0.010	0.013	0.005	0.013
10	8	0.018	0.054	0.042	0.031	0.016	0.041	0.013	0.048	0.020	0.015	0.005	0.022	0.013	0.034	0.018	0.009	0.004	0.018
(c) HAC(1) (Newey-West)																			
0	4	0.008	0.018	0.030	0.018	0.021	0.032	0.006	0.011	0.012	0.014	0.012	0.014	0.009	0.009	0.007	0.008	0.006	0.011
0	8	0.014	0.035	0.048	0.024	0.053	0.054	0.010	0.017	0.031	0.023	0.026	0.024	0.014	0.011	0.012	0.013	0.008	0.011
1	4	0.009	0.037	0.023	0.018	0.015	0.023	0.007	0.025	0.011	0.014	0.007	0.011	0.009	0.014	0.010	0.014	0.006	0.010
1	8	0.016	0.072	0.049	0.049	0.023	0.048	0.012	0.051	0.024	0.027	0.011	0.022	0.010	0.017	0.015	0.013	0.007	0.015
10	4	0.010	0.031	0.023	0.016	0.010	0.023	0.008	0.029	0.009	0.011	0.005	0.009	0.008	0.020	0.010	0.013	0.005	0.014
10	8	0.020	0.058	0.044	0.036	0.013	0.050	0.013	0.051	0.021	0.017	0.012	0.023	0.014	0.036	0.017	0.011	0.007	0.015

Entries are the Kolmogorov-Smirnov statistics testing the null hypothesis that the two distributions are equal. The critical values are: 0.0096 at 5% and 0.0115 at 1%. Asymptotic distributions where computed using 100,000 replications of the representations in theorem 2.1, see appendix 3.B. Finite-sample distributions were computed using 20,000 Monte Carlo replications.

Table 3.3: Maximum Absolute Difference between Finite-Sample CDF and Asymptotic CDF for Design II.

		$T = 100$						$T = 200$						$T = 500$					
μ^2/k	k	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}	$\hat{\theta}_{2S}$	LR_{2S}	J_{2S}	$\hat{\theta}_{CU}$	LR_{CU}	J_{CU}
(a) HOMO																			
0	1	0.004	0.007	-	0.004	0.007	-	0.005	0.011	-	0.005	0.011	-	0.003	0.007	-	0.003	0.007	-
0	4	0.013	0.015	0.008	0.009	0.016	0.011	0.012	0.011	0.004	0.008	0.013	0.007	0.009	0.009	0.005	0.006	0.009	0.008
0	8	0.026	0.027	0.016	0.024	0.036	0.014	0.013	0.012	0.011	0.011	0.020	0.008	0.007	0.008	0.008	0.005	0.011	0.010
1	1	0.008	0.007	-	0.008	0.007	-	0.009	0.011	-	0.009	0.011	-	0.005	0.008	-	0.005	0.008	-
1	4	0.009	0.010	0.007	0.006	0.013	0.006	0.007	0.010	0.008	0.010	0.011	0.010	0.007	0.013	0.005	0.007	0.013	0.006
1	8	0.016	0.019	0.028	0.007	0.024	0.025	0.009	0.011	0.014	0.010	0.016	0.010	0.005	0.009	0.003	0.006	0.011	0.007
10	1	0.007	0.007	-	0.007	0.007	-	0.010	0.011	-	0.010	0.011	-	0.006	0.008	-	0.006	0.008	-
10	4	0.007	0.011	0.010	0.009	0.013	0.009	0.010	0.013	0.007	0.013	0.016	0.006	0.005	0.009	0.006	0.006	0.011	0.006
10	8	0.006	0.023	0.029	0.009	0.023	0.028	0.007	0.017	0.016	0.012	0.017	0.015	0.005	0.010	0.006	0.008	0.011	0.005
(b) HC																			
0	4	0.017	0.020	0.026	0.017	0.012	0.026	0.012	0.009	0.012	0.012	0.008	0.012	0.009	0.005	0.007	0.007	0.007	0.010
0	8	0.026	0.022	0.036	0.022	0.030	0.045	0.018	0.017	0.016	0.020	0.016	0.020	0.009	0.017	0.008	0.012	0.010	0.016
1	4	0.010	0.034	0.017	0.018	0.018	0.021	0.007	0.022	0.012	0.014	0.006	0.016	0.005	0.008	0.008	0.007	0.006	0.009
1	8	0.016	0.063	0.032	0.032	0.018	0.033	0.009	0.040	0.015	0.019	0.008	0.016	0.004	0.022	0.010	0.008	0.005	0.011
10	4	0.008	0.032	0.018	0.013	0.007	0.017	0.009	0.021	0.009	0.011	0.011	0.009	0.004	0.015	0.005	0.005	0.011	0.006
10	8	0.012	0.053	0.029	0.026	0.011	0.032	0.007	0.041	0.014	0.013	0.008	0.015	0.005	0.026	0.009	0.006	0.009	0.009
(c) HAC(1) (Newey-West)																			
0	4	0.017	0.025	0.029	0.016	0.022	0.031	0.013	0.012	0.015	0.015	0.012	0.016	0.009	0.006	0.007	0.007	0.007	0.010
0	8	0.027	0.037	0.035	0.024	0.056	0.045	0.019	0.023	0.017	0.023	0.025	0.021	0.011	0.017	0.008	0.013	0.011	0.015
1	4	0.009	0.044	0.023	0.026	0.018	0.025	0.004	0.027	0.015	0.016	0.009	0.017	0.004	0.015	0.008	0.010	0.005	0.011
1	8	0.013	0.080	0.061	0.051	0.022	0.060	0.009	0.055	0.029	0.030	0.009	0.029	0.006	0.028	0.010	0.012	0.007	0.010
10	4	0.011	0.031	0.023	0.019	0.010	0.023	0.011	0.023	0.013	0.011	0.009	0.012	0.005	0.020	0.006	0.006	0.008	0.007
10	8	0.015	0.057	0.054	0.040	0.013	0.058	0.010	0.044	0.026	0.019	0.009	0.026	0.008	0.037	0.009	0.009	0.004	0.008

Entries are the Kolmogorov-Smirnov statistics testing the null hypothesis that the two distributions are equal. The critical values are: 0.0096 at 5% and 0.0115 at 1%. Asymptotic distributions where computed using 100,000 replications of the representations in theorem 2.1, see appendix 3.B. Finite-sample distributions were computed using 20,000 Monte Carlo replications.

Table 3.4: Maximum Absolute Difference between Finite-Sample CDF and Asymptotic CDF for Design III.

		$T = 100$						$T = 200$						$T = 500$					
μ^2/k	k	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}	$\hat{\theta}_1^{CU}$	$\hat{\theta}_2^{CU}$	J_{CU}	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}	$\hat{\theta}_1^{CU}$	$\hat{\theta}_2^{CU}$	J_{CU}	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}	$\hat{\theta}_1^{CU}$	$\hat{\theta}_2^{CU}$	J_{CU}
(a) HOMO																			
0	2	0.005	0.005	-	0.005	0.005	-	0.012	0.012	-	0.012	0.012	-	0.017	0.017	-	0.017	0.017	-
0	4	0.006	0.006	0.014	0.010	0.011	0.018	0.006	0.008	0.009	0.011	0.010	0.015	0.005	0.004	0.009	0.013	0.013	0.014
0	8	0.007	0.010	0.033	0.009	0.009	0.041	0.006	0.010	0.016	0.007	0.007	0.022	0.008	0.006	0.007	0.006	0.006	0.014
1	2	0.005	0.007	-	0.005	0.007	-	0.014	0.012	-	0.014	0.012	-	0.013	0.014	-	0.013	0.014	-
1	4	0.005	0.011	0.016	0.004	0.008	0.013	0.007	0.008	0.009	0.008	0.007	0.006	0.006	0.006	0.007	0.007	0.006	0.006
1	8	0.009	0.026	0.029	0.009	0.017	0.028	0.006	0.009	0.016	0.010	0.010	0.013	0.006	0.008	0.013	0.004	0.005	0.010
10	2	0.007	0.026	-	0.007	0.026	-	0.010	0.014	-	0.010	0.014	-	0.012	0.011	-	0.012	0.011	-
10	4	0.008	0.050	0.016	0.013	0.022	0.021	0.009	0.025	0.006	0.011	0.017	0.013	0.008	0.015	0.006	0.010	0.015	0.005
10	8	0.006	0.085	0.026	0.010	0.023	0.018	0.009	0.044	0.012	0.009	0.020	0.014	0.007	0.020	0.010	0.006	0.018	0.013
(b) MA-l (West)																			
0	4	0.008	0.009	0.023	0.016	0.016	0.024	0.007	0.007	0.014	0.012	0.011	0.018	0.005	0.005	0.011	0.009	0.009	0.013
0	8	0.012	0.019	0.053	0.027	0.030	0.057	0.012	0.014	0.028	0.017	0.017	0.031	0.008	0.007	0.011	0.015	0.015	0.016
1	4	0.010	0.015	0.024	0.011	0.015	0.016	0.007	0.010	0.014	0.004	0.005	0.009	0.006	0.005	0.009	0.005	0.005	0.007
1	8	0.018	0.035	0.042	0.037	0.046	0.033	0.007	0.014	0.023	0.014	0.018	0.016	0.006	0.009	0.015	0.007	0.008	0.012
10	4	0.010	0.054	0.026	0.012	0.018	0.009	0.010	0.026	0.011	0.011	0.016	0.007	0.008	0.016	0.009	0.010	0.016	0.006
10	8	0.013	0.094	0.041	0.010	0.021	0.016	0.012	0.051	0.018	0.008	0.021	0.015	0.007	0.023	0.013	0.008	0.020	0.015
(c) HAC (Newey-West)																			
0	4	0.008	0.010	0.034	0.020	0.020	0.039	0.006	0.008	0.021	0.021	0.021	0.029	0.004	0.004	0.015	0.022	0.022	0.018
0	8	0.015	0.022	0.068	0.019	0.019	0.068	0.015	0.018	0.035	0.021	0.021	0.036	0.009	0.011	0.016	0.022	0.021	0.018
1	4	0.011	0.017	0.031	0.013	0.018	0.024	0.009	0.012	0.021	0.005	0.006	0.015	0.006	0.006	0.012	0.005	0.005	0.008
1	8	0.020	0.038	0.106	0.053	0.063	0.107	0.012	0.018	0.050	0.028	0.030	0.052	0.008	0.012	0.021	0.016	0.018	0.024
10	4	0.013	0.057	0.033	0.022	0.029	0.032	0.012	0.029	0.017	0.013	0.020	0.021	0.008	0.016	0.010	0.011	0.016	0.010
10	8	0.018	0.098	0.100	0.062	0.069	0.112	0.016	0.053	0.048	0.029	0.032	0.061	0.009	0.025	0.020	0.011	0.026	0.020

Entries are the Kolmogorov-Smirnov statistics testing the null hypothesis that the two distributions are equal. The critical values are: 0.0096 at 5% and 0.0115 at 1%. Asymptotic distributions where computed using 100,000 replications of the representations in theorem 2.1, see appendix 3.B. Finite-sample distributions were computed using 20,000 Monte Carlo replications.

Table 3.5: Maximum Absolute Difference between Finite-Sample CDF and Asymptotic CDF for Design IV.

		$T = 100$			$T = 200$			$T = 500$		
μ^2/k	k	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}	$\hat{\theta}_1^{2S}$	$\hat{\theta}_2^{2S}$	J_{2S}
(a) MA-l										
0	2	0.013	0.035	-	0.010	0.022	-	0.006	0.010	-
0	4	0.008	0.021	0.030	0.004	0.012	0.028	0.006	0.010	0.030
0	8	0.022	0.115	0.032	0.016	0.064	0.021	0.005	0.019	0.014
1	2	0.012	0.055	-	0.011	0.037	-	0.005	0.020	-
1	4	0.019	0.082	0.032	0.015	0.053	0.026	0.016	0.030	0.025
1	8	0.057	0.227	0.029	0.031	0.142	0.019	0.017	0.065	0.027
10	2	0.037	0.185	-	0.021	0.131	-	0.011	0.072	-
10	4	0.082	0.288	0.022	0.051	0.205	0.019	0.025	0.112	0.022
10	8	0.221	0.493	0.027	0.135	0.373	0.021	0.067	0.228	0.028
(b) HAC (Newey-West)										
0	4	0.028	0.053	0.014	0.031	0.044	0.011	0.032	0.042	0.011
0	8	0.109	0.194	0.080	0.097	0.159	0.072	0.095	0.123	0.065
1	4	0.054	0.117	0.019	0.056	0.094	0.014	0.053	0.072	0.014
1	8	0.144	0.285	0.084	0.122	0.227	0.073	0.112	0.166	0.063
10	4	0.120	0.319	0.013	0.088	0.243	0.015	0.069	0.157	0.015
10	8	0.225	0.480	0.073	0.153	0.379	0.068	0.095	0.253	0.070

Asymptotic distributions where computed using 100,000 replications. Finite-sample distributions were computed using 20,000 replications.

sample size (0.020 for $T = 500$).

Figure 3.1 compares the finite-sample distributions of the 2-step (2SLS) and continuously updated (LIML) GMM estimators to the asymptotic approximations given by theorem 2.1, as well as the normal approximations. It is evident that the finite-sample distributions deviate considerably from normality, and that the local-to-zero approximations capture the shape and location of the distributions very well. Note also that 2SLS exhibits higher bias than LIML, but the latter is much more dispersed, thus demonstrating a trade-off between efficiency and bias in the choice of estimators, see also the next subsection.

The final design investigates the implications of serial correlation for the quality of the asymptotic approximations. The results are given in table 3.5.[7] Several comments are in order.

First, we observe significant departures of the exact distributions from the first-order approximations, in sharp contrast to the results of table 3.4. These deviations seem to increase with the concentration parameter and fall with the sample size. This last observation suggests that the

---

[7] The CUE is not reported because, unlike in the previous designs, its local-to-zero asymptotic approximation is unreliable. This is due to the fact that the cases in which the limiting objective function has no minimum occur much more often in this case, violating one of the conditions of theorem 2.1.

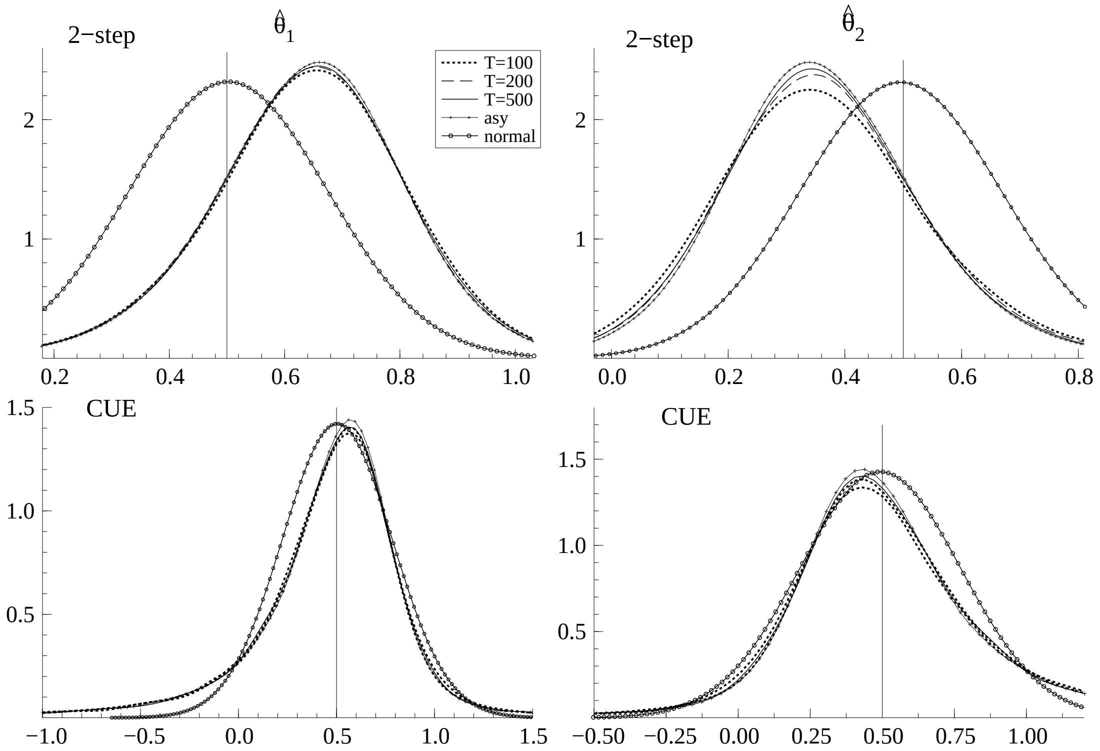


Figure 3.1: Comparison of finite-sample densities of GMM estimators in Design III with local-to-zero and normal asymptotic approximations. Eight instruments have been used ($k = 8$), and the minimum eigenvalue of the concentration parameter matrix is $\mu^2_{min}/k = 1$. The straight line denotes the true value.

source of the discrepancy may lie in higher-order terms that are important even at moderately large samples. A notable exception is when the model is just-identified (in which case all of the estimators are identical). This corresponds to the entries for $k = 2$ in the table, which do not differ substantially from the corresponding results for Design III. This is not surprising, since serial correlation (or heteroscedasticity) does not affect the estimators in the just-identified case.

Second, we see that the approximations are generally somewhat better under West's MA-l estimator, which accounts for the right amount of serial correlation, than under the Newey-West estimator with automatic selection of lag truncation parameter, as we might have expected. A closer inspection reveals also that under the Newey-West estimator the discrepancy between the finite-sample and asymptotic distributions doesn't fall with the sample size, and that the median bias of the GMM estimators tends to be higher than under West's covariance estimator.

An interesting question is also the nature of the observed deviations, i.e., whether the exact distributions are more or less dispersed than the approximations. This cannot be seen from the information in table 3.5, but can be investigated by comparing the distribution functions visually or by looking at summary measures of dispersion. Upon comparing each pair of distribution functions for the reported statistics, we found that in all cases the exact distribution is more dispersed than the asymptotic approximation, see figure 3.2 for some examples. This reinforces the view that the first-order approximations are inadequate and also indicates that the minimum eigenvalue of the concentration parameter may be an incomplete measure of the speed of convergence of finite-sample distributions to the asymptotic approximations in this case.

3.2.1 Bias

Table 3.6 compares the bias of the two types of estimators for the first parameter in Design III, θ_1 and a sample size of 100. The reported statistics are the median bias, also as a proportion of the OLS bias, the median absolute deviation from the true value (MAD), and the interquartile range (IQR). The use of quantile based measures is necessitated by the fact that the CUE doesn't have any moments, and thus moment-based measures such as the MSE are inappropriate.[8]

The following observations stand out. First, we note that in the un-identified case, both estimators exhibit a bias almost identical to the OLS bias, i.e., their distributions are centered on the respective OLS distributions. Moreover, the variability of the 2-step estimator falls with the number of instruments (as shown by both MAD and IQR), but this doesn't happen with the CUE estimator, which seems to be little affected by the addition of more instruments.

Second, in the case of weak identification ($\mu^2/k = 1$), there seems to be a clear trade-off between efficiency and bias, both w.r.t. the number of instruments and w.r.t. the choice of estimator. In the case of the 2-step estimator, the bias seems to rise with the number of instruments, but the dispersion of the distribution falls. On the other hand, this trade-off is absent in the CUE. In particular, the inclusion of more instruments seems to lower the bias (if at all), which is also quite small even under weak identification (contrast 38% to 10% for $k = 4$, or 45% to 4% for $k = 8$). In other words, the CUE exhibits lower bias, even with little information, but it is considerably more variable than the 2-step estimator on the IQR criterion. However, as far as the variability relative to the true value is concerned (measured by MAD), we see that the CUE doesn't do particularly worse in the weakly identified case (compare 0.165 to 0.203 for $k = 8$), thus suggesting that this

---

[8] Such measures as 'approximate MSE criteria' have been reported elsewhere in the literature in comparing those two estimators, see e.g., Donald and Newey (2001). However, we think that those measure are inappropriate since they would tend to exaggerate the variability of the CUE, and would result in an unfair comparison. The MAD is also a natural counterpart to the Median Bias, which is the only meaningful measure of bias for the CUE.

estimator may be preferable to the 2-step in this case.

Third, comparing panel (a) with panels (b) and (c) we see that the impact of the alternative covariance matrix estimators on the bias is minimal. The entries are almost indistinguishable for the 2-step estimator, whereas in the case of the CUE we see a slight increase in precision, at the expense of slightly higher bias. All in all, we may conclude that using HAC estimators when they are not needed doesn't affect the properties of the GMM estimators.[9]

Table 3.6: Analysis of Bias of 2-step and CUE estimator of θ_1 in design III.

		2-step				CUE			
μ^2/k	k	Median bias	% OLS bias	MAD	IQR	Median bias	% OLS bias	MAD	IQR
(a) HOMO									
0	2	0.321	98%	0.641	1.111	0.321	98%	0.641	1.111
0	4	0.326	99%	0.376	0.501	0.306	94%	0.645	1.123
0	8	0.328	100%	0.333	0.304	0.331	101%	0.658	1.141
1	2	0.060	18%	0.318	0.627	0.060	18%	0.318	0.627
1	4	0.125	38%	0.200	0.331	0.033	10%	0.259	0.519
1	8	0.148	45%	0.165	0.214	0.014	4%	0.203	0.409
10	2	-0.001	0%	0.099	0.202	-0.001	0%	0.099	0.202
10	4	0.016	4%	0.069	0.134	-0.002	0%	0.072	0.145
10	8	0.023	7%	0.051	0.094	-0.001	0%	0.051	0.102
(b) MA-l									
0	4	0.324	99%	0.380	0.506	0.324	99%	0.618	1.064
0	8	0.329	100%	0.335	0.314	0.330	101%	0.575	0.946
1	4	0.125	38%	0.203	0.336	0.041	12%	0.263	0.520
1	8	0.149	45%	0.169	0.225	0.035	10%	0.212	0.418
10	4	0.016	4%	0.070	0.137	-0.002	0%	0.071	0.143
10	8	0.024	7%	0.052	0.097	-0.001	0%	0.051	0.102
(c) NW									
0	4	0.325	99%	0.379	0.508	0.312	95%	0.596	0.998
0	8	0.327	100%	0.333	0.320	0.313	95%	0.514	0.834
1	4	0.125	38%	0.205	0.340	0.046	14%	0.263	0.519
1	8	0.149	45%	0.170	0.225	0.046	14%	0.218	0.425
10	4	0.016	4%	0.071	0.139	-0.002	0%	0.072	0.144
10	8	0.024	7%	0.053	0.099	-0.001	0%	0.052	0.103

Monte Carlo standard errors vary between 0.004 and 0.0005 for the estimators that have moments. They cannot be computed for $k = 2$ or the CUE.

Next, we turn to design IV, to investigate the implications of serial correlation for the bias of the estimators. Upon comparison of tables 3.6 and 3.7 several important differences stand out.

---

[9] This finding is consistent with and an extension to the result of Stock and Wright (2000, Section 4.2), where the authors, based on un-reported evidence, state that the choice between heteroscedasticity consistent (HC) or "nonrobust" (HOMO) covariance estimators doesn't affect GMM estimators and tests. In their context of a CCAPM model, there is no serial correlation by construction.

First, we notice that in the purely un-identified case, neither of the two estimators exhibit the full OLS bias, which was evident in the static case of Design III. Apart from that, the bias is similar for both estimators, and the CUE is more dispersed as in the previous case.

More importantly, when we look at the weakly identified case, we see that the CUE exhibits a considerable bias, unlike the static case, and also the bias reduction relative to the 2-step estimator is much smaller. This fact, combined with the higher variability as measured by both the MAD and the IQR suggest that the CUE is not necessarily a preferable choice of estimator in the dynamic case. In other words, the trade-off between efficiency and bias is not in favour of the CUE. This seems to contradict the commonly held view that the latter (LIML) is preferable under weak instruments, as suggested for instance by Staiger and Stock (1997).

However, under strong identification, the above conclusions seem to be reversed. The CUE seems to dominate (mildly) the 2-step estimator on both grounds of efficiency and bias. Even though the difference is small, the last row of panel (a) in table 3.7 shows that the bias of the 2-step estimator is very sensitive to the number of instruments (from 9% to 20% of the OLS bias when the number of instruments grows from 4 to 8), whereas the CUE is relatively immune to expansions of the instrument set. This is potentially important in situations where the number of admissible instruments (most of which are irrelevant) is large, as in the case of dynamic forward-looking models. In particular, when identification is strong, the CUE may be a better choice since it is robust to 'over-instrumenting'.

Finally, by comparing panels (a) and (b) in table 3.7, we see that the 2-step estimator based on West's MA-l HAC dominates the one based on Newey-West, exhibiting lower bias but similar variability. This doesn't seem to be the case for the CUE, especially when identification is weak, thus showing that there is no clear case in favour of West's estimator on a bias-efficiency criterion.

All in all, the main finding of this section is that serial correlation in the structural errors increases the bias of the CUE, and weakens the case for preferring it when identification is weak.

3.3 Finite-sample coverage rates of robust confidence sets

In this section we analyze the implications of using non-strongly exogenous instruments and serial correlation corrections for the finite-sample coverage rates of confidence sets that are robust to weak identification, such as the S-set and K-set discussed in section 2.2.3. Here, we focus on the S-set because it is conceptually simpler, but a similar analysis would apply to the K-set.

Table 3.7: Analysis of Bias of 2-step and CUE estimator of θ_1 in design IV.

		2-step				CUE			
μ^2/k	k	Median bias	% OLS bias	MAD	IQR	Median bias	% OLS bias	MAD	IQR
(a) MA-l									
0	2	-0.225	52%	0.440	0.758	-0.225	52%	0.440	0.758
0	4	-0.307	71%	0.358	0.499	-0.318	74%	0.481	0.799
0	8	-0.299	69%	0.300	0.268	-0.317	73%	0.378	0.486
1	2	-0.148	34%	0.359	0.651	-0.148	34%	0.359	0.651
1	4	-0.204	47%	0.267	0.421	-0.171	39%	0.349	0.636
1	8	-0.219	51%	0.224	0.213	-0.187	43%	0.277	0.428
10	2	-0.004	0%	0.160	0.338	-0.004	0%	0.160	0.338
10	4	-0.042	9%	0.116	0.213	0.004	0%	0.125	0.264
10	8	-0.090	20%	0.107	0.145	-0.001	0%	0.106	0.219
(b) NW									
0	4	-0.321	74%	0.363	0.460	-0.262	60%	0.476	0.854
0	8	-0.344	80%	0.345	0.278	-0.229	53%	0.326	0.606
1	4	-0.227	52%	0.275	0.408	-0.148	34%	0.335	0.617
1	8	-0.255	59%	0.258	0.237	-0.135	31%	0.226	0.368
10	4	-0.056	13%	0.121	0.215	-0.008	1%	0.129	0.263
10	8	-0.090	21%	0.108	0.149	-0.023	5%	0.106	0.204

Monte Carlo standard errors vary between 0.0049 and 0.0009 for the estimators that have moments. They cannot be computed for $k = 2$ or the CUE.

The S-set is based on inverting the (normalized) objective function $\mathcal{Q}_T(\theta, \theta)$, based on the fact that it is asymptotically distributed as $\chi^2(k)$, see theorem 2.3. The coverage rate of a confidence set with nominal $(1-\alpha)\%$ coverage is the probability that this (random) confidence set includes the true parameter value θ_0 under the null. This is equivalent to one minus the null rejection probability of the (Anderson-Rubin) statistic $\mathcal{Q}_T(\theta_0, \theta_0)$ testing the joint hypothesis that $\theta = \theta_0$. Thus, the coverage rates are given by $\Pr\left(\mathcal{Q}_T(\theta_0, \theta_0) < \chi^2_{1-\alpha}(k)\right)$, where $\chi^2_{1-\alpha}(k)$ denotes the usual $\alpha\%$ critical value of a χ^2 test with k degrees of freedom.

The questions of interest here are the following:

1. How are finite-sample coverage rates affected by the use of non-strongly exogenous instruments?
2. What is the impact of HAC covariance estimators?
3. What is the impact of serial correlation?

These will be investigated using the specification of Design IV, under partial identification.[10] The

---

[10] Extending these experiments to the case of weak identification is straightforward, but the results are not essentially different, and are omitted.

model is:

$$y_t = \theta_1 \, y_{t+1} + \theta_2 \, y_{t-1} + u_t \tag{3.6}$$

and the DGP is

$$y_t = \alpha \, y_{t-1} + e_t \tag{3.7}$$

Two sets of instruments will be used: Strongly exogenous $Z_t = (\mathbf{z}_1, \ldots, \mathbf{z}_k)'$ where the instruments are drawn from a k-dimensional VAR(1) ($C \neq 0$ in 3.1); or Lagged endogenous variables $Z_t = y_t, \ldots, y_{t-k}$, (details are given in appendix 3.C).

This simple design is capable of generating various types of serial correlation, according to the true value of θ_1. Since θ_1 and θ_2 are completely un-identified here, we can choose any pair satisfying the condition: $\theta_2 = \alpha/(1 - \alpha \, \theta_1)$, for any given α. The latter governs the dynamics of the process y_t, whereas $\theta_1 \neq 0$ will cause the structural error u_t to be serially-correlated.[11]

Impact of lagged instruments

To address the first question, we set $\theta_1 = 0$ (which would correspond to the hypothesis of no forward-dynamics) and hence u_t is serially uncorrelated. Table 3.8 reports the coverage rates of nominal 90% S-sets based on a homoscedastic covariance estimator (HOMO) and an unfeasible version ($HOMO_0$) which removes the impact of estimating the structural error variance, thus being exactly pivotal in the case of strongly exogenous instruments and under normality.

Table 3.8: Impact of non-strongly exogenous instruments on coverage rates of nominal 90% S-sets.

		Strongly exog. instr.		*Lagged instr.* $\alpha = 0$		*Lagged instr.* $\alpha = 0.5$		*Lagged instr.* $\alpha = 0.9$	
k	T	HOMO	HOMO$_0$[a]	HOMO	HOMO$_0$	HOMO	HOMO$_0$	HOMO	HOMO$_0$
4	50	90.4%	90.0%	92.9%	91.9%	92.6%	91.9%	92.4%	91.6%
4	100	90.2%	90.0%	91.2%	90.7%	91.0%	90.7%	90.9%	90.5%
4	200	90.1%	90.0%	91.1%	91.0%	91.1%	90.9%	91.0%	90.7%
4	500	90.0%	90.0%	90.5%	90.4%	90.2%	90.1%	90.4%	90.2%
8	50	91.5%	90.0%	95.1%	93.5%	95.2%	93.6%	95.2%	93.6%
8	100	90.7%	90.0%	92.8%	92.2%	92.9%	92.3%	92.9%	92.3%
8	200	90.5%	90.0%	91.7%	91.2%	91.5%	91.1%	91.5%	91.1%
8	500	89.9%	90.0%	90.3%	90.2%	90.5%	90.1%	90.5%	90.1%

All statistics are based on 20000 replications, with HOMO$_0$ used as a control variate. The Monte Carlo standard errors vary from 0.0015 to 0.0021.

[a] HOMO$_0$ denotes the unfeasible $\tilde{V} = \sigma_u^2 \, \widehat{\Sigma}_{ZZ}$, whereas HOMO denotes the usual non-robust estimator $\widehat{V} = \widehat{\sigma}_u^2 \, \widehat{\Sigma}_{ZZ}$.

The following observations are noteworthy. First, all of the reported results in table 3.8 show

---

[11] $\theta_1 > 0$ makes u_t *negatively* serially-correlated, and vice versa.

that the confidence sets have coverage rates that are higher than their nominal size when lagged instruments are used, irrespective of whether the (lagged) instruments are serially-correlated ($\alpha \neq 0$) or not.[12] Second, the size distortions are larger at the smallest sample size (T=50) and disappear as the sample size grows, confirming the asymptotic theory. Third, at small samples these distortions are not the result of the imprecise estimation of σ_u^2, since they are evident even for the unfeasible versions (HOMO$_0$). Fourth, the distortions increase with the number of lagged instruments: at $T = 50$, the test under-rejects about 29% of the time with $k = 4$, but this increases to 51% when four more lagged instruments are added (the sample size is kept fixed).

All in all, lagged instruments cause some size distortion which is only relevant in small samples, but tends to increase with the number of instruments.

Some intuition to understand this observed lagged instruments effect is needed. To simplify the analysis, we assume away the uncertainty arising from the estimation of the structural error variance (i.e., work under the assumption that σ_u^2 is known).

In the benchmark case of strongly exogenous instruments, under normality, the objective function is distributed exactly as $\chi^2(k)$. This follows from the fact that, conditional on the complete history of the instruments (which is possible when they are strongly exogenous), $\sum_{t=1}^{T} Z_t\, u_t = Z'u \sim N_k(0, \sigma_u^2\, Z'Z)$. By standardizing the instruments $\tilde{Z}_t = (Z'Z)^{-1/2} Z_t$, we can therefore re-write the moment conditions $\sum_{t=1}^{T} \tilde{Z}_t\, u_t = (Z'Z)^{-1/2}\, Z'\, u \sim N(0, \sigma_u^2\, I_k)$ such that

$$AR_0 \equiv \mathcal{Q}_T(\theta_0, \theta_0) = \frac{u'Z(Z'Z)^{-1}Z'u}{\sigma_u^2} \sim \chi^2(k) \tag{3.8}$$

On the other hand, when the instruments are not strongly exogenous, meaning that Z_{t+i} correlate with u_t for $i > 0$, this standardization violates the conditional normality of the standardized moment functions $\tilde{Z}_t\, u_t$. Under joint normality of u_t and Z_s, Z_s can be orthogonally decomposed into $\check{Z}_s + A_{s,t}\, u_t$, for some nonrandom matrix $A_{s,t}$. Hence, $Z'Z$ can be written as

$$Z'Z = \sum_{s=1}^{T} Z_s Z_s' = \sum (\check{Z}_s + A_{s,t}\, u_t)(\check{Z}_s + A_{s,t}\, u_t)' = \check{Z}'\check{Z}(I + B_{1t}\, u_t + B_{2t}\, u_t^2)$$

where B_{it} are $O_p(T^{-1})$.[13] When Z_t is a scalar, we can re-write the moment function $\tilde{Z}_t\, u_t$ as:

$$\tilde{Z}_t\, u_t = (Z'Z)^{-1/2} Z_t\, u_t = \frac{(\check{Z}'\check{Z})^{-1/2} Z_t\, u_t}{(1 + B_{1t}\, u_t + B_{2t}\, u_t^2)^{1/2}} \tag{3.9}$$

Conditioning on $(\check{Z}'\check{Z})^{-1/2} Z_t$, we note that the distribution of the above quantity will deviate from normality by $O_p(T^{-1})$ terms. The nature of this deviation will depend on the polynomial

---

[12] Note that the strongly exogenous instruments are serially-correlated, but this clearly does not affect the size of the test.

[13] In particular, $B_{1t} = (\check{Z}'\check{Z})^{-1} \sum_s (A_{s,t}\, Z_s' + Z_s\, A_{s,t}')$ and $B_{2t} = (\check{Z}'\check{Z})^{-1} \sum_s A_{s,t}\, A_{s,t}'$.

in the denominator. We can argue heuristically that, since this polynomial is always positive and $B_{2t} > 0$, it will be increasing at large values of u_t, so that the conditional distribution of (3.9) will have thinner tails than the normal in small samples. This seems consistent with the fact that the test statistic of interest $\left(\sum_{t=1}^{T} \tilde{Z}_t\, u_t\right)^2$ under-rejects in small samples.

However, there are important limitations of this approach. First, it is not straightforward to argue that the small-sample effect on the conditional distribution of $(\tilde{Z}_t\, u_t)$ in (3.9) carries forward to the unconditional one, and therefore to the distribution of $\sum_{t=1}^{T} \tilde{Z}_t\, u_t$. Second, it is not easy to see how this small-sample effect varies with the number of instruments, in order to explain the finding that the distortion rises with k.[14]

An alternative approach would be to perform an asymptotic expansion to the distribution of the Anderson-Rubin statistic, with a view to characterizing the finite-sample departure from a $\chi^2(k)$. Focusing on the scalar case of $Z_t = y_{t-1}$, we can equivalently perform an expansion of the statistic $\left(\sum_t y_{t-1}^2\right)^{-1/2} \sum_t y_{t-1}\, u_t$, to order T^{-1} to investigate its tail behaviour compared to $\left(\sum_t z_t^2\right)^{-1/2} \sum_t z_t\, u_t$, say, where $z_s \perp u_t$, for all t, s. This can be done along the lines of Phillips (1977) or Rothenberg (1984), but it is beyond the scope of the present study, and will be omitted.

Impact of serial correlation

Next, we turn to the impact of serial correlation in the structural residuals, which necessitates the use of a HAC covariance estimator in the objective function. Table 3.9 presents the finite-sample coverage rates of the S-sets using three types of serial correlation: positive, zero and negative. These results correspond to y_t being positively serially-correlated ($\alpha = 0.5$ in the DGP (3.7)), but they are not sensitive to the value of α, compare with table 3.10.[15]

The following observations can be made. First, the use of the Newey-West (NW) covariance estimator estimator results in serious distortion in the confidence sets, and the use of pre-whitening and recoloring doesn't seem to improve it (see columns NW and NW-PW in the table). This is true in all cases, even when the serial correlation corrections are not needed, but more so in the case of negative serial correlation, where the size distortion remains considerable (of the order of 50%) even at a sample size of 500. Compared with the coverage rates using MA-l we see that the NW-based sets do much worse. The message is that the use of NW covariance estimators will cause the robust confidence sets to be too conservative, and that inference based on them will

---

[14] When $k > 1$, the polynomial in the denominator is a matrix, making it harder to characterize the tails of the conditional distribution of (3.9).

[15] Coverage rates also for the cases $\alpha = -0.5$ and $\alpha = 0.9$ where not significantly different from the ones presented.

suffer from large type II error.

Second, we look at the coverage rates associated with West's covariance estimator (MA-l). This performs far better than NW, but still involves considerable distortion at small samples under negative serial correlation. This distortion is not due to the fact that the MA coefficient in the MA-l estimator is imprecisely estimated, since setting it equal to the true value (column MA-l_0) doesn't seem to improve coverage. In the worst case, for a sample size of 50 using 8 instruments, the Anderson Rubin statistic will reject 2.8% of the time at the 10% level.

In general, we notice that the coverage rates are increasing with the number of instruments and higher under negative serial correlation. This is particularly relevant for partial adjustment models, where the forward-looking parameter is mostly positive, giving rise to negative serial correlation.

All in all, this Monte Carlo study casts doubts on the validity of the S-sets for partial adjustment models at relatively small samples, particularly when errors are serially-correlated, and recommends against the use of NW covariance estimators.

3.4 The size of the Hansen-Sargan test

In this section, we study the finite-sample size of the Hansen-Sargan test under weak identification and investigate the implications of serial correlation. The theoretical results of section 2.5 show that this statistic is non-pivotal under weak identification, but a boundedly pivotal version can be derived using the CUE, theorem 2.6. We wish to investigate how practical this suggestion is, especially when there is serial correlation which has to be accounted for using a HAC estimator.

The Monte Carlo results are available for all the designs, but we only report the results for designs III and IV for brevity. Table 3.11 gives the rejection frequencies of the test at nominal 10% and 5% levels, for the two versions of the test J_{2S} and J_{CU}, using alternative HAC estimators. Since no heteroscedasticity and serial correlation is present, the homoscedastic versions (2SLS and LIML in panel (a)) are sufficient. In panel (a), we observe that both versions of the test under-reject in the un-identified case. More importantly, the LIML-based J_{CU} test rejects significantly less than the 2SLS one J_{2S}, but the under-rejection decreases as the sample and concentration parameters rise. The J_{2S} test appears closer to its nominal size in the weakly identified case, but over-rejects as the number of instruments grows. Moreover, this over-rejection doesn't fall with the sample when the concentration parameter is fixed to 1. In all cases, the rejection frequencies rise

Table 3.9: Coverage Rates of nominal 90% S-sets using different HAC estimators, $\alpha = 0.5$ in (3.7).

		k= 4					k= 8				
SC[a]	T	NW	NW–PW	MA-l	MA-l_0[b]	HOMO$_0$[c]	NW	with PW	MA-l	MA-l_0	HOMO$_0$
Strongly exog. instr.											
+ve	50	98.5%	98.1%	92.4%	92.7%	90.0%	100.0%	98.4%	94.9%	94.6%	90.0%
+ve	100	94.9%	96.2%	91.3%	91.2%	90.0%	99.7%	99.5%	92.6%	92.5%	90.0%
+ve	200	91.7%	94.0%	90.6%	90.6%	90.0%	95.1%	96.9%	91.0%	91.1%	90.0%
+ve	500	90.0%	92.4%	90.2%	90.2%	90.0%	91.4%	94.2%	90.3%	90.4%	90.0%
Lagged instr.											
+ve	50	99.7%	97.3%	93.7%	93.4%	91.6%	100.0%	96.5%	96.6%	96.8%	93.6%
+ve	100	95.9%	94.6%	91.8%	91.6%	90.6%	100.0%	99.3%	93.8%	93.9%	92.2%
+ve	200	93.1%	92.3%	91.0%	90.8%	90.6%	96.6%	95.3%	91.6%	91.9%	91.0%
+ve	500	91.5%	91.0%	90.6%	90.5%	90.3%	93.6%	92.5%	90.4%	90.5%	90.3%
Strongly exog. instr.											
0	50	98.7%	97.5%	92.1%	91.7%	90.0%	100.0%	98.5%	94.6%	93.7%	90.0%
0	100	95.8%	95.5%	91.1%	90.8%	90.0%	99.8%	99.2%	92.1%	92.0%	90.0%
0	200	92.8%	92.7%	90.4%	90.3%	90.0%	96.3%	95.9%	91.1%	90.8%	90.0%
0	500	91.0%	91.0%	90.1%	90.1%	90.0%	93.3%	93.2%	90.4%	90.3%	90.0%
Lagged instr.											
0	50	99.8%	97.4%	93.3%	92.6%	91.9%	100.0%	97.0%	96.4%	95.8%	93.6%
0	100	96.2%	94.8%	91.2%	91.1%	90.7%	99.9%	99.3%	93.6%	93.5%	92.3%
0	200	92.8%	92.2%	91.1%	91.2%	90.9%	96.1%	95.2%	91.7%	91.6%	91.1%
0	500	90.9%	90.7%	90.1%	90.3%	90.1%	92.4%	92.1%	90.5%	90.4%	90.1%
Strongly exog. instr.											
-ve	50	99.7%	98.1%	93.8%	92.9%	90.0%	100.0%	98.5%	95.9%	95.1%	90.0%
-ve	100	98.6%	97.1%	92.5%	91.6%	90.0%	100.0%	99.7%	93.6%	93.1%	90.0%
-ve	200	96.9%	95.0%	90.9%	90.6%	90.0%	99.2%	98.0%	91.9%	91.5%	90.0%
-ve	500	95.7%	93.6%	90.3%	90.2%	90.0%	97.8%	95.8%	90.9%	90.7%	90.0%
Lagged instr.											
-ve	50	99.9%	97.6%	95.2%	94.4%	91.9%	100.0%	95.8%	97.6%	97.2%	93.3%
-ve	100	98.0%	96.2%	93.2%	92.4%	91.1%	100.0%	99.4%	94.6%	94.0%	91.9%
-ve	200	95.9%	93.9%	91.7%	91.4%	90.8%	98.2%	96.5%	92.2%	91.7%	90.4%
-ve	500	94.9%	92.9%	90.9%	90.7%	90.2%	96.1%	94.2%	90.9%	90.7%	90.4%

Monte Carlo standard errors (MCSE) vary between 0.12% and 0.21%.

[a] The three specifications of serial correlation are: $\theta_1 = -0.6$ (*positive*), $\theta_1 = 0.6$ (*negative*) or $\theta_1 = 0$ (zero) in equation (3.6).

[b] This is West's MA-l estimator with the residual MA coefficient estimate replaced by its true value.

[c] As in table 3.8.

Table 3.10: Coverage Rates of nominal 90% S-sets using different HAC estimators, $\alpha = 0$ in (3.7).

		k= 4					k= 8				
SC[a]	T	NW	NW–PW	MA-l	MA-l_0[b]	HOMO$_0$[c]	NW	with PW	MA-l	MA-l_0	HOMO$_0$
Strongly exog. instr.											
+ve	50	98.4%	98.1%	92.8%	92.7%	90.0%	100.0%	98.3%	95.1%	94.6%	90.0%
+ve	100	94.8%	96.5%	91.3%	91.4%	90.0%	99.7%	99.6%	92.6%	92.5%	90.0%
+ve	200	91.3%	94.2%	90.6%	90.6%	90.0%	95.1%	97.1%	91.2%	91.2%	90.0%
+ve	500	89.8%	92.7%	90.1%	90.3%	90.0%	91.2%	94.5%	90.4%	90.5%	90.0%
Lagged instr.											
+ve	50	99.8%	97.3%	94.2%	93.6%	91.6%	100.0%	96.3%	96.7%	96.6%	93.5%
+ve	100	96.9%	95.1%	92.0%	91.7%	90.9%	100.0%	99.2%	93.7%	93.7%	92.2%
+ve	200	94.2%	92.5%	91.1%	90.8%	90.7%	97.4%	95.6%	91.8%	92.0%	91.1%
+ve	500	92.6%	91.3%	90.7%	90.7%	90.3%	94.7%	92.9%	90.4%	90.4%	90.3%
Strongly exog. instr.											
0	50	98.7%	97.5%	92.1%	91.7%	90.0%	100.0%	98.5%	94.6%	93.7%	90.0%
0	100	95.8%	95.5%	91.1%	90.8%	90.0%	99.8%	99.2%	92.1%	92.0%	90.0%
0	200	92.8%	92.7%	90.4%	90.3%	90.0%	96.3%	95.9%	91.1%	90.8%	90.0%
0	500	91.0%	91.0%	90.1%	90.1%	90.0%	93.3%	93.2%	90.4%	90.3%	90.0%
Lagged instr.											
0	50	99.8%	97.6%	93.6%	92.8%	91.9%	100.0%	97.1%	96.5%	95.9%	93.5%
0	100	96.6%	95.2%	91.4%	91.3%	90.7%	99.9%	99.3%	93.6%	93.3%	92.2%
0	200	92.7%	92.1%	91.1%	91.1%	91.0%	96.1%	95.2%	92.0%	91.7%	91.2%
0	500	91.0%	90.7%	90.3%	90.3%	90.4%	92.4%	92.0%	90.3%	90.3%	90.2%
Strongly exog. instr.											
-ve	50	99.6%	97.9%	93.2%	92.5%	90.0%	100.0%	98.5%	95.4%	94.7%	90.0%
-ve	100	98.1%	96.5%	91.4%	91.1%	90.0%	100.0%	99.6%	93.0%	92.8%	90.0%
-ve	200	96.3%	94.3%	90.3%	90.3%	90.0%	98.9%	97.4%	91.4%	91.4%	90.0%
-ve	500	95.0%	93.0%	90.4%	90.4%	90.0%	97.1%	95.0%	90.6%	90.6%	90.0%
Lagged instr.											
-ve	50	99.8%	97.2%	94.6%	93.5%	91.9%	100.0%	96.0%	97.4%	96.8%	93.3%
-ve	100	97.0%	95.2%	92.6%	91.8%	91.1%	100.0%	99.2%	94.0%	93.7%	92.0%
-ve	200	94.0%	92.3%	91.4%	91.2%	90.7%	97.5%	95.7%	91.6%	91.5%	90.5%
-ve	500	92.9%	91.5%	90.6%	90.5%	90.4%	94.7%	92.8%	90.9%	90.8%	90.5%

MCSE vary between 0.10% and 0.21%.

[a] The three specifications of serial correlation are: $\theta_1 = -0.6$ (*positive*), $\theta_1 = 0.6$ (*negative*) or $\theta_1 = 0$ (zero) in equation (3.6).

[b] This is West's MA-l estimator with the residual MA coefficient estimate replaced by its true value.

[c] As in table 3.8.

with the number of over-identifying restrictions $k-2$, but they remain mostly below the nominal level of significance.[16]

Finally, comparing panel (a) with (b) and (c) we see that using the MA-l covariance estimator when it is not needed doesn't make much difference, whereas the NW estimator causes more under-rejection in small samples, even under strong identification, and is very sensitive to the number of instruments used (last row of the table).

Table 3.11: Size of the Hansen-Sargan test of over-identifying restrictions in Design III.

		$T=100$				$T=200$				$T=500$			
μ^2/k	k	J_{2S}		J_{CU}		J_{2S}		J_{CU}		J_{2S}		J_{CU}	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
(a) HOMO													
0	4	6.1%	2.9%	1.1%	0.2%	5.8%	2.6%	1.0%	0.3%	5.5%	2.3%	1.0%	0.2%
0	8	6.8%	3.0%	0.7%	0.2%	6.5%	3.0%	0.8%	0.2%	6.4%	2.9%	0.8%	0.2%
1	4	10.3%	5.3%	5.4%	2.0%	9.9%	5.3%	5.1%	2.1%	10.0%	5.0%	4.9%	1.8%
1	8	12.0%	6.3%	5.5%	2.2%	12.1%	6.4%	5.6%	2.3%	12.1%	6.4%	5.7%	2.2%
10	4	10.8%	5.3%	8.8%	4.2%	10.5%	5.3%	9.3%	4.4%	10.5%	5.3%	9.8%	4.6%
10	8	11.3%	5.6%	8.9%	4.3%	11.0%	5.6%	9.4%	4.5%	11.0%	5.6%	9.9%	4.8%
(b) MA-l													
0	4	6.0%	2.7%	1.1%	0.3%	5.7%	2.5%	1.0%	0.3%	5.7%	2.3%	1.0%	0.2%
0	8	6.5%	2.6%	0.7%	0.1%	6.2%	2.7%	0.7%	0.2%	6.5%	3.0%	0.8%	0.2%
1	4	10.2%	4.9%	4.8%	1.6%	9.9%	5.2%	4.8%	1.9%	9.7%	4.9%	4.6%	1.8%
1	8	10.8%	4.8%	3.9%	1.2%	11.5%	5.7%	4.9%	1.7%	11.8%	6.2%	5.3%	2.0%
10	4	10.4%	4.9%	8.7%	4.1%	10.6%	5.1%	9.2%	4.3%	10.4%	5.1%	9.7%	4.6%
10	8	10.0%	4.4%	8.6%	4.1%	10.5%	5.1%	9.1%	4.3%	10.7%	5.3%	9.6%	4.5%
(c) NW													
0	4	5.0%	1.7%	0.7%	0.1%	5.2%	2.1%	0.8%	0.2%	5.4%	2.1%	0.9%	0.2%
0	8	1.2%	0.1%	0.1%	0.0%	3.7%	1.1%	0.3%	0.0%	5.2%	2.0%	0.6%	0.1%
1	4	8.2%	3.2%	3.2%	0.8%	9.1%	4.3%	4.1%	1.4%	9.5%	4.5%	4.4%	1.5%
1	8	2.4%	0.3%	0.3%	0.0%	7.1%	2.5%	2.1%	0.5%	9.8%	4.3%	3.9%	1.2%
10	4	8.5%	3.2%	8.2%	3.8%	9.4%	4.3%	8.7%	4.0%	9.9%	4.7%	9.2%	4.2%
10	8	2.2%	0.3%	7.0%	2.8%	6.5%	2.2%	7.4%	3.0%	8.9%	3.9%	7.8%	3.1%

Asymptotic MC standard errors: 0.21% for the 10% level, and 0.15% for the 5% level.

Turning now to Design IV in table 3.12, we see a similar story. The rejection frequencies appear to be slightly closer to the nominal levels, but still the tests under-reject in most cases. The distortion in the null rejection probabilities of J_{CU} appears to be smaller but still too large, and the MA-l-based tests outperform the NW-based tests.

All in all, we see that weak identification causes the J-statistic to lead to conservative inference, especially when a general correction for serial correlation is used (NW). This is consistent with the discussion in Staiger and Stock (1997) as far as the LIML estimator is concerned, but doesn't agree with their finding that J_{2SLS} rejects too often. The opposite seems to be true here, and, if anything that statistic appears to be closer to the nominal size under weak identification. However, there is no reason to expect it to be so in general, as theorem 2.6 suggests. There are cases in

---

[16] This observed under-rejection on the null hypothesis suggests that the main problem will be low power under alternative hypotheses, consistent with the numerical results reported in the next chapter.

Table 3.12: Size of the Hansen-Sargan test of over-identifying restrictions in Design IV.

		$T = 100$				$T = 200$				$T = 500$			
μ^2/k	k	J_{2S}		J_{CU}		J_{2S}		J_{CU}		J_{2S}		J_{CU}	
		10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%
(a) MA-l													
0	4	7.3%	3.3%	2.7%	1.1%	7.5%	3.5%	3.3%	1.3%	8.2%	3.8%	3.7%	1.5%
0	8	7.3%	3.0%	3.0%	0.9%	8.4%	3.9%	4.7%	1.9%	9.5%	4.6%	5.6%	2.5%
1	4	8.9%	4.1%	3.6%	1.4%	9.3%	4.5%	4.2%	1.7%	9.7%	4.9%	4.3%	1.8%
1	8	9.4%	4.2%	3.9%	1.4%	11.1%	5.4%	5.7%	2.4%	12.4%	6.3%	6.7%	2.8%
10	4	11.2%	5.9%	6.7%	2.8%	11.5%	6.2%	7.7%	3.4%	12.2%	6.7%	8.1%	3.8%
10	8	10.9%	5.3%	5.8%	2.3%	12.4%	6.6%	8.1%	3.6%	13.8%	7.5%	8.5%	4.0%
(b) NW													
0	4	5.9%	2.6%	2.2%	0.7%	6.1%	2.7%	2.7%	0.9%	6.5%	3.0%	2.7%	1.1%
0	8	4.9%	1.8%	1.6%	0.3%	6.0%	2.6%	2.4%	0.9%	7.2%	3.3%	3.2%	1.3%
1	4	7.0%	3.1%	3.2%	1.0%	7.6%	3.4%	3.8%	1.3%	7.7%	3.9%	3.6%	1.4%
1	8	6.0%	2.4%	2.3%	0.6%	7.4%	3.4%	3.4%	1.2%	8.7%	4.3%	4.2%	1.7%
10	4	9.5%	4.5%	6.6%	2.5%	10.1%	5.1%	7.7%	3.5%	10.1%	5.5%	7.7%	3.6%
10	8	6.5%	2.7%	4.1%	1.4%	8.0%	3.8%	5.9%	2.5%	9.3%	4.8%	6.8%	3.0%

Asymptotic MC standard errors: 0.21% for the 10% level, and 0.15% for the 5% level.

which that test may indeed over-reject, as suggested for example in the experiments undertaken by Staiger and Stock (1997), so the two findings are not contradictory. What they show is that it is impossible to reach general conclusion about the extent and direction of size distortions, since they depend on unknown nuisance parameters.

Finally, we turn to the suggestion made in theorem 2.6 for a boundedly pivotal test based on the J_{CU} statistic. This amounts to comparing J_{CU} not to the usual $\chi^2(k-2)$ critical values (where two parameters are estimated in those two designs) but to a $\chi^2(k-1)$, whose degrees of freedom is the number of over-identification relative to the *well-identified* parameters. We illustrate the conclusion of the theorem by comparing the distribution of the J_{CU} statistic to a $\chi^2(k-1)$, for all of the cases reported in table 3.12 above, see figure 3.3. Evidently, the quantiles of the statistic never exceed those of the bounding distribution, but we notice that the difference is quite significant in all cases. In other words, this the test appears to be too conservative, and even more so than the 'unbounded' test, which also under-rejected in all the cases presented in tables 3.11 and 3.12.

These results raise doubts about the ability of this test to diagnose mis-specification in weakly identified situations. A rejection will clearly signify mis-specification (and with a very low type I error), but it will be unlikely. Nevertheless, this is the only valid test of the over-identifying restrictions under weak identification, to date, so clearly more work is needed in this area.

3.5 Size and power of the excess serial-correlation test

We present some Monte Carlo evidence on the size and power of the excess serial-correlation test developed in section 2.5.3. The experiments are based on a modification of design I to allow for serial correlation in the structural residuals. In particular, the model and DGP are:

$$\begin{aligned} y_t &= \theta\, Y_t + u_t \\ u_t &= \varphi\, u_{t-1} + \varepsilon_t + \psi\, \varepsilon_{t-1} \qquad (3.10) \\ Y_t &= \Pi'\, Z_t + v_t \end{aligned}$$

We set Π (and hence the concentration parameter) high to ensure identification, since the test is valid only in that case. This design is well-suited to address both the size and the power of the test. The null hypothesis is that the structural errors exhibit up to first-order autocorrelation ($\varphi = 0$) under the alternative of higher-order serial correlation. The instruments are strongly exogenous $Z_t = \mathbf{z}_t$ to preserve their validity under the alternative hypothesis. When lags of y_t are used as instruments, the test may also have power against violation of the orthogonality conditions under the alternative.[17] We abstract from this case because we want to study the pure power function of the test under its alternative.

Various versions of the tests are presented, in order to study its size at different degrees of freedom. As in standard serial-correlation tests, a choice has to be made about the number of lags to be used in constructing the test under the alternative. The null and alternative hypotheses are given by (2.37), and clearly require the specification of the number of lags to be tested for under the alternative s. Since s is the degrees of freedom of the distribution of the test, there is clearly a trade-off between allowing for more lags to maximize the chance of detecting higher-order serial correlation, and the risk of over-parameterization which would affect power. Under the alternative, the test has a non-central χ^2 distribution. When higher-order autocorrelations are non-zero, including them in the statistic will possibly raise the non-centrality parameter, but also the degrees of freedom, thus having an ambiguous effect on the power.

Table 3.13 presents Monte Carlo results on the finite-sample size of SC tests of the hypothesis of first order serial correlation against excess serial correlation of orders 2 to 6.

We notice that the finite-sample size of the test is reasonably good. The most parsimonious version SC(1,1) is almost perfectly sized, but increasing s causes under-rejection in small samples.

---

[17] When $\varphi = 0$, lags of y_{t-2} onwards are valid instruments, but when $\varphi \neq 0$ the error exhibits infinite order autocorrelation, and no lags of y_t would be valid instruments.

Table 3.13: Size of various excess serial correlation tests at 10% level of significance.

k	ψ	T	SC(1,1)	SC(1,2)	SC(1,3)	SC(1,4)	SC(1,5)
4	-0.6	100	10.0%	9.2%	8.5%	8.0%	7.7%
4	-0.6	200	9.9%	9.7%	9.3%	8.9%	9.0%
4	-0.6	500	10.3%	10.3%	9.8%	10.1%	10.1%
4	0	100	9.6%	9.5%	8.8%	8.3%	8.0%
4	0	200	10.2%	9.6%	9.8%	9.4%	9.2%
4	0	500	10.5%	9.9%	9.7%	9.8%	9.9%
4	0.6	100	10.3%	9.5%	9.2%	9.0%	8.1%
4	0.6	200	10.2%	10.1%	9.9%	9.7%	9.4%
4	0.6	500	9.7%	9.8%	9.6%	9.9%	9.7%
8	-0.6	100	9.8%	9.2%	8.3%	7.9%	7.1%
8	-0.6	200	9.9%	9.5%	9.3%	8.8%	9.1%
8	-0.6	500	10.4%	10.4%	10.0%	10.1%	10.0%
8	0	100	9.6%	9.2%	8.5%	7.9%	7.8%
8	0	200	10.4%	9.7%	9.8%	9.5%	9.3%
8	0	500	10.4%	10.0%	9.6%	9.8%	9.9%
8	0.6	100	10.3%	9.4%	9.0%	8.6%	7.9%
8	0.6	200	10.0%	10.2%	9.8%	9.9%	9.3%
8	0.6	500	9.8%	9.7%	9.6%	9.8%	9.8%

Asymptotic MCSE: 0.30% for the 10% level.

Moreover, the size seems to be unaffected by the number of instruments used, or the nature of the serial correlation under the null (though the rejection frequencies are slightly smaller under negative serial correlation).

Next, we turn to the power of the test under local alternatives, indexed by the autoregressive parameter φ in the residual equation (3.10). This choice of alternative specification is chosen over sequences of MA(q+s) specifications for two reasons. First, because the AR formulation allows for infinite order serial correlation but the power function is one-dimensional and easier to visualize; and second, because this alternative specification is more realistic in practice, where the excess serial correlation is usually the result of omitted dynamics, and the omitted variables are usually well approximated by low-order autoregressive processes.

Figure 3.4 plots the power function of SC(1,s), for $s = 1, \dots, 5$ under the above alternatives. Due to the structure of the model, the null hypothesis is composite here, and corresponds to the pair $\{\psi = \psi_0, \varphi = 0\}$ and $\{\psi = \psi_0, \varphi = -\psi_0\}$ in (3.10). The former is obvious, whereas the latter reduces the ARMA error process (3.10) to white noise, which also fulfills the null hypothesis (ψ_0 is set equal to -0.5 for the results presented here).

The resulting power function is non-standard, exhibiting two troughs at 0 and $0.5 = -\psi_0$. Evidently, the test is consistent at the tails, and none of the versions is uniformly most powerful. Also, the increase in s seems to reduce power, showing that parsimony is to be preferred in this case, but this is specific to the type of alternative chosen and it seems to be a minor issue.[18] However, the most noteworthy feature is a distinct lack of power in the region $(0, -\psi_0)$. This is an important weakness of the test since the cases in which it has little power are likely to be the most relevant ones in forward-looking macro models, where the errors are negatively correlated by construction, and macroeconomic variables are usually positively autocorrelated. It also raises doubts about the utility of this test as an alternative to the Hansen-Sargan test of the over-identifying restrictions when lagged regressors are used as instruments.[19]

3.6 The QMD Identification test

In this section, we investigate the finite-sample size of the Quasi Minimum Distance (QMD) identification test and compare it to the (naive) Reduced Rank test, which is not robust to serial correlation or heteroscedasticity in the first-stage regression residuals. The experiments are based on a modification of Design III. The model and the first stage regressions are:

$$\begin{aligned} y_t &= \theta_1 \, Y_{1t} + \theta_2 \, Y_{2t} + u_t \\ \begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} &= \Pi' \, Z_t + \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix} \\ v_{1t} &= \varepsilon_t + \psi \, \varepsilon_{t-1} \end{aligned}$$

The model contains two endogenous regressors, and the DGP is such that the error in the auxiliary regression for one of them exhibits first-order serial correlation. This situation is chosen to resemble a dynamic rational expectations model where one of the endogenous regressors is y_{t+1} and its

---

[18] It is actually a benefit of the test, which suffers little power loss due to over-parameterization.

[19] These Monte Carlo results are limited in scope, and are given mainly to verify the asymptotic theory. Before recommending the use of such a test for forward-looking models it would be necessary to investigate its finite-sample properties in such situations where the hypothesized MA(q) serial correlation is an artefact of the model, due to the forecasting horizon. This can be investigated using the Monte Carlo Design IV. Preliminary experimentation suggests that the test is considerably under-sized in that case, and that the size distortions appear to fall slowly with the sample size.

forecast error on $(t-1)$-dated information exhibits at least MA(1) serial correlation, see chapter 4 for details. Π is set such that only $\theta_1 + \theta_2$ is well-identified, whereas $\theta_1 - \theta_2$ is completely unidentified under the null of partial identification and weakly identified under a local alternative, with the minimum eigenvalue of the concentration parameter equal to 1.

The QMD test will be based on a Newey-West covariance estimator, despite the fact that a more efficient estimator can be found in this design. The reason is that, unlike in the case of estimating a forward-looking model, where the serial correlation pattern in the residual is restricted by the orthogonality conditions, the first-stage regression is not meant to be a proper model of the dynamics of the endogenous regressors. Hence, higher-order serial correlation cannot be ruled out on the basis of correct specification, e.g., when an endogenous regressor Y_t follows an AR(p) process but only the first p-1 lags are used as instruments.

Table 3.14: Comparisons of rejection frequencies for the QMD and Reduced Rank regression identification tests and a *local* alternative.

		Under the null				Under the alternative			
k	T	QMD 10%	QMD 5%	RR 10%	RR 5%	QMD 10%	QMD 5%	RR 10%	RR 5%
(a) $\psi = 0$									
2	100	11.1%	4.8%	10.1%	4.8%	40.6%	25.8%	40.7%	28.3%
2	200	10.7%	5.0%	10.1%	5.1%	41.2%	28.0%	40.5%	29.4%
2	500	10.2%	5.2%	10.0%	5.1%	41.1%	28.7%	40.9%	29.2%
4	100	9.4%	2.8%	10.7%	5.6%	37.4%	18.7%	47.9%	34.9%
4	200	10.1%	4.1%	10.5%	5.2%	44.7%	28.7%	49.0%	36.4%
4	500	10.3%	4.5%	10.2%	4.7%	47.1%	33.3%	48.8%	35.9%
(b) $\psi = -0.6$									
2	100	9.7%	4.2%	4.2%	1.5%	44.2%	29.2%	34.1%	22.0%
2	200	10.3%	5.0%	4.3%	1.6%	46.3%	33.1%	34.5%	21.9%
2	500	10.4%	5.0%	3.9%	1.4%	49.2%	36.3%	34.3%	21.3%
4	100	7.2%	1.8%	2.8%	0.9%	40.0%	20.0%	34.4%	21.9%
4	200	9.1%	3.5%	2.4%	0.9%	50.5%	33.9%	34.9%	21.9%
4	500	10.4%	5.1%	2.3%	0.7%	57.4%	43.1%	34.5%	21.2%
(c) $\psi = 0.6$									
2	100	13.0%	6.1%	16.2%	9.2%	35.3%	21.8%	39.3%	28.6%
2	200	12.7%	6.3%	15.9%	9.5%	35.2%	23.0%	39.6%	29.0%
2	500	12.3%	6.7%	16.0%	9.8%	33.8%	23.3%	39.2%	29.0%
4	100	12.2%	4.3%	21.1%	12.9%	33.5%	16.4%	50.0%	38.5%
4	200	13.2%	5.9%	21.5%	12.7%	38.8%	23.4%	50.6%	39.2%
4	500	12.9%	6.1%	21.2%	12.6%	39.5%	26.5%	51.0%	39.0%

Asymptotic MC standard errors: 0.30% for the 10% level, and 0.22% for the 5% level.

Table 3.14 reports the rejection frequencies for the QMD and Reduced Rank regression tests for the hypothesis of partial identification of degree 1, i.e.,

$$\begin{aligned} \mathrm{H}_0 : &\quad \mathrm{rank}(\Pi) = 1 \\ \mathrm{H}_1 : &\quad \mathrm{rank}(\Pi) = 2 \end{aligned}$$

Under the local alternative, the minimum eigenvalue of $\Pi'\Pi$ is 1. Panel (a) corresponds to the case of no serial correlation and given that there is no heteroscedasticity either, the likelihood ratio reduced rank test (RR) is correctly sized and optimal. This is verified by the entries in the table. The QMD test, which unnecessarily corrects for serial correlation and heteroscedasticity is well-sized in most cases, and almost as powerful as the likelihood ratio test.

Panels (b) and (c) report rejection frequencies under negative and positive serial correlation, in which the RR test is inappropriate: it under-rejects under negative serial correlation, and over-rejects under positive serial correlation. The QMD test on the other hand has better size properties, although it tends to be distorted in the same direction as the RR test. The size distortion falls with the sample size, unlike the RR test, appears to be sensitive to the number of instruments, and is bigger under positive than under negative autocorrelation.

Even though its size is closer to the nominal level than the naive RR test, the QMD test is still far from accurate in finite samples. The observed finite-sample distortions in the null rejection probabilities may be the result of the inefficient estimate of the reduced rank matrix $\widehat{\Pi}$ in (2.31). There is some evidence that using the (unrestricted) full rank estimate $\widehat{\Pi}_p$ (OLS) aggravates the distortions. The restricted estimator used here is the pseudo MLE under no serial correlation or heteroscedasticity, which is also inefficient, but less so than the unrestricted one. Hence, more precision may be gained by using a two step GMM estimator, derived by minimizing the 'auxiliary' objective function (2.31) over all reduced rank matrices Π_r.

Another source of the problem, familiar from the results of section 3.2, is the HAC covariance estimator. The one used here is Newey-West with automatic bandwidth selection, and this was shown to have detrimental effects on the coverage of S-sets in section 3.3. Although we could try alternative nonparametric estimators, it seems unlikely to gain much improvement without using a parametric one, such as West's MA-l. However as argued above, such an approach requires more assumptions about the structure of the autocorrelation than are implied by the model at hand, which is an incomplete description of the local DGP. So, it seems impossible to pursue this route without modelling the endogenous variables directly.

Another important feature of the table is the high rejection frequencies under the local alternative. We see that the proposed identification test has considerable power in a case where conventional asymptotic theory still provides a poor description to the finite sampling distributions of estimators and test statistics, as was shown in section 3.2: see figure 3.1. This corroborates a similar finding of Staiger and Stock (1997, section 6.C) in this more general setting.

All in all, the QMD test seems to be a reasonable first attempt at a pretest of partial identification that is robust to serial correlation and heteroscedasticity, but it should be used cautiously both because its finite-sample rejection frequencies under the null are still inaccurate when the number of instruments is large, and because rejection does not necessarily imply strong identification.

3.7 Summary of key findings

1. The first-order asymptotic approximations of theorem 2.1 to the finite-sample distributions GMM estimators seems reasonably accurate both when the instruments are not strongly exogenous and when only combinations of the parameters are well-identified, verifying the theorem. HAC estimators of the GMM weighting matrix make little difference when not needed, but evidence suggests that the approximations are better with the parametric compared to the nonparametric corrections when errors are negatively serially-correlated. In a dynamic model with serially-correlated errors, such as a partial adjustment mechanism, the approximations deteriorate considerably even at moderately large samples, and higher-order terms are no longer negligible.

2. Analysis of bias of the 2-step versus the CUE relative to the bias of Least Squares revealed a bias-efficiency trade-off in the choice of estimators in all cases. The common finding that the bias of the CUE is generally smaller and less sensitive to the number of instruments than the 2S-GMM, at the expense of higher variability, is corroborated in static models, but not in a dynamic model with serially-correlated errors. In that case, the CUE doesn't appear to be preferable under weak instruments, since its bias is not significantly smaller than the 2S estimator, but it is considerably more dispersed. Also, the bias rises with the number of instruments and falls with the concentration parameter, as the asymptotic theory suggests.

3. The coverage rate of the S-set, which is robust to weak identification, exceeds its nominal level when lagged instruments are used, as well as under serial correlation. The use of the Newey-West nonparametric HAC estimator is hopeless for the S-set, since its coverage rate appears close to 100%, leading to extremely conservative inference and a large type II

error. The S-set is also sensitive to over-instrumenting, since its coverage rate rises with the number of (irrelevant) non-strongly exogenous instruments. All in all, these results raise doubts about the utility of this inferential approach for forward-looking macro-models.

4. The Hansen-Sargan test is also found to be under-sized for weak instruments under both 2S and the CUE, when errors exhibit negative serial correlation. This result does not generalize, since the asymptotic results of the previous chapter do not place any restrictions on the size distortions, and Monte Carlo studies elsewhere found the 2S-based test to be oversized in certain cases. Under strong identification, the test is correctly sized in the idealized static case with strongly exogenous instruments, but undersized under negative serial correlation with lagged instruments. Last, the finding that the CUE-based statistic is boundedly pivotal is corroborated, but doesn't appear useful in testing for mis-specification under weak instruments. This remains an open area of research, where contributions are much needed.
5. The test for excess serial correlation appears to be accurately sized, when conducted parsimoniously, but raising the degrees of freedom causes under-rejection in small samples. When the null hypothesis is one of negative serial correlation at lag 1, the test exhibits a distinct lack of power under alternatives that are relevant for forward-looking macro models, thus reducing its appeal as a parsimonious alternative to a Hansen-Sargan test for dynamic mis-specification.
6. Finally, the QMD test developed in the previous chapter has better size properties than a naive reduced rank test for identification, but it still suffers from size distortions in finite samples, especially as the number of instruments rises. More importantly, the test will have power against local alternatives under which normal asymptotic approximations are still unreliable. This shows that the use of conventional asymptotics after a pre-test for the lack of identification rejects is problematic.

Appendix

3.A Monte Carlo specificity and variance reduction

It is important to look for ways to reduce the specificity and increase the accuracy of Monte Carlo experiments. Based on Hendry (1983), we first look for invariances of the simulated statistics with respect to any nuisance parameters. Then, we propose a new control variate for the simulation of rejection frequencies, which dominates the ones proposed so far and is guaranteed to offer variance reduction of the simulated rejection probabilities.

3.A.1 Invariances

In this section, we discuss normalizations that can be carried out in order to limit the number of nuisance parameters which are relevant for the experiments. This will reduce computational cost dramatically.

We first normalize the instrumental variables, by setting $\Sigma_{\mathbf{zz}} = \boldsymbol{I}_k$ in 3.1, without loss of generality. The estimate $\widehat{\theta}$ as well as the resulting test statistics are invariant to the variance of the instruments Σ_{ZZ} since $P_Z = P_{ZA}$ for any symmetric positive definite matrix A.[20]

Similarly, the J-statistic is also invariant to the variance sub-matrix of the regressors in the first stage regression Σ_{vv}, since the structural residuals are unaffected by the orthonormalization of the regressors prior to estimation.[21] This last orthonormalization will, of course, affect both the parameters of interest θ, and the nuisance parameter Π. To see this, consider substituting $Y_t^* = A\,Y_t$ for Y_t in the model (3.2). The model and first stage regression will become:

$$\begin{aligned} y_t &= \theta_1' A^{-1} Y_t^* + \theta_2' Z_t + u_t \\ Y_t^* &= A\Pi' Z_t + A\,v_t \end{aligned}$$

In the new co-ordinate system for Y_t, both θ_1 and Π are different. By choosing $A = \Sigma_{vv}^{-1/2}$, we can orthogonalize the residuals v_t. Note that the concentration parameter is already normalized w.r.t. Σ_{vv}, so it will be unchanged.

Finally, the variance of the structural error σ_u could also be normalized to 1, where possible, since this is merely a change in the units of measurement of the regressand.

3.A.2 A new Control Variate

A control variable is a device designed to reduce the variation of a simulation experiment. In the case of power function estimation, the number of experiments required may be quite large, depending on the size of the domain of the function and the desired degree of smoothness. Oftentimes, even after allowing for invariances there remains a large set of values over which the function

---

[20] $P_{ZA} = Z\,A(A\,Z'Z\,A)^{-1}\,A\,Z' = Z\,A\,A^{-1}(Z'Z)^{-1}\,A^{-1}\,AZ' = P_Z$.

[21] In general, $\widehat{u} = M\,y$, where M is a residual projection matrix. In the IV case, $M = M_{Y \perp P_Z Y} = Y(Y'P_ZY)^{-1}Y'P_Z = I - Y\,A\,(A'Y'P_ZY\,A)^{-1}A'Y'P_Z$, for any non-singular matrix A.

needs to be simulated. Even moderate reductions in intra-experiment variance could potentially reduce the total computation time dramatically. Control variates can also yield important insights directly into the distributional behaviour.

Suppose we are interested in calculating the power function of a test statistic τ_Q of interest. The type of control variable one often uses is a test static τ_R, say, whose power function is known, and which is highly correlated with τ_Q. Let Q^{α}_{crit} and R^{α}_{crit} denote the critical regions for each test, for some nominal size α under the null hypothesis and define the indicator function $X = 1_{\{\tau_X \in X^{\alpha}_{crit}\}}$ for $X = \{Q, R\}$, i.e.,

$$Pr(X = 1|\mathrm{H}_0) = Pr(\tau_X \in X^{\alpha}_{crit}\,|\mathrm{H}_0) = \alpha.$$

Now, if the tests are correlated, we will frequently observe $Q = R = 1$ and $Q = R = 0$. This gives us a convenient way of simulating Q. Since $Pr(R)$ is known, and is equal to π say, the controlled experiment would be:

$$\widehat{p} = N^{-1}\sum q_j + \pi - N^{-1}\sum r_j \tag{3.A.1}$$

instead of the usual $\sum q_j/N$, where $\sum q_j/N$ and $\sum r_j/N$ are the average simulated rejection frequencies of the test statistics τ_Q and τ_R, respectively.

In this experiment, we shall use a different control variable.[22] First, notice that the distribution of (Q, R) is bivariate Bernoulli, with the density presented in the table below.

Variables	$R = 0$	$R = 1$
$Q = 0$	ϕ_{00}	ϕ_{01}
$Q = 1$	ϕ_{10}	ϕ_{11}

We can then simulate our probability of interest using the following conditional decomposition:

$$\begin{aligned} Pr(Q = 1) &= Pr(Q = 1\,|\,R = 1)Pr(R = 1) + Pr(Q = 1\,|\,R = 0)Pr(R = 0) \\ &= p_1\pi + p_0(1 - \pi) \end{aligned} \tag{3.A.2}$$

Since π is known we only need to simulate the p_is. Will this guarantee a variance reduction against the simple estimator $\bar{p} = \sum q_j/N$? The answer of course depends on the choice of estimators for p_i. We can prove the following proposition:

Proposition 3.1. *The estimator defined by*

$$\tilde{p} = \pi\tilde{p}_1 + (1 - \pi)\tilde{p}_0 \tag{3.A.3}$$

---

[22] I am grateful to Prof. Hendry for this suggestion.

where $\tilde{p}_i$ *are the maximum likelihood estimators of* p_i*,*

$$\tilde{p}_1 = \frac{\sum q_j r_j}{\sum r_j}, \quad \tilde{p}_0 = \frac{\sum q_j(1-r_j)}{\sum(1-r_j)}$$

dominates asymptotically (in the number of replications) the unconditional estimator $\sum q_j/N$ *by:*

$$A\mathrm{var}(\bar{p}) - A\mathrm{var}(\tilde{p}) = \frac{\pi(1-\pi)(p_1-p_0)^2}{N}$$

Proof. First, let us write down the likelihood of the Bernoulli experiment. The conditional density of Q given R is then given by:

$$f(q|r) = \begin{cases} p_1^q(1-p_1)^{1-q} & r=1, \quad q \in \{0,1\} \\ p_0^q(1-p_0)^{1-q} & r=0, \quad q \in \{0,1\} \end{cases} \tag{3.A.4}$$

Since the simulations are based on independent realizations of the set of time-series, the Monte Carlo sample of test outcomes is a set of N independent Bernoulli trials. Hence, the log-likelihood function, conditional on R, is:

$$\begin{aligned} l(q|r;p_1,p_0) &= \sum \big[q_i r_i \log p_1 + (1-q_i) r_i \log(1-p_1) \\ &\quad + q_i(1-r_i)\log p_0 + (1-q_i)(1-r_i)\log(1-p_0)\big]. \end{aligned} \tag{3.A.5}$$

The score vector and average information matrix can be easily found to be:

$$s(p_1,p_0;\mathbf{q},\mathbf{r}) = \begin{pmatrix} \frac{\sum q_i r_i}{p_1} - \frac{\sum(1-q_i)r_i}{1-p_1} \\ \frac{\sum q_i(1-r_i)}{p_0} - \frac{\sum(1-q_i)(1-r_i)}{1-p_0} \end{pmatrix} \tag{3.A.6}$$

$$\mathcal{I}(p_1,p_0) = \begin{pmatrix} \frac{\pi}{p_1(1-p_1)} & 0 \\ 0 & \frac{1-\pi}{p_0(1-p_0)} \end{pmatrix} \tag{3.A.7}$$

where we have used the definitions:

$$\mathrm{E}[QR] = Pr(Q=1,R=1) = Pr(Q=1|R=1)Pr(R=1) = p_1\pi$$

and similarly for the other three cases. The MLE estimators follow immediately by setting the score to 0. The asymptotic variance of the MLE estimators is the inverse of the average information, and hence the asymptotic variance of $\tilde{p}$ is:

$$\begin{aligned} NA\mathrm{var}(\tilde{p}) &= \pi^2 \frac{p_1(1-p_1)}{\pi} + (1-\pi)^2 \frac{p_0(1-p_0)}{1-\pi} \\ &= \pi p_1(1-p_1) + (1-\pi)p_0(1-p_0) \end{aligned} \tag{3.A.8}$$

On the other hand, the (scaled) variance of the unconditional estimator, $\bar{p} = \sum q_j/N$ is simply:

$$\begin{aligned} N A\text{var}(\bar{p}) &= p(1-p) \\ &= (\pi p_1 + (1-\pi)p_0)\,(\pi(1-p_0) + (1-\pi)(1-p_0)) \qquad (3.A.9) \end{aligned}$$

Subtracting (3.A.8) from (3.A.9), we obtain:

$$N[A\text{var}(\bar{p}) - A\text{var}(\tilde{p})] = \pi(1-\pi)(p_1-p_0)^2 \geq 0$$

as required. □

Finally, the following result shows why the conditional control variable is preferable to the most frequently used variance reduction technique defined in equation (3.A.1).

Corollary 3.2. *The estimator $\tilde{p}$ defined in (3.A.3) dominates the estimator $\widehat{p}$, defined in (3.A.1), for any choice of control variable R. Moreover, unlike the former, the latter cannot guarantee a variance reduction against the unconditional $\bar{p}$ estimator.*

Proof. The variance of $\widehat{p}$ is

$$\begin{aligned} N A\text{var}(\widehat{p}) &= p(1-p) + \pi(1-\pi) - 2\pi(p_1 - p) \\ &= p(1-p) + \pi(1-\pi)(1 - 2p_1 + 2p_0) \qquad (3.A.10) \end{aligned}$$

where the last term in the first line is due to the covariance of Q and R. From the previous proposition, $A\text{var}(\tilde{p}) = p(1-p) - \pi(1-\pi)(p_1-p_2)^2$. Hence,

$$\begin{aligned} N[A\text{var}(\widehat{p}) - A\text{var}(\tilde{p})] &= 1 - 2p_1 + 2p_2 + (p_1-p_0)^2 \\ &= (1 - p_1 + p_0)^2 \geq 0 \end{aligned}$$

□

The above results clearly show that the gains in efficiency increase with the correlation between the target and the control variables, i.e., as $(p_1 - p_0)$ rises. Of course, since the distribution of τ_R is known and that of τ_Q is not known, the conditioning cannot be perfect. Nevertheless, clever choices of τ_R (e.g. abstracting from dynamic sampling) may yield valuable gains.

3.B Computational appendix

Here, we give the procedures used to compute the implied model parameters and the required asymptotic approximations to the distributions of the various GMM estimators. The moment conditions will be linear, but the structural errors could be serially-correlated.

Let us collect all of the variables in a $(1+n+k)\times 1$ vector $\Upsilon_t = (y_t, Y_t', Z_t')'$, where n is the dimension of the endogenous regressors Y_t. Denote all the regressors in the generic structural model (3.2) by $\bar{Y}_t' = (Y_t', Z_t')\, S$, where $S = (S_Y', S_Z')'$ is a general $(n+k)\times p$ selection matrix that will pick the regressors out of the set $(Y_t', Z_t')'$, i.e.,[23]

$$\bar{Y}_t = S' \begin{pmatrix} Y_t \\ Z_t \end{pmatrix} = \underset{p\times n}{S_Y'}\, Y_t + \underset{p\times k}{S_Z'}\, Z_t$$

Next, we need to derive the autocovariance function of the whole data vector Υ_t. We will denote this by

$$\Gamma_{\Upsilon,j} = \mathrm{E}\left(\Upsilon_t \Upsilon_{t-j}'\right), \quad j \in [-J, J]. \tag{3.B.11}$$

This notation will refer to the asymptotic ACF, i.e., when $T = \infty$. In the case of some parameters of the DGP being local-to-zero, this ACF will correspond to those parameters being set *equal* to zero.

We shall also make use of the 'finite-T' second moments of Υ_t, which will be denoted by $\Sigma_{\Upsilon\Upsilon,T} \to \Sigma_{\Upsilon\Upsilon} = \Gamma_{\Upsilon,0}$, where $\Sigma_{\Upsilon\Upsilon} = \Sigma_{\Upsilon\Upsilon,\infty}$. Hence, the 'finite-sample' covariance between the regressors and the instruments is

$$\Sigma_{Z\bar{Y},T} = \begin{pmatrix} 0 & 0 & I_k \end{pmatrix} \Sigma_{\Upsilon\Upsilon,T} \begin{pmatrix} 0 \\ S \end{pmatrix}.$$

The coefficient $\bar{\Pi}_T$ of the first-stage regression (3.5) is given by:

$$\bar{\Pi}_T = (\underbrace{\Sigma_{ZZ,T}^{-1}\Sigma_{ZY,T}}_{\Pi_T} \;\; I_k)\, S = \Sigma_{ZZ,T}^{-1}\Sigma_{Z\bar{Y},T} \to \bar{\Pi}, \quad \text{as} \quad T \to \infty$$

(the latter being the coefficient that is implied when the local-to-zero parameters are set to zero). Notably, when this leads to under-identification of degree $p_1 = p - p_2$, there exist $k \times p_2$ and $p \times p_2$ full rank matrices B and d respectively, such that:

$$\bar{\Pi}_\infty = B\, d'.$$

The matrix d and its orthogonal complement $d_\perp$ can be determined (up to rotation) straightforwardly by a singular value decomposition of $\bar{\Pi}_\infty$. This yields:

$$\Phi = \begin{pmatrix} \bar{d}_\perp & \bar{d} \end{pmatrix} = \begin{pmatrix} d_\perp (d_\perp' d_\perp)^{-1} & d\,(d'd)^{-1} \end{pmatrix}$$

---

[23] For instance, $\bar{Y}_t = \left(Y_t', Z_{1t}'\right)'$ would correspond to

$$S_Y = \left(I_n, 0_{n\times k_1}\right) \quad \text{and} \quad S_Z = \begin{pmatrix} 0_{k_1\times n} & I_{k_1} \\ 0_{k_2\times k_1} & 0_{k_2\times n} \end{pmatrix}.$$

and hence, one re-parameterization of θ into weakly and well-identified parameters can be given by (2.5), i.e.,

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \Phi^{-1}\,\theta \quad \text{and} \quad \phi(\alpha,\beta) = \Phi \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \Phi_1\,\alpha + \Phi_2\,\beta = \bar{d}_\perp\,\alpha + \bar{d}\,\beta. \tag{3.B.12}$$

From that, we can derive

$$C^* = \sqrt{T_0}\,\left(\bar{\Pi}_{T_0} - \bar{\Pi}_\infty\right) \quad \text{and} \quad C = \begin{pmatrix} C_\alpha & C_\beta \end{pmatrix} = C^*\,\Phi.$$

The first-stage regression errors will also prove useful:

$$\bar{v}_t = \bar{Y}_t - (\bar{\Pi}_T^*)' Z_t = \kappa' \Upsilon_t.$$

Also, define

$$\iota(\theta) = \begin{pmatrix} 1 \\ -S\,\theta \end{pmatrix}, \quad \bar{\iota}(\theta) = \begin{pmatrix} 1 \\ -S_Y\,\theta \\ \Sigma_{ZZ}^{-1}\,(\Sigma_{ZY}\,S_Y\,\theta - \Sigma_{Zy}) \end{pmatrix}, \quad \kappa = \begin{pmatrix} \underset{1\times p}{0} \\ S_Y \\ -\Sigma_{ZZ}^{-1}\,\Sigma_{ZY} S_Y \end{pmatrix} \tag{3.B.13}$$

and $\tilde{\kappa}(\theta) = (\bar{\iota}(\theta),\ \kappa)$. Using the notation in (3.B.13), the various quantities can be written as (see table 2.1 on page 14 for definitions and descriptions):

$$\begin{array}{lcl} h_t(\theta) & = & y_t - \theta' \bar{Y}_t = \iota(\theta)' \Upsilon_t \\ u_t & = & h_t(\theta_0) = \iota(\theta_0)' \Upsilon_t \\ f_t(\theta) & = & Z_t\,h_t(\theta) = Z_t\,\Upsilon_t'\,\iota(\theta) = Z_t\,u_t + Z_t\,\bar{Y}_t'\,(\theta_0 - \theta) \\ \mathrm{E}\left[h_t(\theta) |\ Z_t\right] & = & \mathrm{E}\left[\iota(\theta)'\Upsilon_t |\ Z_t\right] = \iota(\theta)' \Sigma_{\Upsilon Z} \Sigma_{ZZ}^{-1} Z_t = \left[(\Sigma_{Zy} - \Sigma_{ZY}\,S_Y\theta)\,\Sigma_{ZZ}^{-1} - S_Z\theta\right]' Z_t \\ \bar{h}_t(\theta) & = & h_t(\theta) - \mathrm{E}\left[h_t(\theta) |\ Z_t\right] = \iota(\theta)'\Upsilon_t - \iota(\theta)' \Sigma_{\Upsilon Z} \Sigma_{ZZ}^{-1} Z_t = \bar{\iota}(\theta)' \Upsilon_t. \end{array} \tag{3.B.14}$$

Next, to verify assumption D, we compute $T^{-1} \sum_{t=1}^T \mathrm{E}(f_t(\theta) | \mathcal{F}_t)$ and its asymptotic approximation to order $O_p(T^{-1/2})$.

$$\frac{1}{T} \sum_{t=1}^T \mathrm{E}\left(f_t(\theta)\,| Z_t\right) = \frac{1}{T} \sum_{t=1}^T Z_t\,\mathrm{E}(u_t\,|\,Z_t) \underbrace{- \frac{1}{T} \sum_{t=1}^T Z_t\,\mathrm{E}\left(\bar{Y}_t'\,|\,Z_t\right)\,(\theta - \theta_0)}_{m_T(\theta)} \tag{3.B.15}$$

Under correct specification (assumption D) the first term will be identically equal to zero. The second term will be:

$$\begin{aligned}
m_T(\theta) &= -\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t \Sigma_{ZZ,T}^{-1}\Sigma_{Z\bar{Y},T}\,(\theta-\theta_0) = \frac{Z'Z}{T}\bar{\Pi}_T\,(\theta-\theta_0)\\
&= -\left[\Sigma_{ZZ}\bar{\Pi} + \left(\frac{Z'Z}{T}-\Sigma_{ZZ}\right)\bar{\Pi} + \frac{Z'Z}{T}\left(\bar{\Pi}_T-\bar{\Pi}\right)\right](\theta-\theta_0)\\
&= -\left[\Sigma_{ZZ}\,B\,d' + \left(\frac{Z'Z}{T}-\Sigma_{ZZ}\right)\,B\,d' + \Sigma_{ZZ}\frac{C^*}{\sqrt{T}} + \left(\frac{Z'Z}{T}-\Sigma_{ZZ}\right)\frac{C^*}{\sqrt{T}}\right](\theta-\theta_0)\\
&= \underbrace{-\left[\Sigma_{ZZ}\,B + \left(\frac{Z'Z}{T}-\Sigma_{ZZ}\right)\,B\right](\beta-\beta_0)}_{m_{2T}(\beta)} - \frac{1}{\sqrt{T}}\underbrace{\left[\Sigma_{ZZ}\,C^* + \left(\frac{Z'Z}{T}-\Sigma_{ZZ}\right)\,C^*\right](\theta-\theta_0)}_{m_{1T}(\theta)}
\end{aligned}$$

Under the stationarity and ergodicity of Z_t, $\frac{Z'Z}{T}-\Sigma_{ZZ}=O_p(T^{-1/2})$, and hence,

$$\begin{aligned}
m_{1T}(\theta) &= -\Sigma_{ZZ}\,C^*\,(\theta-\theta_0)+o_p(1)\\
m_1(\phi(\alpha,\beta)) &= -\Sigma_{ZZ}\,[C_\alpha(\alpha-\alpha_0)+C_\beta(\beta-\beta_0)]\\
m_{2T}(\beta) &= -\Sigma_{ZZ}\,B\,(\beta-\beta_0)+o_p(1)\\
\mathcal{J}_2(\beta_0) &= \nabla_\beta\, m_2(\beta_0) = -\Sigma_{ZZ}\,B
\end{aligned}$$

The implied 'true' parameter θ_0 can be derived by solving $m_T(\theta)=0$, which, using (3.B.15) implies:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} Z_t\,\mathrm{E}(u_t\,|\,Z_t) &= \frac{1}{T}\sum_{t=1}^{T} Z_t\,Z_t'\,\Sigma_{ZZ,T}^{-1}\,\Sigma_{Z\Upsilon,T}\,\iota(\theta_0)=0\\
&\Rightarrow\ \Sigma_{Zy,T}-\Sigma_{Z\bar{Y},T}\theta_0=0\\
&\Rightarrow\ \theta_0=(\Sigma_{Z\bar{Y},T}'\Sigma_{Z\bar{Y},T})^{-1}\Sigma_{Z\bar{Y},T}'\Sigma_{Zy,T} \qquad (3.B.16)
\end{aligned}$$

Next, for the distribution theory we need to specify the following:

$$\begin{aligned}
g_T(\theta) &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T} Z_t\Upsilon_t'\,\iota(\theta) = \tfrac{1}{T}\sum_{t=1}^{T} Z_t u_t - \tfrac{1}{T}\sum_{t=1}^{T} Z_t\,\bar{Y}_t'\,(\theta-\theta_0)\\
\Psi_T(\theta_0) &= \tfrac{1}{\sqrt{T}}\textstyle\sum_{t=1}^{T} Z_t\,h_t(\theta_0) = \tfrac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t u_t = \Psi(\theta_0)+o_p(1)\\
\Xi_T(\theta_0) &= T^{-1/2}\textstyle\sum_{t=1}^{T}\left(Z_t\bar{Y}_t' - Z_tZ_t'\bar{\Pi}_T^*\right) = T^{-1/2}\sum_{t=1}^{T} Z_t\Upsilon_t'\,\kappa \qquad (3.B.17)\\
&= T^{-1/2}\textstyle\sum_{t=1}^{T} Z_t\,\bar{v}_t' = \Xi(\theta_0)+o_p(1)\\
\Psi_T(\theta) &= \Psi(\theta_0)-\Xi(\theta_0)\,(\theta-\theta_0)+o_p(1)
\end{aligned}$$

Using the definitions (3.B.13),

$$\underbrace{\begin{pmatrix}\Psi_T(\theta_0) & \Xi_T(\theta_0)\end{pmatrix}}_{k\times(1+p)} = T^{-1/2}\sum_{t=1}^{T} Z_t\,\Upsilon_t'\tilde{\kappa}(\theta_0)$$

To simplify the notation, define the $k(1+p)$-dimensional random vector $\boldsymbol{\xi}$ by:

$$\boldsymbol{\xi}=\begin{pmatrix}\boldsymbol{\xi}_u\\ \boldsymbol{\xi}_v\end{pmatrix}=\begin{pmatrix}\Psi(\theta_0)\\ \mathrm{vec}\,(\Xi(\theta_0))\end{pmatrix}\sim N_{k(1+p)}\,[0,\mathbf{V}(\theta_0)]$$

defined in assumption B$'$. Note also that

$$\mathrm{vec}\left(\begin{pmatrix}\Psi_T(\theta_0) & \Xi_T(\theta_0)\end{pmatrix}\right) = T^{-1/2}\sum_{t=1}^{T}\tilde{\kappa}(\theta_0)'\Upsilon_t \otimes Z_t = \boldsymbol{\xi} + o_p(1)$$

Thus, since Υ_t is a weakly stationary and conditionally homoscedastic process with autocovariance function $\Gamma_{\Upsilon,j}$, the limiting $k(1+p)$-dimensional variance matrix $\mathbf{V}(\theta_0)$ can be derived using (2.22):

$$\mathbf{V}(\theta_0) = \sum_{j=-J}^{J}\tilde{\kappa}(\theta_0)'\;\Gamma_{\Upsilon,j}\;\tilde{\kappa}(\theta_0)\times\Gamma_{Z,j} \tag{3.B.18}$$

with the lag truncation parameter J determined as in (2.23). The top left $k\times k$ sub-matrix of $\mathbf{V}(\theta_0)$ is simply the variance of the moment conditions evaluated at θ_0, namely $V(\theta_0)$. We will also need this at $\theta\neq\theta_0$, i.e., we will need the limiting variance of $\Psi_T(\theta)$. The latter can be written as:

$$\Psi_T(\theta) = \begin{pmatrix}\Psi_T(\theta_0) & \Xi_T(\theta_0)\end{pmatrix}\begin{pmatrix}1\\ \theta_0-\theta\end{pmatrix} = \boldsymbol{\xi}_u + \left[(\theta_0-\theta)'\otimes I_k\right]\boldsymbol{\xi}_v + o_p(1)$$

so that its variance $V(\theta)$ is simply:

$$V(\theta) = \begin{pmatrix}I_k & (\theta_0-\theta)'\otimes I_k\end{pmatrix}\mathbf{V}(\theta_0)\begin{pmatrix}I_k\\ (\theta_0-\theta)\otimes I_k\end{pmatrix}. \tag{3.B.19}$$

Finally, we define

$$\begin{aligned}\tilde{W}(\alpha) &= V(\phi(\alpha,\beta_0))^{-1}, \quad \text{where}\\ V(\phi(\alpha,\beta_0)) &= \begin{pmatrix}I_k & (\alpha_0-\alpha)'\bar{d}_\perp'\otimes I_k\end{pmatrix}\mathbf{V}(\theta_0)\begin{pmatrix}I_k\\ \bar{d}_\perp(\alpha_0-\alpha)\otimes I_k\end{pmatrix}.\end{aligned} \tag{3.B.20}$$

GMM objective function Let $\bar{\theta}_T = \bar{\theta}_T(\theta)$ and $\bar{\theta} = \mathrm{plim}(\bar{\theta}_T) = \bar{\theta}(\theta)$ for brevity. Given a weighting matrix $W_T\left(\bar{\theta}_T\right)$ which is a consistent estimator of the inverse of $V(\bar{\theta})$, as defined in (3.B.19), the GMM objective function is:

$$\mathcal{Q}_T(\theta;\bar{\theta}_T) = \iota(\theta)'\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Upsilon_t Z_t'\right)W_T(\bar{\theta}_T)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Z_t\Upsilon_t'\right)\iota(\theta)$$

By theorem 2.1 and using (3.B.12), (3.B.14) and (3.B.20), the limiting objective function is

$$\begin{aligned}\bar{\mathcal{Q}}(\alpha,b;\bar{\alpha}) &= \Big[\overbrace{\boldsymbol{\xi}_u - \left((\alpha-\alpha_0)'\,\bar{d}_\perp'\otimes I_k\right)\boldsymbol{\xi}_v}^{\Psi(\phi(\alpha,\beta_0))} + \overbrace{\Sigma_{ZZ}\,C_\alpha\,(\alpha_0-\alpha)}^{m_1(\phi(\alpha,\beta_0))} + \overbrace{(-\Sigma_{ZZ}\,B)\,b}^{\mathcal{J}_2(\beta_0)\,b}\Big]'\\ &\times\;\tilde{W}(\bar{\alpha})\;\left[\boldsymbol{\xi}_u - \left((\alpha-\alpha_0)'\,\bar{d}_\perp'\otimes I_k\right)\boldsymbol{\xi}_v - \Sigma_{ZZ}\,C_\alpha\,(\alpha-\alpha_0) - \Sigma_{ZZ}\,B\,b\right]\end{aligned}$$

where $b=\sqrt{T}(\beta-\beta_0)$, and $\bar{\alpha}=\alpha_{m-1}$ for the m-step estimator and $\bar{\alpha}=\alpha$ for the CUE. To reduce the computational burden of simulating the distribution of the above quantity at its minimum over (α, b), we concentrate the objective function w.r.t. b, so that

$$\begin{aligned}
\mathcal{Q}^*(\alpha) &= \left[\boldsymbol{\xi}_u - \left((\alpha-\alpha_0)'\,\bar{d}_\perp' \otimes I_k\right)\boldsymbol{\xi}_v - \Sigma_{ZZ}\,C_\alpha\,(\alpha-\alpha_0)\right]' M(\alpha) \\
&\quad \times \left[\boldsymbol{\xi}_u - \left((\alpha-\alpha_0)'\,\bar{d}_\perp' \otimes I_k\right)\boldsymbol{\xi}_v - \Sigma_{ZZ}\,C_\alpha\,(\alpha-\alpha_0)\right], \\
M(\alpha) &= \tilde{W}(\alpha) - \tilde{W}(\alpha)\Sigma_{ZZ}\,B\left[B'\,\Sigma_{ZZ}'\,\tilde{W}(\alpha)\Sigma_{ZZ}\,B\right]^{-1} B'\,\Sigma_{ZZ}'\,\tilde{W}(\alpha), \qquad (3.B.21) \\
b^* &= -\left[B'\,\Sigma_{ZZ}'\,\tilde{W}(\alpha^*)\Sigma_{ZZ}\,B\right]^{-1} B'\Sigma_{ZZ}'\tilde{W}(\alpha^*) \\
&\quad \times \left[\boldsymbol{\xi}_u - \left((\alpha^*-\alpha_0)'\,\bar{d}_\perp' \otimes I_k\right)\boldsymbol{\xi}_v - \Sigma_{ZZ}\,C_\alpha\,(\alpha^*-\alpha_0)\right]
\end{aligned}$$

Using the fact that $\Psi(\theta) = \boldsymbol{\xi}_u + [\boldsymbol{\xi}_v]_{k\times p}\,(\theta_0 - \theta)$, where $[\cdot]_{n\times m}$, shapes an nm-vector into an $n\times m$ matrix, the concentrated objective function becomes:

$$\begin{aligned}
\mathcal{Q}^*(\alpha) &= \left[\boldsymbol{\xi}_u - [\boldsymbol{\xi}_v]_{k\times p}\,\bar{d}_\perp\,(\alpha-\alpha_0) - \Sigma_{ZZ}\,C_\alpha\,(\alpha-\alpha_0)\right]' \\
&\times \quad M(\bar{\alpha})\left[\boldsymbol{\xi}_u - [\boldsymbol{\xi}_v]_{k\times p}\,\bar{d}_\perp\,(\alpha-\alpha_0) - \Sigma_{ZZ}\,C_\alpha\,(\alpha-\alpha_0)\right]
\end{aligned}$$

In the case of a multi-step estimator, this is quadratic in α and hence the minimization problem can be carried out analytically to yield:

$$\begin{aligned}
\alpha_m^* &= \alpha_0 + \left[\left([\boldsymbol{\xi}_v]_{k\times p}\,\bar{d}_\perp + \Sigma_{ZZ}\,C_\alpha\right)' M(\alpha_{m-1}^*)\,\left([\boldsymbol{\xi}_v]_{k\times p}\,\bar{d}_\perp + \Sigma_{ZZ}\,C_\alpha\right)\right]^{-1} \\
&\times \quad \left([\boldsymbol{\xi}_v]_{k\times p}\,\bar{d}_\perp + \Sigma_{ZZ}\,C_\alpha\right)' M(\alpha_{m-1}^*)\,\boldsymbol{\xi}_u
\end{aligned} \qquad (3.B.22)$$

A note on numerical solution of the CUE In the case of the CUE, this simplification is not possible. In fact, in that case the objective function behaves much like a ratio of second order polynomials, is clearly bounded from below, but need not always have a minimum in α, for some realizations of $\boldsymbol{\xi}$. This possibility arises with positive, albeit small probability. For a simple linear two parameter model with one degree of under-identification (α being a scalar), four typical realizations of the (negative of the) limiting objective function were simulated and presented in figure 3.5. The non-monotonicity is common in all graphs and could pose problems depending on the initial value. To avoid reaching a local minimum, the procedure is preceded by a small grid search. More rare is the occurrence of panel (d) where no minimum exists. This happened less than 2% of the time in our simulations.

Other parameters In the case of a single parameter, a measure of the 'endogeneity' of the regressor $\bar{Y}_t$ is given by the correlation between the structural error u_t and the first-stage regression forecast error $\bar{v}_t$, as was used for example in the Monte Carlo studies of Staiger and Stock (1997).

In the many parameter case, a generalization of this could be $\Sigma_{\bar{v}\bar{v}}^{-1}\Sigma_{\bar{v}u}$, i.e., the regression coefficient of u_t on $\bar{v}_t$. However, there are two problems with this. First, when exogenous regressors are present, $\Sigma_{\bar{v}\bar{v}}$ is singular and this cannot be computed. Secondly, if we wanted to restrict attention only on the forecast error of the 'endogenous' regressors, these need not always be clearly discernible from the exogenous ones (if, e.g., the selection matrix is such that combinations of Y_t and Z_t are used as regressors). Also, we expect that the impact of any nuisance parameter on estimator distributions will be primarily through the weakly identified parameter combinations α. Thus, we will consider a measure of the endogeneity of the combination of regressors corresponding to the weakly identified parameters $\alpha = d'_\perp \theta$:

$$\lambda = (d'_\perp \, \Sigma_{\bar{v}\bar{v}} \, d_\perp)^{-1} d'_\perp \, \Sigma_{\bar{v}u} = \left(d'_\perp \, \kappa' \, \Sigma_{\Upsilon\Upsilon} \, \kappa \, d_\perp\right)^{-1} d'_\perp \, \kappa' \, \Sigma_{\Upsilon\Upsilon} \, \iota(\theta_0).$$

Next, we compute the eigenvalues of the generalized concentration parameter. The only complication is that, when exogenous regressors are present in the model, the first stage regression error variance $\Sigma_{\bar{v}\bar{v}}$ is singular, so the concentration parameter matrix cannot be computed in the usual way (hence the need to focus attention on the endogenous parameters only). However, we may still compute the eigenvalues of this generalized concentration parameter, by solving the generalized eigenproblem:[24]

$$|T \, \Sigma'_{Z\bar{Y},T} \, W \, \Sigma_{Z\bar{Y},T} - \lambda \, \Sigma_{\bar{v}\bar{v}} / \sigma_u^2| = 0$$

using the Schur Decomposition.[25] When $\Sigma_{\bar{v}\bar{v}}$ is singular, some of the generalized eigenvalues will be 'infinite', thus demonstrating that the signal from the instruments relative to the noise in those directions is infinite (another way of viewing this is that the canonical correlations of exogenous regressors with themselves is unitary).

Mis-specification When the model is not correctly specified, but there exists a unique non-random minimum to the limiting objective function, this can be thought of as the pseudo-true parameter θ_p.[26] In general, $\theta_p \neq \theta_0$, unless the mis-specification is local-to-zero, i.e., disappears asymptotically. In all other cases, θ_p will differ for different weighting matrices W, and also for different estimation methods:

$$\begin{array}{llcl} (m\text{-step}) & \theta_p^{(m)} & = & \left(\Sigma'_{Z\bar{Y}} W \left(\theta_p^{(m-1)}\right) \Sigma_{Z\bar{Y}}\right)^{-1} \Sigma'_{Z\bar{Y}} W \left(\theta_p^{(m-1)}\right) \Sigma_{Zy}, \quad m = 1, 2, \ldots \\ (\text{CUE}) & \theta_p^{(c)} & = & \text{argmin}_{\theta \in \mathbf{\Theta}} \, (\Sigma_{Zy} - \Sigma_{Z\bar{Y}} \, \theta)' \, W(\theta) (\Sigma_{Zy} - \Sigma_{Z\bar{Y}} \, \theta) \end{array} \tag{3.B.23}$$

---

[24] This definition is meaningful provided $\bar{v}_t$ is conditionally homoscedastic.

[25] I thank Jurgen Doornik for implementing this decomposition in the latest version of Ox, version 3.2.

[26] For the pseudo-true parameter to be non-random, the model must be well-identified in the sense that $\theta = \beta$, i.e., that $m_T(\theta) = m_{2T}(\beta)$ in (3.B.15) above. Otherwise, the minimizer of the limiting objective function will be random.

where $W\left(\theta_p^{(0)}\right) = \Sigma_{ZZ}^{-1}$ or any other arbitrary non-singular matrix.

Under *local* mis-specification (assumption D′), the first term in (3.B.15) will not vanish for finite T:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} Z_t\,\mathrm{E}(u_t\,|\,Z_t) &= \frac{1}{T}\sum_{t=1}^{T} Z_t\,Z_t'\,\Sigma_{ZZ}^{-1}\,\Sigma_{Z\Upsilon,T}\,\iota(\theta_0) \\
&= \underbrace{\Sigma_{Z\Upsilon}\,\iota(\theta_0)}_{=0} + \left(\frac{Z'Z}{T} - \Sigma_{ZZ}\right)\Sigma_{ZZ}^{-1}\underbrace{\Sigma_{Z\Upsilon}\,\iota(\theta_0)}_{=0} + \frac{Z'Z}{T}\Sigma_{ZZ}^{-1}\underbrace{\left(\Sigma_{Z\Upsilon,T} - \Sigma_{Z\Upsilon}\right)}_{\mathrm{C}/\sqrt{T}}\,\iota(\theta_0) \\
&= \underbrace{T^{-1/2}\,\mathrm{C}\,\iota(\theta_0)}_{O(T^{-1/2})} + \underbrace{\left(\frac{Z'Z}{T} - \Sigma_{ZZ}\right)\Sigma_{ZZ}^{-1}\,T^{-1/2}\,\mathrm{C}\,\iota(\theta_0)}_{O_p(T^{-1})} = \zeta_T + o_p(T^{-1/2}).
\end{aligned}$$

So, the local-to-zero mis-specification can be computed as:

$$\zeta_T = \left(\Sigma_{Z\Upsilon,T} - \Sigma_{Z\Upsilon}\right)\,\iota(\theta_0)$$

Last, we can compute an approximate measure of the *mis-specification parameter* defined in (2.34) by:

$$\nu^2 = T\,\zeta_T'\left[V(\theta_0)^{-1} - V(\theta_0)^{-1}B\,d'\left(d\,B'\,V(\theta_0)^{-1}\,B\,d'\right)^{-1}\,d\,B'\,V(\theta_0)^{-1}\right]\zeta_T \tag{3.B.24}$$

3.C Monte Carlo designs

Here, we give the details of the four Monte Carlo designs used in this chapter and elsewhere.

Design I

This is a static one-parameter model with one endogenous variable

$$\text{Model:}\quad y_t = \theta\,Y_{1t} + u_t,$$

where $y_t = \mathrm{y}_{1t}$ and $Y_t = \mathrm{y}_{2t}$, and the DGP given in (3.1) is also static, i.e. $A_1 = A_2 = 0$ and $C = 0$:

$$\begin{aligned}
\overbrace{\begin{pmatrix}\mathrm{y}_{1t}\\ \mathrm{y}_{2t}\end{pmatrix}}^{\mathrm{y}_t} &= \overbrace{\begin{pmatrix}0_{1\times 8}\\ \Pi'\end{pmatrix}}^{A_z'}\mathbf{z}_t + \mathbf{e}_t \\
\mathbf{e}_t &\sim NID_2\left[0,\begin{pmatrix}1 & \lambda\\ \lambda & 1\end{pmatrix}\right] \\
\mathbf{z}_t &\sim NID_k(0, I).
\end{aligned}$$

The instruments are taken from $\mathbf{z}_t$ (strongly exogenous), i.e. $Z_t = (\mathbf{z}_{1t}, \ldots, \mathbf{z}_{kt})'$ for $k = 1, 4$ and 8. The identification of θ depends on the $k \times 1$ matrix Π, which is the second column of A_z, and the concentration parameter is a scalar, $\mu^2_{min} = \Pi'\Pi$, given the normalization $\Sigma_{ZZ} = \Sigma_{\mathbf{vv}} = I_k$, in this case. No identification corresponds to $\Pi = 0$, and weak identification to $\Pi = O(T^{-1/2})$.

There are several different configurations for Π that would correspond to the same concentration per instrument (μ^2_{min}/k), e.g.,

$$\Pi = \left(\frac{1}{T}\right)^{1/2} \begin{pmatrix} \mu_{min} \\ \vdots \\ \mu_{min} \end{pmatrix} \quad \text{or} \quad \Pi = \left(\frac{k}{T}\right)^{1/2} \begin{pmatrix} \mu_{min} \\ 0 \\ \vdots \end{pmatrix}$$

In other words, we could 'spread out' the information across the instruments;[27] or we may gather all of the information in one instrument, the remaining instruments being irrelevant. To check whether different choices of Π have any bearing on the statistics of interest, we compared the resulting distributions using the above two configurations and plotted a few examples in figure 3.6. As seen from those graphs, the choice of Π is inconsequential. In other words, only the amount of information matters, not its source.

Given the above, we will proceed by making only one of the instruments relevant (embodying all the relevant information), so that the first stage regression for Y_{1t} is:

$$Y_t = [(k/T)^{1/2}\mu_{min}]\,\mathbf{z}_{1t} + v_{1t}$$

Design II

This design is used to investigate the impact of using non-strongly exogenous instruments. It differs from design I in that the instruments are now lags of the dependent variable $Z_t = y_{t-1}, \ldots, y_{t-k}$, for $k = 1, 4$ and 8.

The model is the same as before (static), but the DGP is dynamic ($A_z = 0$ but $A_1 \neq 0$), the information arising through Granger-causality of $y_t = \mathrm{y}_{1t}$ for the endogenous regressor $Y_t = \mathrm{y}_{2t}$:

$$\overbrace{\begin{pmatrix} \mathrm{y}_{1t} \\ \mathrm{y}_{2t} \end{pmatrix}}^{\mathrm{y}_t} = \overbrace{\begin{pmatrix} 0 & 0 \\ a_T & 0 \end{pmatrix}}^{A_1} \mathrm{y}_{t-1} + \mathbf{e}_t$$

where $\mathbf{e}_t$ is distributed as for Design I, $\mathbf{z}_t$ is now irrelevant, and $a_T = (k/T)^{1/2}\mu_{min}$ is set so as to achieve the desired value for the concentration parameter. This follows from the first stage

---

[27] Here, we do it evenly, but any other scheme would do.

regression which now is:

$$Y_t = [(k/T)^{1/2}\mu_{min}]\, y_{t-1} + v_{1t}$$

Finally, $\theta_0 = 0$ as before.

Design III

Design III investigates the case of a linear combination of parameters being identified. The model now has two parameters, and is of the form:

$$\text{Model:} \quad y_t = \theta_1\, Y_t + \theta_2\, Z_{1t} + u_t$$

where $y_t = \mathrm{y}_{1t}$, $Y_t = \mathrm{y}_{2t}$ and $Z_{1t} = \mathbf{z}_{1t}$. The instruments are the strongly exogenous variables $Z_t = (\mathbf{z}_{1t}, \ldots, \mathbf{z}_{kt})$, $k = 2, 4$ and 8.

The DGP is static $A_1 = A_2 = 0$ and $C = 0$ as in Design I, but the A_z matrix is now different:

$$A_z = \begin{pmatrix} \theta_1 + \theta_2 & 1 \\ a_T\, \theta_1 & a_T \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $a_T = (c\, k/T)^{1/2}\mu_{min}$ so as to get the desired value of the concentration parameter.[28] This can be seen from the first stage regression for Y_t, which is:

$$Y_t = \mathbf{z}_{1t} + [(c\, k/T)^{1/2}\mu_{min}]\, \mathbf{z}_{2t} + v_{1t}$$

The degree of under-identification is at most 1 (since $\mathbf{z}_{1t}$ is an exogenous regressor), and the well-identified parameter combination β is $\theta_1 + \theta_2$, whereas $\theta_1 - \theta_2 = \alpha$ is weakly identified. The normalization $\alpha_0 = 0$ can be achieved by setting $\theta_1 = \theta_2 = 0.5$, say.

Design IV

This is an extension of Designs II and III, where the model is dynamic and akin to a partial adjustment mechanism:

$$\text{Model:} \quad \underset{\mathrm{y}_{1t}}{y_t} = \theta_1 \underset{\mathrm{y}_{1,t+1}}{Y_t} + \theta_2 \underset{\mathrm{y}_{1,t-1}}{Z_{1t}} + u_t,$$

with additional instruments $Z_{2t} = (\mathbf{z}_{1t}, \ldots, \mathbf{z}_{1,t-k+1})'$ or $Z_{2t} = (\mathrm{y}_{1,t-2}, \ldots, \mathrm{y}_{1,t-k})$, $k = 2, 4$ and 8. The DGP is also dynamic:

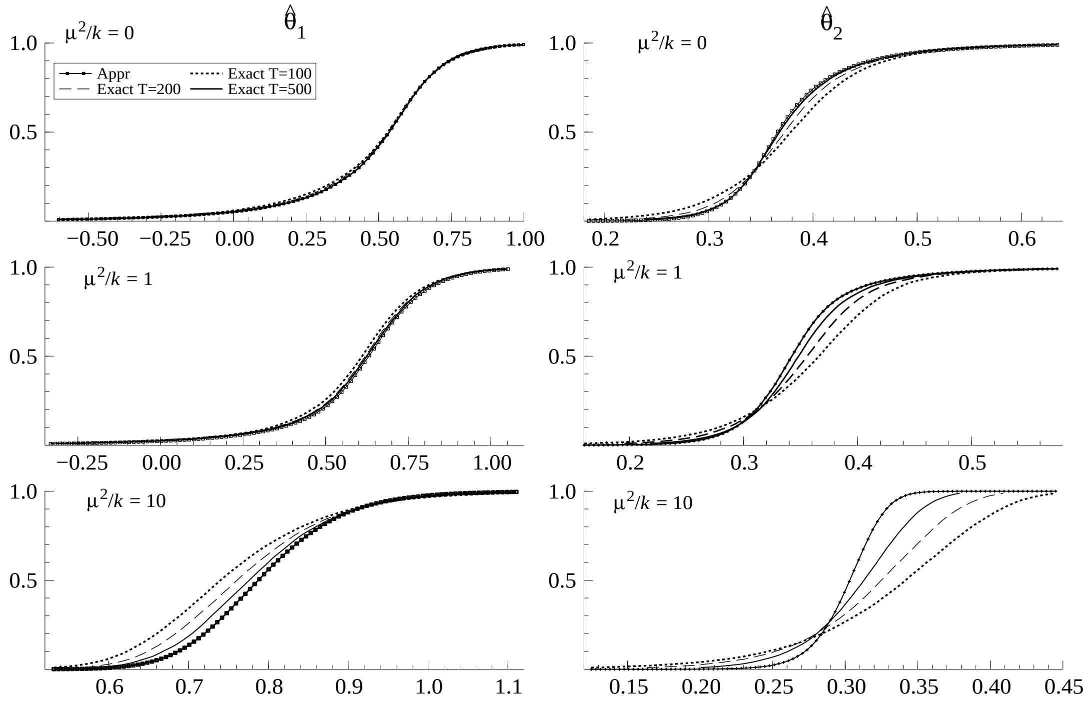
$$\begin{aligned} \mathrm{y}_{1t} &= a\, \mathrm{y}_{1,t-1} + b_T\, \mathbf{z}_{1t} + e_{1t} \\ \mathbf{z}_{1t} &= \rho\, \mathbf{z}_{1,t-1} + v_{1t} \end{aligned}$$

---

[28] $c = k/(k-1)$ is a correction factor needed to set the minimum eigenvalue of the concentration parameter to the desired level. It accounts for the fact that one of the instruments is an exogenous regressor.

where $b_T = [(k/(cT))^{1/2}\mu_{min}]$, where c is a correction factor. One of the eigenvalues of the concentration parameter is infinite, reflecting the fact that y_{t-1} is an exogenous regressor, whereas the other is approximately equal to μ_{min}. The well-identified parameter combination is $(0.25\,\theta_1 + \theta_2)$ whereas the weakly identified parameter combination is $(\theta_1 - 0.25\,\theta_2)$. Also, we set $a = 0.5$ and $\rho = 0.7$ such that the implied plims of $\widehat{\theta}_{OLS}$ are $\theta_1 = \theta_2 = 0.4$, whereas the true values are 0.83 and 0.29, respectively, and the endogeneity parameter is 0.48. Experimenting with different values of a and ρ didn't affect the qualitative results reported above.

(a) 2-step GMM (West)



(a) 2-step GMM (Newey-West)

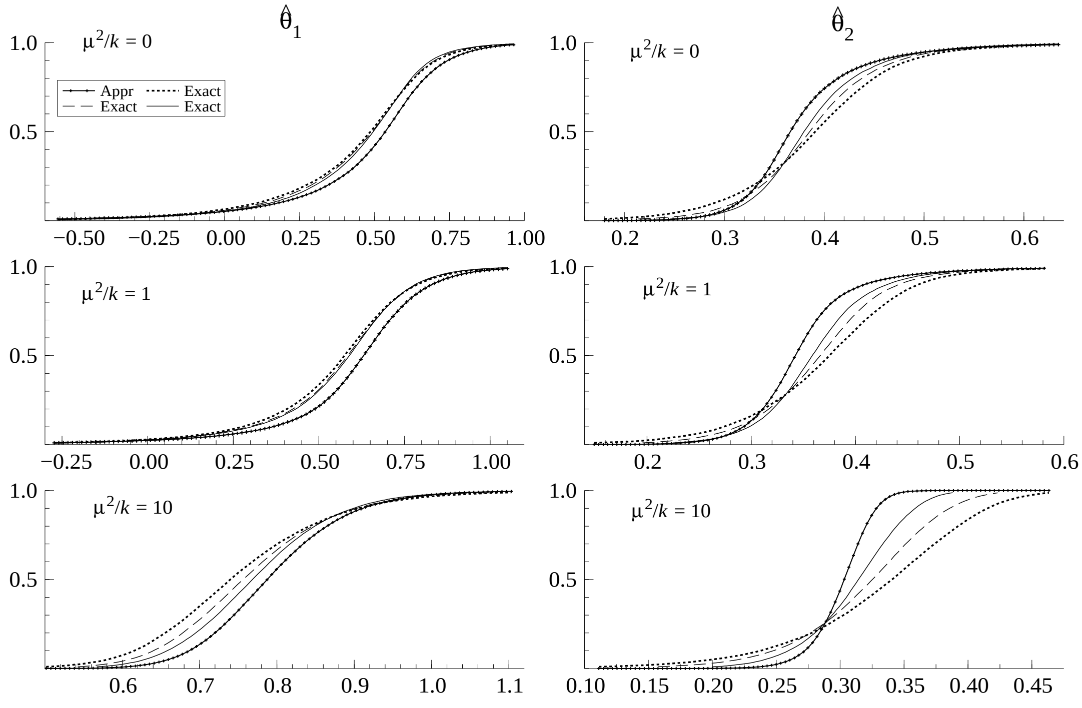


Figure 3.2: Comparison of asymptotic and finite-sample CDFs for the 2-step GMM estimators in Design IV, using West's MA-l covariance estimator (top graph), and Newey-West (bottom graph). The number of instruments k is 8.

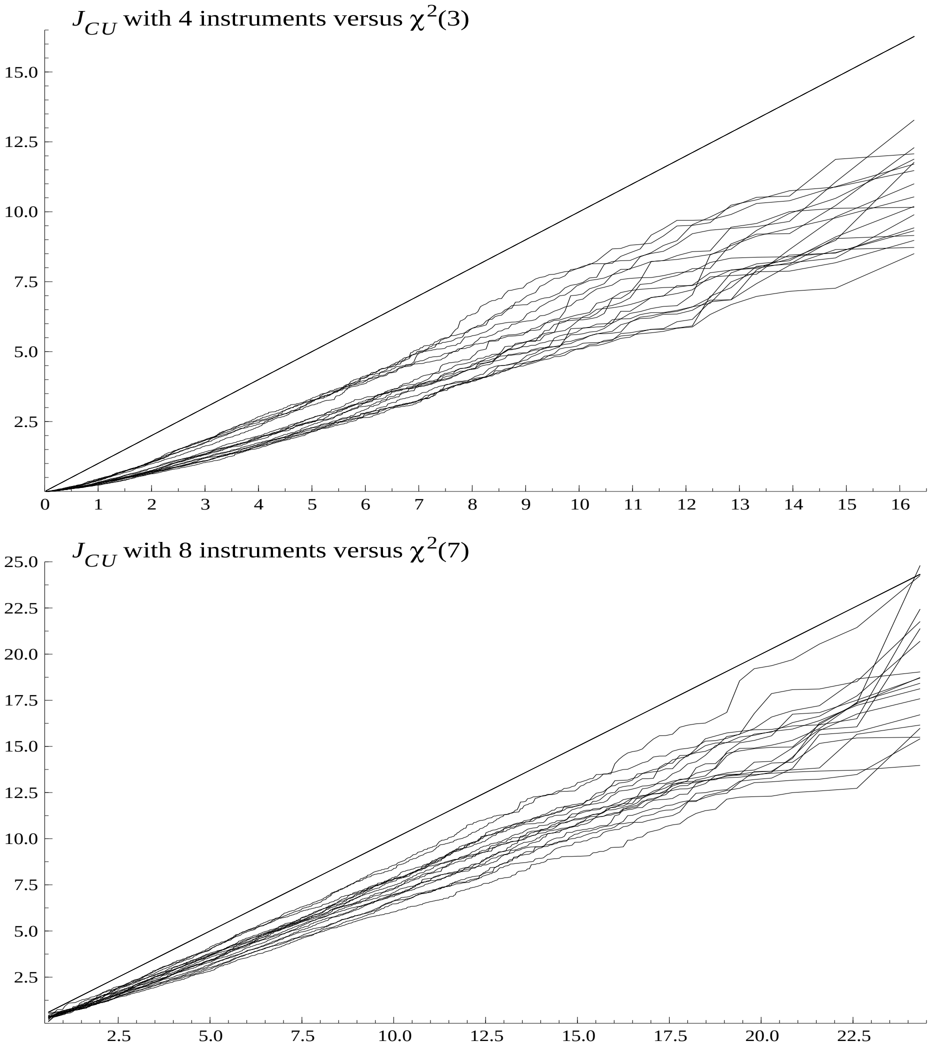


Figure 3.3: QQ plots of the J_{CU} statistic versus $\chi^2(k-1)$ for all cases in Design IV.

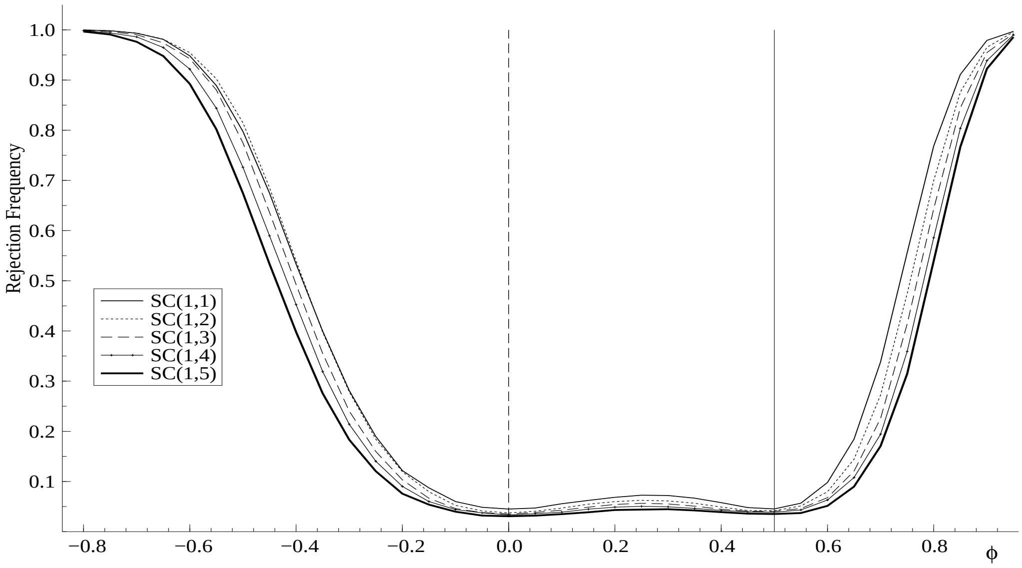


Figure 3.4: Power of excess serial-correlation test against the alternative of AR(1) serial correlation. The null hypothesis is no serial correlation between lags 2 to $q = 2, \ldots, 6$. Straight lines denote the values where the null hypothesis is true.

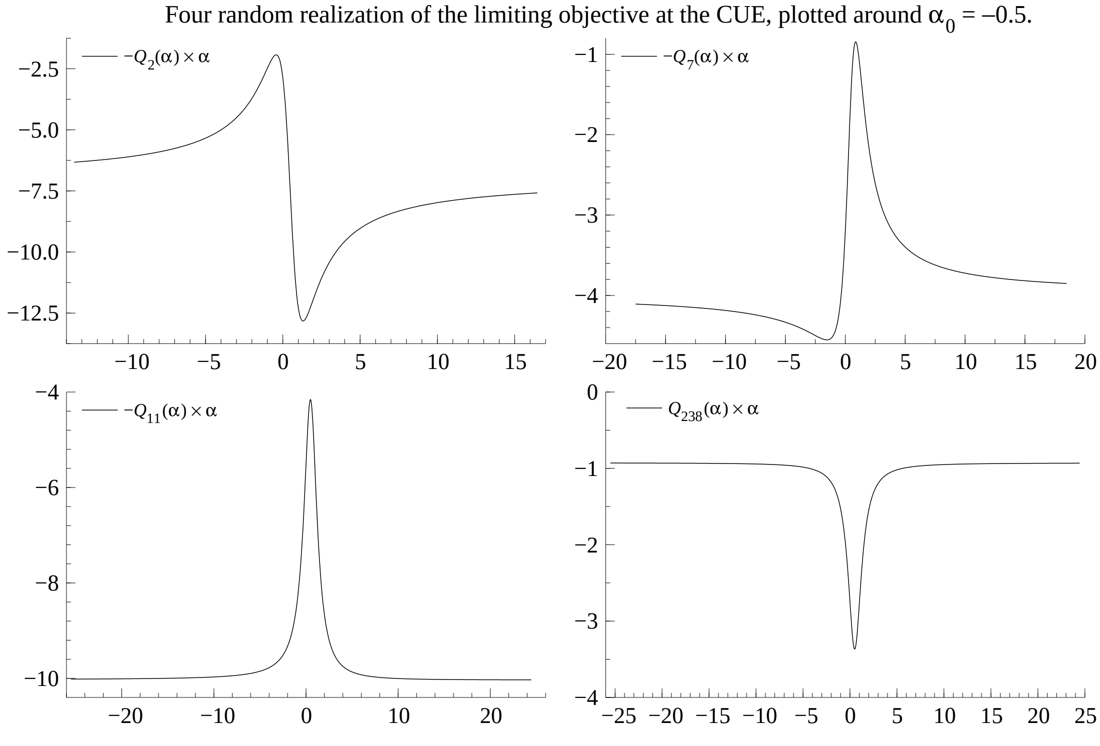
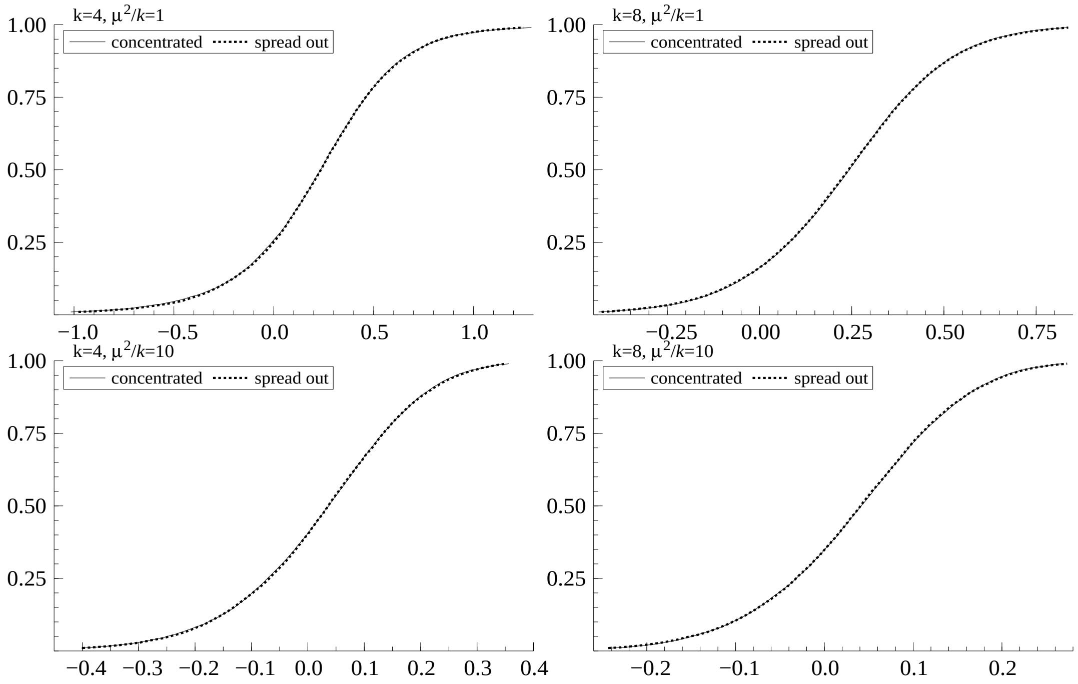


Figure 3.5: Different realizations of the concentrated objective function.

(a) 2SLS



(b) LIML

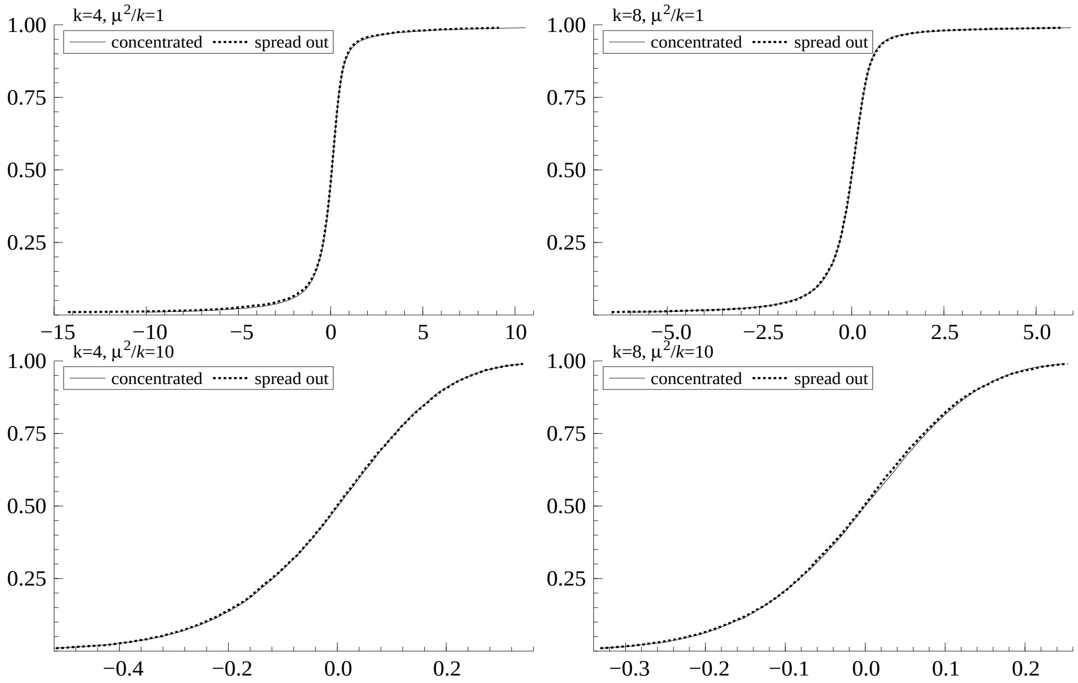


Figure 3.6: Comparison of GMM estimator distributions for concentrated versus spread out information in Design I. The sample size is 100, 20000 replications have been used, and the estimators are 2-Step and CUE. The distributions are indistinguishable.

Chapter 4

The New Keynesian Phillips Curve

In this chapter, we present a critique of the single equation GMM approach to the estimation of forward-looking models with reference to the estimation of a New Keynesian Phillips curve. Our analysis will focus on the paper by Galí and Gertler (1999) (henceforth GG). However, it is clear that the same methodology can be applied to any forward-looking structural model.

First, we will address the issue of identification of the structural parameters that has been overlooked due to the incompleteness of the single-equation formulation. In particular, we will uncover an important and empirically relevant special case in which the structural model is only *partially* identified, despite the apparent sufficiency of moment restrictions. Thus we demonstrate that the conditioning process cannot be weakly exogenous, as it carries important information about the identifiability of the parameters of interest.

The identification analysis given here is based on economic theory, in particular, the application of rational expectations. This 'structural' approach can be seen as an alternative to the pure data-based approach of pre-testing for identification in the first-stage regression. The analysis of the Phillips curve reveals that the two approaches can give contradictory results, and demonstrates the dangers of relying purely on a pre-test of identification. This situation arises when a model is identified through imposing incorrect restrictions that give rise to mis-specification.

Then, we will study the power of specification testing and the ensuing properties of GMM estimators and test statistics when the moment restrictions are invalid either (i) due to the presence of excess serial correlation, or (ii) because the restrictions implied by the assumption of rational expectations are invalid. This is motivated by the very high negative serial correlation pattern in the estimated residuals of all the models in Galí and Gertler (1999). It goes without saying that such serial correlation is enough to invalidate the estimation method (although not necessarily the model – see discussion below). Thus, it is interesting to see what biases it may impart on the estimated parameters, when it is overlooked.

Since the Hansen-Sargan test of overidentifying restrictions is the main specification test used under this methodology, it is important to study its power when the model is dynamically mis-specified, and hence, whether this test is appropriate for testing rationality of expectations. The analysis of the previous chapter reveals considerable size distortions when HAC estimator of the GMM weighting matrix are used. Here we give further evidence on the power of the test, which is decreasing sharply in the number of irrelevant instruments (over-instrumenting) and general corrections for serial correlation (over-corrections).

The above weakness in detecting mis-specification, combined with the genuine identification problems with forward-looking formulations outlined below, help explain why several different over-identified forward-looking Phillips curves can be estimated, all being apparently well-identified and correctly specified. In other words, the lack of identification of the most general forward-looking model gives rise to a large class of almost observationally equivalent models with different policy implications. This shows clearly why the debate on forward-lookingness cannot be informed using such models.

The structure of the paper is as follows. Section 4.1 introduces the hybrid Phillips Curve model of GG and describes their baseline econometric formulation. Section 4.2 discusses the issues of partial identification and weak instruments with reference to that model. Section 4.3 addresses the issue of mis-specification, and studies its implications for the Hansen-Sargan test, and the resulting biases of estimators and test statistics. Finally, section 4.4 concludes. Estimates of "nuisance parameters" and auxiliary models are given in the appendix.

4.1 The Phillips curve: New versus Old

'Phillips curve' is the name economists use to refer to an equation that describes the evolution of prices (or of the inflation rate) in a macroeconomic system. This is usually a difference equation, that involves a handful of macroeconomic variables assumed to influence price-setting behaviour.

A simple example is:

$$\pi_t = \delta x_t + \phi\pi_{t-1} + \zeta_t \tag{4.1}$$

where x_t could be, say, the output gap, or some marginal cost measure, depending on the specifics of the model, and ζ_t could be thought of as an inflation (supply) shock. This is an example of a traditional Phillips curve, or, in the words of Galí and Gertler, the "old" Phillips curve.

More recent models of inflation dynamics have emphasized forward-looking optimizing behaviour on the part of price-setters. This, combined with some form of friction in prices, has led to a new formulation of the price determination equation, the "New" Phillips curve.

A typical example of a New Phillips curve is one in which firms are monopolistically competitive, and face some constraints on price adjustment. They would optimally choose to set their prices as a constant markup over marginal costs, unless they have to keep them unchanged, which happens with probability θ.[1] This gives rise to the following equation for the determination of inflation:

$$\pi_t = \lambda mc_t + \beta \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) \tag{4.2}$$

where $\lambda = (1-\theta)(1-\beta\theta)/\theta > 0$, $\beta \in [0,1]$ is a discount factor, and mc_t denotes the percent deviation of marginal costs from their 'steady state'. Also, $\mathrm{E}(\cdot|\mathcal{F}_t)$ denotes expectations conditional on the *agents'* information set up to time t, $\mathcal{F}_t$. In other words, firms should take into account the possibility that they might be unable to set future prices optimally when setting the current prices. The operational counterpart of equation (4.2) in GG is derived by substituting real unit labour costs for mc_t (in deviation from their steady state value). Hence, the econometric formulation becomes:

$$\pi_t = \lambda s_t + \beta \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) \tag{4.3}$$

Adding Lagged Inflation: a hybrid Phillips curve

Model (4.2) is a pure forward-looking Phillips curve. Thus, it can be seen as a limiting case of a more general model that accommodates both forward- and backward-looking price-setting behaviour. Moreover, as it stands it is difficult to reconcile with the data. These considerations prompted a number of researchers to put forward a hybrid version of new and old Phillips curves, see, for example, Fuhrer and Moore (1995) and Buiter and Jewitt (1985):

$$\pi_t = \delta x_t + \gamma \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + (1-\gamma)\pi_{t-1} \tag{4.4}$$

---

[1] This is the so-called Calvo (1983) formulation.

GG proposed a new hybrid version, which is motivated by the idea of combining both forward- and backward-looking price-setting behaviour within the optimizing framework of the previous section. Firms are still monopolistically competitive, and face a constraint on price adjustment, but only a fraction of them is engaging in 'rational' intertemporal optimizing behaviour. The remaining firms use a simple adaptive rule for updating their prices. This leads to a simple generalization of (4.3), which is very similar to (4.4):[2]

$$\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1}. \tag{4.5}$$

Their model also provides a link between what they call the "reduced form" parameters $(\lambda, \gamma_f, \gamma_b)$ in (4.5) and some "deep" or structural parameters (β, θ, ω):

$$\begin{aligned} \lambda &\equiv (1-\omega)(1-\theta)(1-\beta\theta)\phi^{-1} \\ \gamma_f &\equiv \beta\theta\phi^{-1} \\ \gamma_b &\equiv \omega\phi^{-1} \\ \phi &\equiv \theta + \omega[1-\theta(1-\beta)]. \end{aligned} \tag{4.6}$$

In fact, we shall not use the same terminology as Galí and Gertler to distinguish between these two sets of parameters. Instead, we will refer to both sets as structural parameters, and we will retain the name 'reduced form parameters' for the parameters of the *solution* to the structural model.

These structural parameters have appealing interpretations: $\beta \in [0,1]$ is the discount factor; $\theta \in [0,1]$ governs the degree of price inertia ($\theta = 1$ corresponds to maximum rigidity – no price change); and $\omega \in [0,1]$ is the fraction of backward-looking firms. Another appealing feature of their model is that it nests the pure forward-looking Phillips curve (4.3), which is simply derived by setting $\omega = 0$ in (4.6).

However, their model is not well-suited to accommodate the other limiting hypothesis of the traditional old Phillips curve. Setting $\omega = 1$ (the fraction of backward-looking firms) makes inflation a pure random walk, with no feedback from any other variables. The only other way their model can nest the old curve (4.1) is by setting $\gamma_f = 0$ in (4.5). But, in terms of the structural parameters, this would imply either $\beta = 0$ or $\theta = 0$, from (4.6), both highly implausible given their interpretations.

Another disadvantage of the above specification is that the 'deep' parameters (β, θ, ω) are not *globally* identified. This follows from the fact that the mapping between the two sets of

---

[2] The only difference is the lack of a restriction on the coefficients on inflation. In that sense, the Galí-Gertler model (4.5) nests the previous hybrid version (4.4).

structural parameters in (4.6) is not invertible, i.e., there is more than one $(\beta, \theta, \omega) \in \Re^3$ for every $(\lambda, \gamma_f, \gamma_b) \in \Re^3$. This causes problems for the direct estimation of (β, θ, ω) by GMM, and hence for Monte Carlo experimentation on those estimators.

Lack of global identification is not of particular interest, because whenever a discrete number of optima is identified, economic theory will usually be informative in selecting a unique parameterization, as is the case here.[3] However, this was not realized by Galí and Gertler (1999), whose optimization procedure does not take into account the lack of global identification.

For the aforementioned reasons, in this chapter we focus on the parameterization $(\lambda, \gamma_f, \gamma_b)$, and we analyze the problems associated with lack of *local* identification.

Estimating the model by GMM

In our analysis, we will use a slightly generalized (and more realistic) version of the Galí-Gertler hybrid model (4.5), namely:

$$\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t \tag{4.7}$$

where $\epsilon_t \sim NID(0, \sigma_\epsilon^2)$ is adapted to $\mathcal{F}_t$, and it is an innovation with respect to $\mathcal{F}_{t-1}$.[4] The reason for this generalization is that using equation (4.5) would imply that the joint distribution of π_t and s_t is singular, with all the stochastic variation being driven by the latter. That restriction, namely $\sigma_\epsilon = 0$, is unnecessary for the validity of the model and need not be imposed. Another reason for retaining an additional (latent) variable in the Phillips curve is that it will enable us to discuss the implications of omitted dynamics for the properties of the estimators.

The above model cannot be estimated directly due to the fact that $\mathrm{E}(\pi_{t+1}|\mathcal{F}_t)$ is a latent variable. Therefore, we replace this expectation with the actual future realization in (4.7) in order to derive the GMM *estimating equation*:

$$\pi_t = \lambda s_t + \gamma_f \pi_{t+1} + \gamma_b \pi_{t-1} + e_t \tag{4.8}$$

where the 'structural residual' e_t is given by:

$$e_t = \epsilon_t - \gamma_f \eta_{t+1} \tag{4.9}$$

---

[3] It can be shown that there are three solutions for (β, θ, ω) in terms of $(\lambda, \gamma_f, \gamma_b)$ in equation (4.6). So even if the latter are globally identified, the former cannot be. Furthermore, the resulting deep parameters are not guaranteed to be real for all possible values of $(\lambda, \gamma_f, \gamma_b)$. However, if we restrict attention to the unit cube, then it is possible to show there is a unique real solution to the GMM program. Thus, the economically motivated restriction of $(\beta, \theta, \omega) \in [0, 1]^3$ will suffice to identify the parameters.

[4] In Rational Expectations terminology, s_t, ϵ_t are the forcing variables, the latter being unobserved, and π_t is the endogenous or decision variable, see the exposition in appendix A.

where $\eta_{t+1} \equiv \pi_{t+1} - \pi_{t+1|t}$ is the forecast error in predicting future inflation, and hence a mean innovation process with respect to information up to time t, $\mathcal{F}_t$. This process can be explicitly derived, given a solution to the model. For example, define the 1-step forecast errors $v_{t+1} = s_{t+1} - \mathrm{E}(s_{t+1}|\mathcal{F}_t)$ and $u_{t+1} = \pi_{t+1} - \mathrm{E}(\pi_{t+1}|\mathcal{F}_t)$. Then, under the solution (4.17) that we derive below, $u_t = \frac{1}{1-\gamma_f \delta}\epsilon_t$ and:

$$\begin{aligned} \eta_{t+1} &\equiv \pi_{t+1} - \pi_{t+1|t} = \alpha s_{t+1} + \delta \pi_t + u_{t+1} - \alpha \mathrm{E}(s_{t+1}|\mathcal{F}_t) - \delta \pi_t \\ &= \alpha v_{t+1} + u_{t+1} \end{aligned} \tag{4.10}$$

Hence, the structural residual e_t becomes

$$e_t = \epsilon_t - \gamma_f \alpha v_{t+1} - \gamma_f u_{t+1} = (1 - \gamma_f \delta) u_t - \gamma_f \alpha v_{t+1} - \gamma_f u_{t+1}.$$

We note that the IV residual may exhibit negative serial correlation at lag 1 (if $\sigma_\epsilon^2 \neq 0$), without invalidating the model. In other words, the model could accommodate up to first-order negative serial correlation in the estimated residuals.[5]

Digression: Residual Serial Correlation The serial correlation of the GMM residual is not specific to the driving process for inflation, but is intrinsic in any forward-looking rational expectations model. Since this point is important, it is worth elaborating upon briefly.[6]

Suppose all the variables involved in the analysis (regressors and instruments) can be adequately characterized in companion form by a VAR(1), (approximate local DGP):

$$X_t = A\, X_{t-1} + U_t$$

where the endogenous regressor is $y_t = X_{1,t}$. Then,

$$X_{t+1} = A^2\, X_{t-1} + U_{t+1} + A\, U_t$$

and letting $a = A'_{1\cdot}$, $\tilde{a} = \left(A^2\right)'_{1\cdot}$ denote (the transpose of) the first row of those matrices, respectively,

$$y_{t+1} = \tilde{a}'\, X_{t-1} + u_{t+1} + a'\, U_t$$

The GMM regression is:

$$y_t = \beta\, y_{t+1} + (\underbrace{a - \beta\, \tilde{a}}_{\gamma})'\, X_{t-1} + \underbrace{u_t - \beta\, u_{t+1} - \beta\, a'\, U_t}_{e_t}$$

---

[5] This is not, strictly speaking, the case in the original Galí-Gertler model, where $\sigma_\epsilon^2 = 0$. However, the authors are concerned there with 12th-order serial correlation, which is incompatible with their model.

[6] For instance, Fuhrer and Moore (1995) estimate a forward-looking model of inflation by Maximum Likelihood and report high and significant negative serial correlation, but appear unable to interpret it.

with identifying restrictions on γ. Thus, we note that, even when $U_t \sim iid$, e_t exhibits 1st order serial correlation in general, by construction. That serial correlation is negative whenever $\beta > 0$.

Going back to the GG model, equation (4.8) is an IV regression, with valid moment conditions of the form:

$$\mathrm{E}\left[(\pi_t - \lambda s_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1})\mathbf{z}_t\right] = 0 \tag{4.11}$$

for any $\mathbf{z}_t \in \mathcal{F}_t \backslash \{\pi_t\}$.[7] It appears that the parameters of interest $(\lambda, \gamma_f, \gamma_b)$ could be estimated by IV, or any other multi-step GMM procedure to take account of the serial correlation (and any potential heteroscedasticity) in the residuals, e_t. GG use a 2-step-2SLS estimator with a 12-lag Newey-West estimate of the covariance matrix. However, any serial correlation beyond lag 1 is sufficient to invalidate their methodology.

Data

The data is from the US economy during the sample period 1960:Q1 to 1997:Q4, see appendix 4.A.1 for details. Inflation, measured as the quarterly growth rate of the US GDP deflator, and the labour share in deviation from its steady state are plotted in figure 4.1. Both variables seem

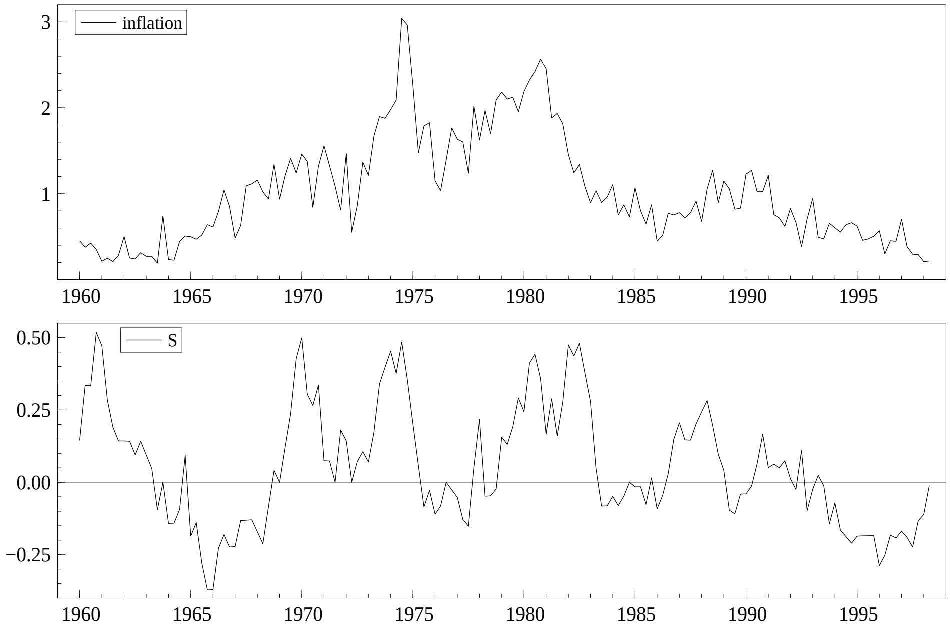


Figure 4.1: Inflation and the labour share.

---

[7] π_t cannot be an admissible instrument for model (4.7), except in the special case $\sigma_\epsilon = 0$, since π_t is not orthogonal to ϵ_t, by construction.

highly positively autocorrelated, and resemble a random walk. The empirical analysis below shows that the dynamic specification for both variables has roots near unity: s_t is best described by an AR(1) model with a coefficient of 0.9, whereas a simple random walk on inflation has a standard error of 28 basis points, which is not far from the standard error of the reduced form model based on the complete system (22 basis points), see equation (4.A.1) in the appendix.

4.2 Identification analysis

As we saw in chapter 2, the identification analysis of the structural model requires knowledge of the reduced form of the system. This can be estimated directly from the data, or it can be derived from the structural model by postulating a distribution for the forcing variables and solving the system. Here, we use the latter approach for two reasons. First, simple structural equations such as the New Keynesian Phillips curve usually have considerable intuitive appeal, and this grants them some degree of immunity to criticisms of mis-specification. Nevertheless, even in such cases, it is important to examine whether the equation of interest is identified and hence empirically estimable, and whether inference based on it is reliable. Secondly, by assuming that the structural model is correctly specified we can separate the analysis of identification from mis-specification issues, which are usually conflated. Thus, we can focus entirely on detecting pathological situations in which a correctly specified model becomes weakly identified.

To discuss identification, we need to 'complete' the characterization of the structural model under analysis, by postulating a distribution for the non-modelled variable, s_t. We will consider alternative distributions for s_t and discuss their implications for the identification of the parameters of interest, $(\lambda, \gamma_f, \gamma_b)$ or (β, θ, ω).

We shall consider the following leading cases:

$$\text{Case 1}: \quad s_t = \rho s_{t-1} + v_t \tag{4.12}$$

$$\text{Case 2}: \quad s_t = \rho s_{t-1} + \varphi \pi_{t-1} + v_t \tag{4.13}$$

$$\text{Case 3}: \quad s_t = \rho_1 s_{t-1} + \rho_2 s_{t-2} + v_t \tag{4.14}$$

where $v_t \sim NID(0, \sigma_v^2)$ is an innovation w.r.t. $\mathcal{F}_{t-1}$ and it is orthogonal to ϵ_t in (4.7).[8] We also need to specify the information set $\mathcal{F}_t$. In principle, this set contains all of the information that is available to the agents at time t, which is usually more than a handful of macroeconomic variables

---

[8] Normality is assumed for the purposes of simulation.

that the econometrician may have at their disposal. Therefore, this set is non-decreasing and contains at least current and past values of the endogenous variable π_t, and the forcing variables (s_t, ϵ_t):[9]

$$\mathcal{F}_t = (\pi_t, \pi_{t-1}, \ldots; s_t, s_{t-1}, \ldots; \epsilon_t, \epsilon_{t-1}, \ldots; \ldots).$$

Case 1 is the benchmark under-identified case. It is a typical example of the simplest possible dynamic, stationary specification for an exogenous variable. The second and third cases are straightforward extensions of the first one, which introduce richer dynamics in the exogenous process. Case 2 introduces Granger causality of π_t for s_t. As is shown below, this extension is not sufficient to yield identification.

In the third case, identification is achieved, but the 'strength' of the instruments crucially depends on the size of the second-order dynamics.[10] Also, since it nests the unidentified case, case 3 is suitable as a basis for discussion of weak instruments or lack of empirical identification. That is, we can study the behaviour of the estimators when the key identification parameter, ρ_2, becomes insignificant.

The completed model will consist of a pair of equations, one for inflation and the other for the conditioning variable, labour share. In other words, it will contain equation (4.7) together with one of (4.12) - (4.14).

4.2.1 Lack of identification

Case 1: $s_t \sim AR(1)$

The complete structural model under the leading case 1 would be:

$$\begin{aligned} \pi_t &= \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t \\ s_t &= \rho s_{t-1} + v_t, \qquad \epsilon_t \perp v_t \end{aligned} \tag{4.15}$$

The total structural parameters are:

$$\boldsymbol{\theta} = (\lambda, \gamma_f, \gamma_b, \sigma_\epsilon^2, \rho, \sigma_v^2)$$

---

[9] GG use additional variables as instruments, as we discuss in section 4.2.3 below. However, whether this extra information is *relevant* (in predicting future inflation or the labour share) depends on the local DGP for (π_t, s_t), which has not been specified yet. Given the above 3 specifications of the completing process for s_t, we will show below that the implied solution for inflation is such that no other variable should be relevant in predicting future inflation.

[10] This point goes back to Abel and Mishkin (1983), but appears to have been overlooked in the recent monetary economics literature. However, it is worth re-emphasizing it here.

of which $\boldsymbol{\theta}_1 = (\lambda, \gamma_f, \gamma_b, \sigma^2_\epsilon)$ are the *parameters of interest* and $\boldsymbol{\theta}_2 = (\rho, \sigma^2_v)$ are the *nuisance parameters*.

Assuming model-consistent expectations, the solution to this model will depend on the roots μ_i of the polynomial

$$\gamma(L) = 1 - \gamma_f L^{-1} - \gamma_b L \tag{4.16}$$

A necessary condition for the existence of a solution is that at most one of those roots is explosive, i.e., at least μ_1, say, is *not* explosive $|\mu_1| \geq 1$. When there is exactly one explosive root ($|\mu_2| < 1$), the solution is *unique*.[11] These conditions in turn impose restrictions on the set of admissible structural parameters $(\gamma_f, \gamma_b) \in (\Gamma_f, \Gamma_b)$. There are, of course, several different possibilities, but we shall only consider the range of values implicit in the estimates from Galí and Gertler, which are also plausible in many other contexts, namely:

$$\gamma_b, \gamma_f \geq 0 \quad \text{and} \quad \gamma_b + \gamma_f \leq 1$$

Both conditions are implied by the underlying model of GG, given equation (4.6) and the requirement that the 'deep' parameters (β, θ, ω) lie between zero and 1. These restrictions conveniently ensure (i) that the roots are real (and hence, rules out complex solutions), (ii) that a solution always exists, and, (iii) when the inequality is strict, that it is also unique.[12]

The possibility of non-uniqueness $\gamma_f + \gamma_b = 1$ is dealt in section 4.2.1 below. Hence, we can focus only on the unique solution to the structural model (4.15) which will be of the form:

$$\begin{aligned} &\pi_t = \alpha s_t + \delta \pi_{t-1} + u_t \\ &\alpha = \tfrac{\lambda}{1-\gamma_f(\delta+\rho)}, \qquad \delta = \tfrac{1-\sqrt{1-4\gamma_f\gamma_b}}{2\gamma_f}, \qquad u_t = \tfrac{1}{1-\gamma_f\delta}\epsilon_t \end{aligned} \tag{4.17}$$

with $\boldsymbol{\phi} = (\alpha, \delta, \sigma^2_u, \rho, \sigma^2_v)$ being the reduced-form parameters. The solution is the mapping from $\boldsymbol{\theta}$ to $\boldsymbol{\phi}$ and the latter characterize completely the DGP. Clearly, when no further restrictions are imposed on $\boldsymbol{\theta}$, it is of higher dimension than $\boldsymbol{\phi}$, so the mapping is not invertible and there is a multiplicity of $\boldsymbol{\theta}$s that correspond to the same DGP. Hence, the structural parameters cannot be jointly determined given knowledge of the reduced form, and the structural model (4.15) is *under*-identified. Since (ρ, σ^2_v) are in common, the under-identification affects $(\lambda, \gamma_f, \gamma_b)$ and σ^2_ϵ.

---

[11] Conditions for existence and uniqueness of solutions to rational expectations models are discussed in appendix A at the end of the thesis, see proposition A.1.

[12] The two roots are:

$$\mu_{1,2} = \frac{1 \pm \sqrt{1 - 4\gamma_f\gamma_b}}{2\gamma_b}$$

so that the discriminant $\Delta = 1-4\gamma_f\gamma_b \geq (1-2\gamma_b)^2$ is always positive. Hence $1-2\gamma_b+\sqrt{\Delta} \geq 1-2\gamma_b+\sqrt{(1-2\gamma_b)^2} = 0$ and $1 - 2\gamma_b - \sqrt{\Delta} \leq 1 - 2\gamma_b - \sqrt{(1-2\gamma_b)^2} = 0$, showing that $\mu_1 \geq 1$ and $\mu_2 \leq 1$. The last inequality is strict when $\gamma_f + \gamma_b < 1$, yielding a unique solution to the RE model.

To see this differently, let us derive from the DGP (4.17) the conditional model for π_{t+1} given the variables that can be used as instruments, namely, $(s_t, s_{t-1}, \ldots, \pi_{t-1}, \pi_{t-2}, \ldots)$. This is essentially the first-stage regression of the endogenous regressor π_{t+1} on the instruments:

$$\pi_{t+1} = \alpha(\rho + \delta)\, s_t + \delta^2\, \pi_{t-1} + u_{t+1} + \alpha\, v_{t+1} + \delta\, u_t. \tag{4.18}$$

Now, add and subtract $\gamma_f\, \pi_{t+1}$ to the reduced form of inflation (4.17), using (4.18), to get:

$$\pi_t = \gamma_f\, \pi_{t+1} + \underbrace{\alpha[1 - (\delta + \rho)\gamma_f]}_{\lambda}\, s_t + \underbrace{\delta(1 - \gamma_f\, \delta)}_{\gamma_b}\, \pi_{t-1} + \underbrace{(1 - \delta\, \gamma_f)\, u_t - \gamma_f\, u_{t+1} - \gamma_f\, \alpha\, v_{t+1}}_{e_t}. \tag{4.19}$$

Viewing this as a GMM regression model with known (α, δ, ρ), clearly $(\lambda, \gamma_f, \gamma_b)$ are not jointly estimable from (4.19) with instruments $(s_t, s_{t-1}, \ldots, \pi_{t-1}, \pi_{t-2}, \ldots)$. The forward-looking parameter is un-identified because there are no relevant additional instruments in the first stage regression (4.18) (beyond s_t and π_{t-1} which are already used as exogenous regressors) available to estimate it. Also, the other two parameters are functions involving the un-identifiable parameter γ_f, so they, too, will be un-identified. The degree of under-identification is only 1, but it spills over to all of the parameters of interest. This is another example of the generalized assumption D in chapter 2, whereby only some *combinations* of the parameters are well-identified.

In general, to the extent that the reduced form model (solution) is nested within the structural model (as is the case here), there is trivially more than one observationally equivalent (OE) parameterization of the structural model.[13] Therefore, the issue here cannot be how to distinguish between a forward- and a backward-looking specification, which is impossible in this setting.[14]

Distributions of estimators under partial identification

The theoretical results of chapter 2 as well as the Monte Carlo evidence in chapter 3 suggest that all of the GMM estimators in this case will exhibit a 'double inconsistency', being both $O_p(1)$, and exhibiting a large bias in the direction of OLS. Here we give some extra Monte Carlo evidence, tailored to the specifics of the GG model.

The Monte Carlo setting requires the specification of values for all the model's parameters. The parameters of interest are set to the values reported by Galí and Gertler (1999, Table 2),

---

[13] That is, setting $\gamma_f = 0$ in (4.15) yields (4.17), or simply, the solution to a backward-looking model is the model itself.

[14] This may become testable via tests for super-exogeneity, when we allow for the possibility of breaks in the parameters, see Hendry (1988) and Ericsson and Hendry (1999). We will not be concerned with this possibility here.

reproduced in table 4.1.[15] The first unrestricted parameterization is the one used for all the simulation experiments reported below. The restricted parameterization relates to the analysis of identifying restrictions that we discuss in this section. The remaining nuisance parameters $(\sigma_\epsilon, \rho, \sigma_v)$ are estimated from the GG data, see appendix 4.A.2 for details.

Table 4.1: Parameter values for simulation (i) and discussion of identifying restrictions (ii).

	ω	θ	β	γ_b	γ_f	λ	σ_ϵ	ρ	σ_v
(i)	0.486	0.834	0.909	0.378	0.591	0.015	0.18	0.9	0.10
(ii)	0.522	0.838	1.000	0.384	0.616	0.009	-	-	-

The different parameterizations are taken from Galí and Gertler (1999, table 2).

Next, we consider a 2-step GMM estimator of $(\lambda, \gamma_f, \gamma_b)$. This is essentially the 2S-2SLS of Cumby, Huizinga, and Obstfeld (1983), using either a 1-lag (under the Null) or a 12-lag (following GG) HAC estimator of the variance matrix of the moment conditions.[16] The instrument set includes four lags of π_t and s_t, and a constant is included in the estimated equation. In order to replicate the GG analysis exactly, we are not using s_t as an instrument, which essentially means treating s_t as endogenous.[17]

We simulate both the GMM estimators and the respective OLS estimators of the parameters of interest $(\gamma_b, \gamma_f, \lambda)$. The results are given in table 4.2, and figure 4.2.

Discussion of results The following observations stand out from the reported results. First, all the parameter estimates appear to be inconsistent, as anticipated. λ and γ_b are biased upwards relative to the 'true' values used for the simulation, whereas the forward coefficient γ_f is by far the most affected parameter exhibiting a severe downward bias. More importantly, the coefficient estimators exhibit a bias that is almost identical to the OLS bias, in line with the exact distribution

---

[15] These estimates are based on a US GDP deflator-based measure of inflation, which is GG's preferred measure and the one we used for our data analysis later on. They also reported alternative estimates using the non-farm business price deflator or a different normalization of the orthogonality conditions for sensitivity analysis of their estimates, and they found them broadly in line with the ones we show in table 4.1 here.

[16] Similarly to GG, we use the HAC estimator proposed by Newey and West (1987), denoted HAC(i), where i stands for the lag truncation parameter.

[17] Strictly speaking, this is equivalent to estimating:

$$\pi_t = \lambda \mathrm{E}(s_t|\mathcal{F}_{t-1}) + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_{t-1}) + \gamma_b \pi_{t-1} + \varepsilon_t$$

where s_t is endogenous. This is justified by an 'error-in-variables' interpretation of s_t, when the latter is a *proxy* for the true relationship being driven by marginal costs, which is thus measured with 'error'. Note also that the above model is implied by (4.7), since it entails a weaker condition for the process of π_t. In other words, it is a weaker model that does not pin down the contemporaneous correlation between s_t and π_t.

Table 4.2: Monte Carlo experiment of GMM estimators of $(\lambda, \gamma_f, \gamma_b)$ in (4.15) under $s_t \sim AR(1)$.

	$\boldsymbol{\gamma_b}$			$\boldsymbol{\gamma_f}$			$\boldsymbol{\lambda}$		
Sample	mean	st. dev.	bias	mean	st. dev.	bias	mean	st. dev.	bias
50	0.409	0.118	0.031	0.428	0.335	-0.163	0.053	0.273	0.038
100	0.423	0.102	0.045	0.426	0.319	-0.165	0.046	0.150	0.031
150	0.428	0.100	0.050	0.420	0.319	-0.171	0.045	0.113	0.030
300	0.430	0.094	0.052	0.423	0.302	-0.168	0.044	0.078	0.029
500	0.433	0.096	0.055	0.419	0.305	-0.172	0.045	0.069	0.030
1000	0.433	0.095	0.055	0.419	0.299	-0.172	0.044	0.060	0.029

The MCSE are (decreasing in T) γ_b: 0.0012 - 0.0009, γ_f: 0.0033 - 0.0030, λ: 0.0027 - 0.0006.
Parameter values for simulation: table 4.1, row (i).
plims of OLS estimators: $\gamma_{b,ols} = 0.43$, $\gamma_{f,ols} = 0.43$, $\lambda_{ols} = 0.042$.

results for the exogenous instruments case.

Second, the variance of the estimator distributions falls very slowly with the sample size (compared to OLS), demonstrating that the estimators are of order $O_p(1)$ rather than $O_p(T^{-1/2})$. This is certainly true for the weakly identified parameter γ_f but also affects γ_b, the coefficient of the exogenous regressor π_{t-1}. Even though, the distribution of $\widehat{\lambda}$ seems to be imploding, the latter converges to the wrong value, the plim of the OLS estimator. Moreover, this implosion is misleading, since $\widehat{\lambda} = O_p(1)$ (indeed, increasing the sample $T > 1000$ doesn't reduce the variance of the estimator) and it simply indicates the prevalence of higher-order terms in small samples, as the analysis of chapter 3 suggested.

All of these results are consistent with the local-to-zero asymptotic approximations that are presented in chapter 2. Both the theoretical results given there and the simulation results here show that partial identification of the parameters leads to *all* of the estimators being $O_p(1)$. Specifically, in this model there are three structural parameters to be estimated but only two relevant instruments, s_{t-1} and π_{t-1}. This means that two *functions* of the parameters are consistently estimable, but none of these functions corresponds exactly to γ_f, γ_b or λ. Therefore, this partial lack of identification spills over to all of the estimated parameters.

Potentially identifying restrictions

Next, we discuss a number of theoretically motivated restrictions in terms of their implications for identification, and the extent to which they are testable.

The first restriction of particular interest is the following:

$$\mathrm{H}_\beta : \beta = 1 \quad \Leftrightarrow \quad \gamma_f + \gamma_b = 1 \tag{4.20}$$

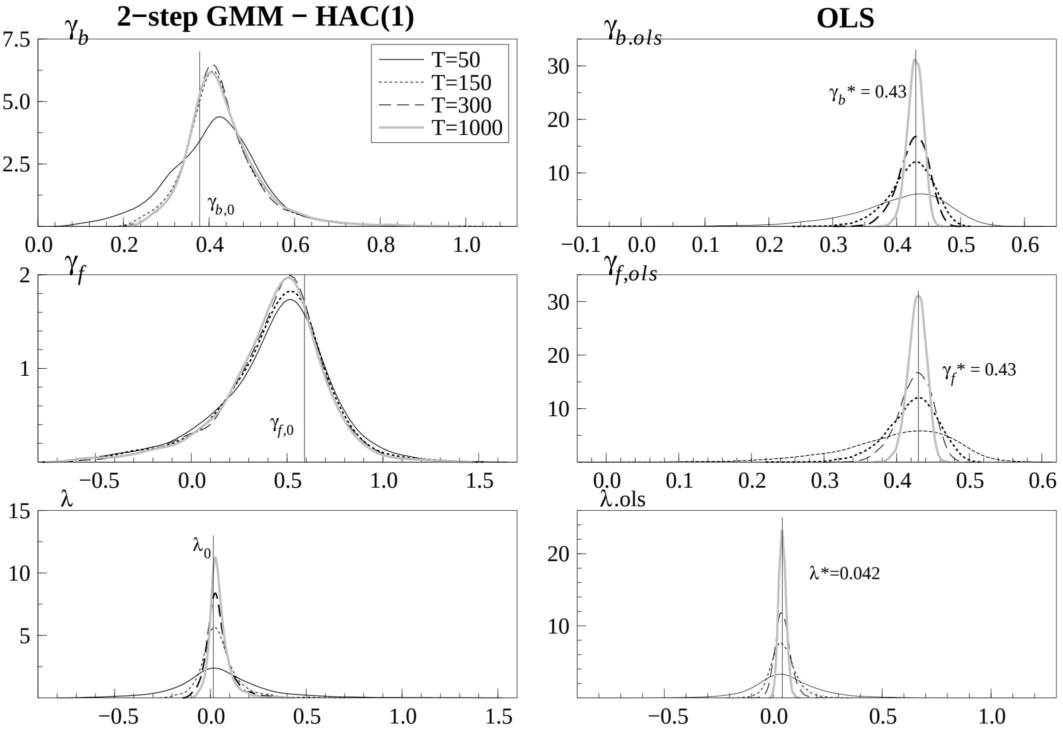


Figure 4.2: Simulated distributions of GMM and OLS estimators.

Galí and Gertler emphasise this case because it links their model to the prototype hybrid Phillips curve of Buiter and Jewitt (1985). Also, since it reduces the number of structural parameters, we are interested in knowing whether it is sufficient for identification.

Imposing restriction (4.20) on (4.15) yields (letting $\gamma = \gamma_f$):

$$\begin{aligned} \pi_t &= \lambda s_t + \gamma \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + (1-\gamma)\pi_{t-1} + \epsilon_t \\ s_t &= \rho s_{t-1} + v_t. \end{aligned} \tag{4.21}$$

The analysis of section 4.2.1 showed that this restriction, together with $\gamma > 0$, guarantees existence of a solution of the form (4.17), but may also imply non-uniqueness. The lag polynomial (4.16) together with its roots now simplifies to:

$$\gamma(L) = 1 - \gamma L^{-1} - (1-\gamma)L, \qquad \text{with roots } \mu_1 = 1 \text{ and } \mu_2 = \frac{\gamma}{1-\gamma} \tag{4.22}$$

As far as uniqueness is concerned, we distinguish the following two interesting cases: $\gamma < 1/2$ and $\gamma \geq 1/2$.

Case $\gamma < \frac{1}{2}$: This implies $\mu_2 < 1$, and hence the solution is unique. The economic interpretation of this would be that the fraction of forward-looking agents-firms is smaller than the backward-looking ones. Or alternatively, partial adjustment (friction) dominates forward-looking behaviour.

The implied solution is:

$$\pi_t = \alpha s_t + \delta \pi_{t-1} + u_t \quad \text{with} \quad \alpha = \frac{\lambda}{1-\gamma-\gamma\rho}, \quad \delta = 1, \quad u_t = \frac{\epsilon_t}{1-\gamma} \tag{4.23}$$

Note that the coefficient on lagged inflation in the reduced form is 1, and independent of the structural parameters. This immediately implies that inflation has a unit root, which makes the restriction testable upon estimating the system (or the reduced form for inflation and performing a unit-root test). However, the structural model is still *under*-identified, since there are now 5 structural parameters $(\lambda, \gamma, \sigma_\epsilon, \rho, \sigma_v)$ and only 4 reduced form parameters $(\alpha, \sigma_u, \rho, \sigma_v)$.

Case $\gamma \geq \frac{1}{2}$: This is the case that best fits the GG data (see bottom line of table 4.1). This time the solution is non-unique, since none of the roots of the $\gamma(L)$ polynomial are explosive, and there is one non-predetermined variable. However, we can still solve the system forward using the unitary root of the (4.22) polynomial, since $|\rho| < 1$.[18] That solution would again be of the form (4.17)

$$\pi_t = \alpha s_t + \delta \pi_{t-1} + u_t \quad \text{with} \quad \alpha = \frac{\lambda}{\gamma(1-\rho)}, \quad \delta = \frac{1-\gamma}{\gamma}, \quad u_t = \frac{\epsilon_t}{\gamma} \tag{4.24}$$

Now the structural parameters become just-identified, although the strength of the identifying restriction remains an issue. To investigate it, consider the following reparameterization of the structural equation (4.21):

$$\Delta \pi_t = \lambda s_t + \gamma \Delta_2 \pi_{t+1|t} + \epsilon_t \tag{4.25}$$

where $\Delta_2 = 1 - L^2$. Obviously, the strength of identification depends on the correlation between $\Delta_2 \pi_{t+1}$ and the extra instrument π_{t-1}. To study this, we derive the forecasting equation for $\Delta_2 \pi_{t+1}$ given s_t and $(t-1)$-dated information, based on the solution (4.17):

$$\Delta_2 \pi_{t+1|t} = \alpha(\rho + \delta) s_t + (\delta^2 - 1)\pi_{t-1}$$

Clearly, the identifying power of π_{t-1} for γ depends on δ. As δ gets closer to 1 (equivalently, γ gets closer to 1/2) identification weakens.

As we emphasized earlier, the under-identification of the unrestricted model implies that there is a *set* of values $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ for the structural parameters $\boldsymbol{\theta}$ that are consistent with the same reduced form (DGP), $\boldsymbol{\phi}_0$. Conceptually, only one of those values, say $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, will be 'true' in the sense

---

[18] Note that the usual 'no bubbles' transversality condition is satisfied, by virtue of the fact that the resulting process for inflation will be stationary.

that it corresponds to the true underlying behavioural relationship. However, we have no way of distinguishing between $\boldsymbol{\theta}_0$ and other values in $\boldsymbol{\Theta}_0$ in a statistical sense. Given that the degree of under-identification here is 1, imposing just one identifying restriction will yield identification (provided it is informative), and will thus ensure that the restricted estimators converge to a particular point $\boldsymbol{\theta}_0^R \in \boldsymbol{\Theta}_0$. However, in general, $\boldsymbol{\theta}_0^R$ need not correspond to the 'true' $\boldsymbol{\theta}_0$.

To illustrate the above point, we simulate the data using the parameter values of the unrestricted specification (top line of table 4.1) rather than the restricted one (bottom line of table 4.1), since in the latter case $\boldsymbol{\theta}_0^R = \boldsymbol{\theta}_0$. To confirm that the restriction (4.20) is identifying, we derive the concentration parameter which is $1.94 \times T$, revealing strong identification.[19] The mean, standard deviation and mean bias of the restricted (2-step) estimators of (γ, λ) are estimated by simulation and are reported in table 4.3. The implied probability limits ($\boldsymbol{\theta}_0^R$) of the restricted GMM estimators can be found using the second moments of the data (which are known given the reduced form) as explained in appendix 3.B, see equation (3.B.16). Alternatively, we can derive the implied parameter values using equation (4.19) together with the restriction $\gamma_b = 1 - \gamma_f = 1 - \gamma$:

$$1 - \gamma = \delta - \delta^2 \gamma \quad \Rightarrow \quad \gamma = \frac{1}{1+\delta} \quad \text{and} \quad \lambda = \alpha\,[1 - (\delta + \rho)\gamma].$$

Either way, given the values $(\alpha_0 = 0.11, \delta_0 = 0.57, \rho_0 = 0.9)$ for the reduced form parameters,[20] we obtain $\gamma_0^R = 0.634$ and $\lambda_0^R = 0.007$.

The results reported in table 4.2 reveal convergence of the restricted GMM estimators to those implied true values. Since these are not the values that we used to simulate the data, the identifying restriction leads to inconsistency as we argued above. In this case, the imposition of the restriction tends to raise the forward-looking parameter (in line with the actual evidence reported by Galí and Gertler, compare the top and bottom lines in table 4.1). Finally, the fact that the restriction is just-identifying is also demonstrated by the mis-specification parameter being exactly zero.[21]

This was not an exhaustive account of what happens in the case $\gamma \geq 1/2$. As we pointed out at the beginning, the solution is non-unique and the generic 'backward' solution to the structural

---

[19] See appendix 3.B for details on the computation of concentration and mis-specification parameters.

[20] These values are derived from the solution to the model (4.17) substituting in the values from the top line table 4.1 for $\gamma_f, \gamma_b, \lambda$.

[21] Using the notation and setup of appendix 3.B, we can derive the second moment matrix of the data, $\Sigma_{\Upsilon\Upsilon}$. Then, it is straightforward to verify that $m(\theta_0^R) = \Sigma_{Z\Upsilon}\iota\theta_0^R = 0$, at $\theta_0^R = (\gamma_0^R, \lambda_0^R) = (0.634, 0.007)$. (these tedious computations are performed using the `LtZasymptotics` class, outlined in appendix A section 1.3 at the end).

Table 4.3: Monte Carlo experiment of GMM estimators of (λ, γ) in (4.21), under $s_t \sim AR(1)$.

	γ			$\boldsymbol{\lambda}$		
Sample	mean	st. dev.	bias	mean	st. dev.	bias
50	0.609	0.093	0.018	0.012	0.150	-0.003
100	0.620	0.063	0.029	0.010	0.064	-0.005
150	0.624	0.050	0.033	0.008	0.043	-0.007
300	0.630	0.035	0.039	0.008	0.024	-0.007
500	0.632	0.027	0.041	0.008	0.016	-0.007
1000	0.635	0.019	0.044	0.007	0.010	-0.008
Parameter values for simulation: top line of table 4.1. Implied plims of estimators $\gamma_0^R = 0.634$, $\lambda_0^R = 0.007$.						
MCSE γ: 0.0013 - 0.0003, λ: 0.0021 - 0.0001.						

equation (4.21) is:

$$\begin{aligned}\pi_t &= -\frac{\lambda}{\gamma} s_{t-1} + \frac{1}{\gamma}\pi_{t-1} - \frac{1-\gamma}{\gamma}\pi_{t-2} - \frac{1}{\gamma}\epsilon_{t-1} + \xi_t \\ \Rightarrow \Delta\pi_t &= \alpha_1 s_{t-1} + \delta\Delta\pi_{t-1} - \frac{1}{\gamma}\epsilon_{t-1} + \xi_t \qquad (4.26)\end{aligned}$$

where $\alpha_1 = -\frac{\lambda}{\gamma}$, $\delta = \frac{1-\gamma}{\gamma}$, and ξ_t is an indeterminate martingale difference sequence w.r.t. the information set $\mathcal{F}_{t-1}$. The 'forward' solution (4.24) can be seen as a special case of (4.26) when we set $\xi_t = \frac{1}{\gamma}\epsilon_t + \alpha v_t$.[22] Of course, ξ_t could be any other martingale process, and hence, in general, the class of backward solutions is an Autoregressive Distributed lag Moving Average process, ADMA(2,1,1).

Two points are worth noting. First, denoting the composite error term $(\xi_t - \gamma^{-1}\epsilon_{t-1})$ by $\tilde{u}_t$, $\Delta\pi_{t-1}$ is *not* weakly exogenous for δ in the equation

$$\Delta\pi_t = \alpha_1 s_{t-1} + \delta\Delta\pi_{t-1} + \tilde{u}_t,$$

whenever ξ_t correlates with ϵ_t, since $\mathrm{E}[\epsilon_{t-1}\,\Delta\pi_{t-1}] \neq 0$ in that case. In other words, we cannot estimate the reduced form parameters (α_1, δ) by means of a simple regression of $\Delta\pi_t$ on s_{t-1} and $\Delta\pi_{t-1}$.[23]

---

[22] Substituting ξ_t into (4.26) and using the definitions of $s_t = \rho s_{t-1} + v_t$ and $\alpha = \frac{\lambda}{\gamma(1-\rho)}$ yields:

$$(1-L)\left(1-\frac{1-\gamma}{\gamma}\right)\pi_t = \frac{\lambda}{\gamma}s_{t-1} + \frac{1}{\gamma}(1-L)\epsilon_t + \frac{\lambda}{\gamma(1-\rho)}(s_t - \rho s_{t-1}) = \frac{\lambda}{\gamma(1-\rho)}(1-L)s_t + \frac{1}{\gamma}(1-L)\epsilon_t$$

and the common factor $(1-L)$ cancels to yield (4.24). Hence, the latter is a COMFAC restriction on the general solution (4.26).

[23] A GMM regression correcting for MA(1) serial correlation would seem appropriate here, albeit possibly inefficient. Indeed, such an approach has been used as an alternative to maximum likelihood for the estimation of ARMA models. On the other hand, if a distribution for ξ_t is specified, it may be possible to estimate the parameters of the reduced form (4.26) by maximum likelihood.

Secondly, the indeterminacy of ξ_t gives rise to the possibility of over-identification. Suppose, for instance, that $\xi_t = \epsilon_t$. Then, the solution (4.26) could be written as an infinite distributed lag equation:

$$\begin{aligned} \frac{1-\delta\, L}{1-\gamma^{-1}\, L}\Delta\pi_t &= \frac{\alpha_1}{1-\gamma^{-1}\, L} s_{t-1} + \epsilon_t, \quad \text{or} \\ \Delta\pi_t &= \tilde{\delta}_1\, \Delta\pi_{t-1} + \tilde{\delta}_2\, \Delta\pi_{t-2} + \ldots + \tilde{\alpha}_1\, s_{t-1} + \tilde{\alpha}_2\, s_{t-2} + \ldots + \epsilon_t \end{aligned}$$

where $\tilde{\delta}_i$ and $\tilde{\alpha}_i$ can be determined. Thus, the first-stage regression of the endogenous variable $\Delta_2\pi_{t+1}$ in the restricted model (4.25) would be

$$\Delta_2\pi_{t+1} = (\tilde{\delta}_1 - 1)\, \Delta\pi_t + \tilde{\delta}_2\, \Delta\pi_{t-1} + \ldots + \tilde{\alpha}_1\, s_t + \tilde{\alpha}_2\, s_{t-1} + \ldots \tag{4.27}$$

Evidently, all the lags of π_t and s_t would be relevant instruments.[24]

The indeterminacy of ξ_t is particularly problematic, since the complete specification of the structural system, (4.15), is insufficient to determine the local DGP (reduced form). Since different specifications of the reduced form have different implications for the identification of the structural model (4.7), the above identification analysis cannot be carried out without additional information. Such information could be obtained by modelling the reduced form directly, as is done in section 4.3 below.

A second restriction of interest is the pure forward-looking specification:

$$\mathrm{H}_\omega : \ \omega = 0 \Rightarrow \gamma_b = 0$$

In this case, the solution will involve setting $\delta = 0$ in (4.17), i.e.,

$$\pi_t = \alpha s_t + u_t$$

Then, the number of structural parameters $(\lambda, \gamma_f, \sigma_\epsilon, \rho, \sigma_v)$ will still be in excess of the reduced form ones $(\alpha, \sigma_\epsilon, \rho, \sigma_v)$, leaving the system un-identified. However, note that this is only true under the assumption that the model is correctly specified, i.e., when the null hypothesis H_ω is correct. If we imposed $\gamma_b = 0$ incorrectly, we would of course get a just-identified (but mis-specified) system. In that case, the specification test of over-identifying restrictions would not reject, and the implied estimate for the forward-looking coefficient would be greater than 1.

---

[24] Even the unrestricted specification (4.8) would be over-identified in this case. This can be easily verified by re-arranging (4.27) so as to derive the forecasting equation for π_{t+1}. In that equation, lags π_{t-2} and s_{t-1} onwards will be informative in predicting π_{t+1}, thus rendering the model over-identified (contrast this with (4.18) above).

Case 2: $s_t \sim ARDL(1,1)$

In the second leading case (4.13), where π_t Granger-causes s_t, the complete structural model becomes:

$$\begin{aligned} \pi_t &= \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t \\ s_t &= \rho s_{t-1} + \varphi \pi_{t-1} + v_t \end{aligned} \tag{4.28}$$

This time, the solution of the model no longer depends on the roots of the polynomial (4.16), but rather on the roots of the matrix polynomial

$$\Gamma(L) = \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \gamma_f & 0 \\ 0 & 0 \end{pmatrix} L^{-1} + \begin{pmatrix} \gamma_b & 0 \\ \varphi & \rho \end{pmatrix} L$$

So, the nuisance parameters (φ, ρ) come into play in determining the solution to the model. The possibility now arises that a solution may not exist (too many explosive roots, when φ is 'too big').[25] However, when φ is not too big, the conclusions of the simple AR(1) case hold, i.e., the model has a unique solution of the form (4.17) and hence the structural model remains un-identified. Thus we see that adding a feedback from inflation to s_t is not sufficient to yield identification.

Violation of weak exogeneity We also see another reason why s_t cannot be weakly exogenous for the parameters of interest $(\gamma_f, \gamma_b, \lambda)$: the range of values these parameters can take is not independent of parameters of the marginal process (φ, ρ). In other words, the parameters of interest $(\gamma_f, \gamma_b, \lambda)$ are *not variation free* with respect to the parameter of the marginal process, under the requirement that both variables be non-explosive. A necessary condition for these two sets to be variation free is that the marginal process parameters do not affect the roots of the characteristic RE polynomial which govern the existence of a solution. This in turn requires that the forcing variables (s_t here) are *not Granger-caused* by the endogenous variables (π_t). [26]

When weak exogeneity is violated, the resulting information loss may be potentially important in developing confidence intervals for the estimated parameters, especially when the point estimates are near the boundary of the admissible parameter space. It is also not unlikely that, in a full system analysis, point estimates for $(\gamma_f, \gamma_b, \lambda)$ and (φ, ρ) are such that the resulting forward-looking model does not have a (non-explosive) solution. The restrictions implicit in the existence

---

[25] The definition of a solution to a rational expectations model A.1 as well as the conditions for existence of a solution are given in the final appendix A, see proposition A.1.

[26] This condition can be test by means of a Granger-causality test in the first stage regression.

condition required for rational expectations models to be meaningful are rarely examined in the empirical literature.

In fact, by empirically estimating an ARDL(1,1) model for s_t on the GG data we notice that it is parsimoniously encompassed by the previous AR(1) specification (F(1, 149) = 1.33 with p-value 0.25). Also, the estimated feedback from lagged inflation is very small ($\widehat{\varphi} = 0.05$) and insignificant. Therefore, there is no need for further experiments in this case, as it is not empirically relevant.

4.2.2 A nearly identified system

We now proceed to the third leading case (4.14). This time the completed structural model is:

$$\begin{aligned} \pi_t &= \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t \\ s_t &= \rho_1 s_{t-1} + \rho_2 s_{t-2} + v_t, \qquad \epsilon_t \perp v_t \end{aligned} \tag{4.29}$$

Since s_t again receives no feedback from lagged inflation, the solution to this model will be determined by the roots of the polynomial (4.16), and will be of the form:[27]

$$\pi_t = \alpha_1 s_t + \alpha_2 s_{t-1} + \delta \pi_{t-1} + u_t \tag{4.30}$$

Now the number of reduced form parameters is the same as the number of structural ones, so the model is identified on the order condition. In fact, it is also identified on the rank condition when $\rho_2 \neq 0$, through the instrument s_{t-1}, as is verified by the forecasting equation for π_{t+1}:

$$\pi_{t+1} = (\alpha_1\rho_1 + \alpha_2 + \delta\alpha_1)s_t + (\alpha_1\rho_2 + \delta\,\alpha_2)s_{t-1} + \delta^2\pi_{t-1} + \underbrace{u_{t+1} + \delta\, u_t + \alpha_1\, v_{t+1}}_{\zeta_t} \tag{4.31}$$

(contrast this with (4.18) in the previous un-identified cases).

Moreover, the concentration parameter is non-zero here, unlike the previous two cases. This can be computed upon knowledge of the second moments in the data, as is explained in appendix 3.B. Using the parameter estimates from GG, the concentration parameter is of the order of 10^{-3} for a sample size of 100. This is remarkably small, in the light of the results in the previous two chapters, where it is found that a concentration parameter with at least two digits is required for the conventional asymptotic theory to work, also consistent with much earlier work by Anderson and Sawa (1979).

---

[27] δ and u_t are as in the AR(1) case (4.17), whereas

$$\alpha_1 = \frac{\lambda(1-\delta\gamma_f)}{[1-\gamma_f(\delta+\rho_1)](1-\gamma_f\delta) - \rho_2\gamma_f^2}, \qquad \alpha_2 = \frac{\lambda\gamma_f\rho_2}{[1-\gamma_f(\delta+\rho_1)](1-\gamma_f\delta) - \rho_2\gamma_f^2}$$

which simplify to (4.17), when $\rho_2 = 0$.

Simulation results

We repeat the experiments of section 4.2.1, in order to compare the behaviour of the estimators in the two settings. Unlike the previous under-identified cases, we would now expect to see convergence to the true values. We would also like to see how sensitive the results might be to the magnitude of the nuisance parameters, and how fast the asymptotic results are attained. The results are reported in table 4.4. The experiment in the top panel is based on the actual parameter values estimated from the data, and reveals weak identification, whereas the lower panel contrasts the results with an artificial situation of strong identification. Figure 4.3 compares the simulated distributions under weak and strong identification.

Table 4.4: Monte Carlo experiment of GMM estimators of $(\lambda, \gamma_f, \gamma_b)$ in (4.29) under $s_t \sim AR(2)$.

(a) Weak identification: $\rho_1 = 0.9$, $\rho_2 = -0.05$, $\sigma_\epsilon = 0.18$, $\sigma_v = 0.1$.
Concentration parameter $= 10^{-5} \times T$.

	γ_b			γ_f			λ		
Sample	mean	st. dev.	bias	mean	st. dev.	bias	mean	st. dev.	bias
50	0.410	0.120	0.032	0.421	0.330	-0.170	0.044	0.297	0.029
100	0.421	0.099	0.043	0.429	0.322	-0.162	0.039	0.164	0.024
150	0.425	0.093	0.047	0.428	0.298	-0.163	0.038	0.119	0.023
300	0.428	0.094	0.050	0.426	0.305	-0.165	0.039	0.084	0.024
500	0.430	0.092	0.052	0.425	0.291	-0.166	0.037	0.063	0.022
1000	0.429	0.094	0.051	0.429	0.293	-0.162	0.036	0.052	0.021

MCSE: γ_b: 0.0017-0.0013, γ_f: 0.0047-0.0041, λ: 0.0042-0.0001

(b) Strong identification: $\rho_1 = 0.9$, $\rho_2 = -0.8$, $\sigma_\epsilon = 0.057$, $\sigma_v = 1$.
Concentration parameter $= 0.459 \times T$.

	γ_b			γ_f			λ		
Sample	mean	st. dev.	bias	mean	st. dev.	bias	mean	st. dev.	bias
50	0.378	0.094	0.000	0.513	0.206	-0.078	0.017	0.008	0.002
100	0.380	0.056	0.002	0.556	0.129	-0.035	0.016	0.005	0.001
150	0.379	0.044	0.001	0.570	0.103	-0.021	0.016	0.004	0.001
300	0.378	0.029	0.000	0.581	0.066	-0.010	0.015	0.002	0.000
500	0.378	0.022	0.000	0.586	0.049	-0.005	0.015	0.002	0.000
1000	0.378	0.015	0.000	0.588	0.034	-0.003	0.015	0.001	0.000

MCSE: γ_b: 0.0013-0.0002, γ_f: 0.0029-0.0005, λ: 0.0001-0.00001

For both panels: Values for structural parameters $(\gamma_b, \gamma_f, \lambda)$ taken from top line of table 4.1.
plims of OLS estimators $\gamma_{b,ols} = 0.43$, $\gamma_{f,ols} = 0.43$, $\lambda_{ols} = 0.042$.

It is not surprising to see that the baseline experiment results do not differ substantially from the previous unidentified case, given the weakness of the additional instrument s_{t-2}. Comparing table 4.4 (a) for case 3 with table 4.2 for the un-identified case 1, we observe only a small reduction

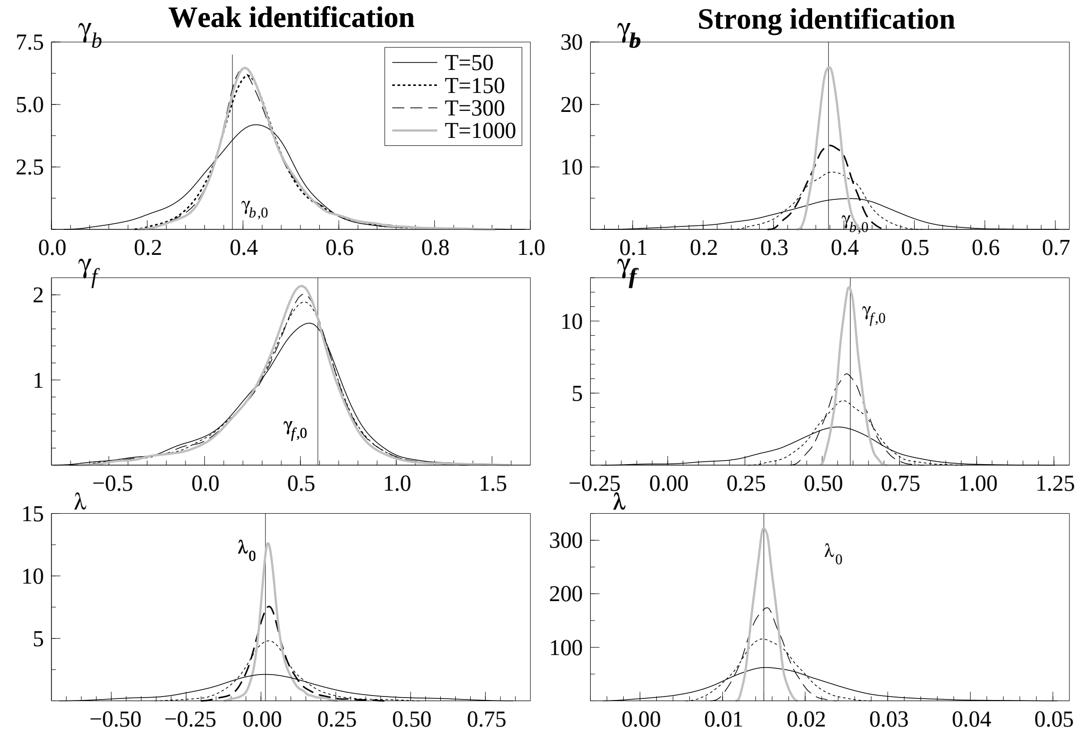


Figure 4.3: Simulated distributions of GMM estimators and weak and strong identification.

in the mean bias of estimators and their standard deviation as a result of $\rho_2 \neq 0$. Similarly, comparing the left columns of figures 4.2 and 4.3 we see almost no difference in the shape of the estimator distributions.

Panel (b) of table 4.4 presents simulation results for a case in which the concentration parameter is high, giving rise to strong identification. The latter is defined generically in (2.15), and, in this case, it depends on the reduced-form parameters $(\alpha_1, \alpha_2, \delta, \rho_1, \rho_2, \sigma_u, \sigma_v)$, which characterize the local DGP, and hence the second moments of the data. These, in turn, depend on the structural parameters $(\gamma_b, \gamma_f, \lambda)$ and the nuisance parameters $(\rho_1, \rho_2, \sigma_v, \sigma_\epsilon)$. The concentration parameter is a highly non-linear function of those nuisance parameters, and it would be counterproductive to derive its exact analytical expression for the purposes of this discussion. However, some intuition can be gained by thinking of it as a signal noise ratio in the first stage regression (4.31), where the signal is the instrument $(\alpha_1\rho_2 + \delta\,\alpha_2)s_{t-1}$ and the noise is $\zeta_t = u_{t+1} + \delta\, u_t + \alpha_1\, v_{t+1}$. It can be shown that the variance of the signal is increasing in $|\rho_2|$, when ρ_1 is kept fixed. Also, since u_t is proportional to ϵ_t (see footnote 27), the variance of the noise is increasing in σ_ϵ^2. The effect of σ_v^2 is ambiguous, since it contributes both to the variance of the noise and to that of the signal, but apparently the latter effect dominates for the parameter values used here.

This discussion helps explain why the concentration parameter based on the actual parameter estimates of $\rho_2 = -0.05, \sigma_v = 0.1$ and $\sigma_\epsilon = 0.18$ is very low (top panel of table 4.4). This is not only due to ρ_2 being small, but also because the standard error in the structural equation σ_ϵ is high relative to the variability of s_t.

To achieve a high value of the concentration parameter, we set $\rho_2 = -0.8$, and also make σ_v^2 100 times bigger and σ_ϵ^2 10 times smaller (none of these changes alone had any significant impact on the concentration parameter). These values differ sharply from the empirical estimates based on GG data.

The analysis of this section shows that if we assume the Galí-Gertler model (4.7) to be *correctly specified*, and under a data coherent representation of the distribution of the forcing variable s_t, the model is weakly identified on past information of π_t and s_t. Even at $T = 1000$ the estimators fail to converge to their true values. This situation is similar to other instances of weak identification, such as the Fuhrer, Moore, and Schuh (1995) linear-quadratic inventory model, where the authors report one case in which 30000 observations are needed for the GMM estimators to converge to the true value.

To corroborate this finding, we next conduct a formal test of identifiability.

4.2.3 Testing identifiability in the GG model

The test is based on the intermediate forecasting regression for the endogenous regressors given the instruments. Which regressors are treated as endogenous by GG depends on their choice of instruments. In line with the theoretical model of GG, we consider an instrument set that includes only lags of the variables in the model. In that sense, s_t is treated as endogenous, although this is not obvious from (4.7).

We give two versions of the test, one using only lags of s_t and π_t as instruments, and the other using additional variables that GG used. Those extra variables are commodity price and wage inflation, a measure of output gap and a long-short interest rate spread, see figure 4.6. Four lags of each of those variables are used as instruments. The results of the QMD test are given in table 4.5. Identification rank refers to the identifiability of the endogenous parameters, λ and γ_f. Rank 0 corresponds to no identification, rank 1 to partial identification (a linear combination of λ and γ_f is identifiable) and finally rank 2 implies full identification, see section 2.4 for details. We also report the (invalid) likelihood ratio Reduced Rank (RR) test for comparison test for comparison.

Table 4.5: Identifiability Tests.

	Small Instrument Set, $k = 8$		Large Instrument Set, $k = 24$	
Rank (r)	QMD	RR	QMD	RR
0	418.31 [0.000]	227.44 [0.000]	612.17 [0.000]	365.80 [0.000]
1	23.55 [0.001]	22.046 [0.001]	144.50 [0.000]	129.24 [0.000]

The p-values are given in square brackets, and they are based on the χ^2 distribution with degrees of freedom $(k - 1 - r)(2 - r)$, see section 2.4 for details.

We see that both reduced rank hypotheses 0 and 1 are strongly rejected in favour of complete identification. This is true for both the smaller and the larger instrument set. It is also evident that the additional instruments (gap, dc, dw, spr) reinforce the identification of the structural parameters.

In the light of these tests, we see that the evidence on identifiability seems to conflict with the theoretical analysis. On the one hand, by looking at the assumed completing model for s_t we concluded that the structural model must be un-identifiable. On the other hand, the test on the strength of the empirical instruments reveals strong evidence in favour of identification.[28] How can we reconcile these two contradictory findings?

The answer seems to lie in the fact that the original structural model may be mis-specified because of omitted dynamics, or other omitted variables, which are then incorrectly used as instruments. These possibilities are explored in the next section.[29]

4.3 Mis-specification analysis

In this section, we will offer an alternative systems analysis on the data set of Galí and Gertler (1999), in order to investigate the specification of their model.

We will be concerned with one particular type of mis-specification, namely the violation of the moment restrictions. This may come about due to: either (a) lags of the dependent variable being incorrectly omitted from the structural equation and used as instruments; or (b) other variables

---

[28] Note from chapter 3 how close the empirical size of the QMD test was to the nominal one, albeit only up to $k = 4$.

[29] This contradiction could also be due to mis-specification of the completing process for s_t. However, as the empirical analysis in the appendix shows, the AR(1) for s_t specification parsimoniously encompasses a General Unrestricted Model (GUM) containing four lags of all the other variables in the GG data set. Therefore, this is unlikely to be the source of the problem.

in the information set being used as instruments incorrectly.

The first cause of mis-specification is likely to show up as higher-order serial correlation in the estimated structural residuals. The second need not have that side effect (although it might, if the omitted variables are themselves autocorrelated), and can be investigated by modelling the reduced form of the complete system.

This type of mis-specification admits two interpretations. It could be seen as a failure of rationality, when the model is assumed to be correctly specified; or as dynamic mis-specification of the model, i.e., the omission of some lagged exogenous variables, which are then used as instruments, incorrectly. If the latter interpretation is adopted, the model can be potentially extended to incorporate the extra dynamics, and satisfy the rational expectations condition. However, this flexibility comes at the cost of creating too many observationally equivalent models, as we show in the final section of the chapter.

4.3.1 Omitted dynamics

To investigate these possibilities, we first examine the serial correlation pattern of the residuals from the GG model, noting that up to first-order is acceptable. Figure 4.4 displays the autocor-

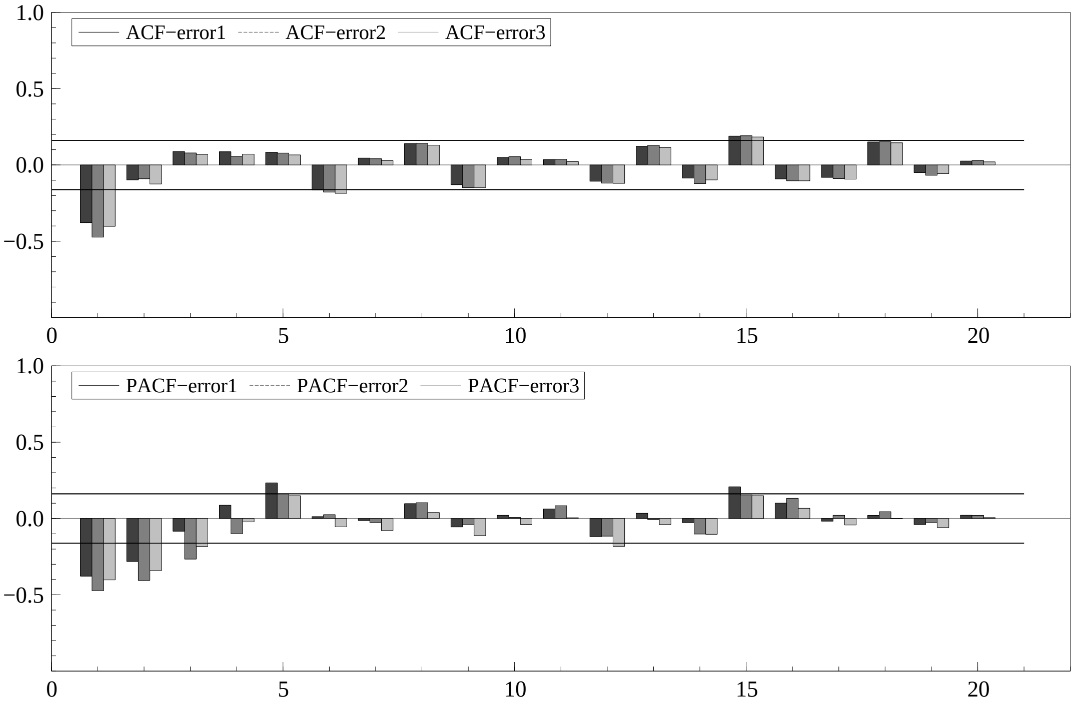


Figure 4.4: Autocorrelation of GG residuals.

relogram and the partial autocorrelogram for the residuals from three different versions of the GG model. The negative first-order autocorrelation is evident, and very significant. There is also some evidence of residual autocorrelation at further lags, albeit not very strong. Both the correlogram and the partial correlogram show that the MA(1) part of the series dominates.

Table 4.6: Residual Serial Correlation Tests from lags 2 to q.

	Model 1		Model 2		Model 3	
q	$\chi^2(q)$	P-value	$\chi^2(q)$	P-value	$\chi^2(q)$	P-value
5	5.42	0.247	7.77	0.100	5.09	0.278
6	8.75	0.119	8.50	0.131	9.21	0.101
7	10.06	0.122	10.87	0.092	10.65	0.100
8	14.83	0.038	15.44	0.031	14.74	0.040
9	17.20	0.028	17.54	0.025	17.40	0.026
10	17.98	0.035	18.24	0.033	18.26	0.032
11	19.28	0.037	20.96	0.021	19.57	0.034
$\mathrm{H}_0: \rho_i = 0$ for all $i > 1$, against $\mathrm{H}_1: \rho_i \neq 0$ for some $i > 1$.						

Table 4.6 reports formal tests for serial correlation performed on those residuals. The null hypothesis of no serial correlation is strongly rejected (not reported), confirming our prior that there is at least first-order serial correlation in the IV regression (4.8), as anticipated from Rational Expectations. We then test for any *excess* (i.e., higher-order) serial correlation, using the test developed in section 2.5.3. The sequence of tests reported in the table shows little evidence of autocorrelation at short lags, as anticipated from the correlogram, but more so over a longer horizon. These results are consistent with the Monte Carlo evidence on the low power of the excess serial correlation test given in chapter 3.

The presence of any serial correlation beyond lag 1 may have two implications for the proposed model (4.7). If it is due to omitted dynamics from that model, it will immediately imply that the latter is mis-specified, and hence the resulting estimation and inference will be misleading. Otherwise, if the structural error ϵ_t in the model (4.7) is itself assumed to be autocorrelated, it will mean that some of the instruments used to estimate it may be invalid, i.e., those that lie within the autocorrelation horizon of the error. Suppose, for instance, that ϵ_t follows a MA(q) process. Then, π_{t-2} to π_{t-q} and possibly other variables up to lag q such as $(s_t, dc_t, \text{etc.})$, would not be valid instruments.

In this analysis, we will focus on the first interpretation of serial correlation which is symptomatic of mis-specification. The reason is that the other interpretation could be dealt with in this framework by selecting the instruments appropriately. Yet, it has to be emphasized that it is methodologically incorrect to assume qth order serial correlation in the structural errors ϵ_t, say, and still use lags 1 to q of the dependent variable(s) as instruments.[30] Moreover, even if such an interpretation is given to the observed serial correlation and the instrument set is adapted accordingly (removing the appropriate lags of y_t), then the "excess serial correlation" which is symptomatic of mis-specification, will refer to the serial correlation beyond what is implicitly allowed for in the selection of instruments. For instance, if the researcher uses instruments dated $t-4$ onwards, excess serial correlation will mean residual serial correlation beyond lag 4, against the alternative of MA(4).

To analyse the implications of excess serial correlation, we need to generalize the DGP, so that the structural model (4.7) becomes dynamically mis-specified relative to it. Given the specificity of Monte Carlo experiments, rather than arbitrarily introducing further dynamics in the DGP, we estimate them from the data. Namely, we model the reduced form as a parsimonious ARDL(p,q) on π_t and s_t and then 'invert' this to find a forward-looking specification that has this reduced form as its solution.

The estimated reduced-form equation for inflation on GG data is

$$\widehat{\pi}_t = \underset{(0.0631)}{0.68}\ \pi_{t-1} + \underset{(0.0617)}{0.241}\ \pi_{t-3} + \underset{(0.161)}{0.554}\ s_t - \underset{(0.162)}{0.476}\ s_{t-2} \tag{4.32}$$

with $\widehat{\sigma}_u = 0.26$ and standard errors given in the parentheses (the omitted lags were insignificant). Symbolically, the reduced form can be written as:

$$\pi_t = \alpha\, s_t + \alpha_2\, s_{t-2} + \delta_1\, \pi_{t-1} + \delta_3\, \pi_{t-3} + u_t \tag{4.33}$$

Equation (4.33) together with an AR(1) specification for s_t enable us to derive $\mathrm{E}(\pi_{t+1}|\mathcal{F}_t)$:

$$\pi_{t+1|t} = \alpha\, \rho\, s_t + \alpha_2 s_{t-1} + \delta_1 \pi_t + \delta_3 \pi_{t-2} \tag{4.34}$$

Given some value for the forward-looking parameter γ_f, we can subtract $\frac{\gamma_f}{1-\gamma_f\delta_1}$ times (4.34) from both sides of equation (4.33) to get an isomorphic forward-looking specification:

$$\pi_t = \lambda s_t + \lambda_1 s_{t-1} + \lambda_2 s_{t-2} + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_1 \pi_{t-1} + \gamma_2 \pi_{t-2} + \gamma_3 \pi_{t-3} + \epsilon_t \tag{4.35}$$

---

[30] In particular, if the serial correlation of the structural errors is assumed to be potentially infinite, *no* lag of the dependent variable could be used as an instrument.

where the structural parameters $(\lambda, \lambda_1, \lambda_2, \gamma_1, \gamma_2, \gamma_3)$ are functions of the reduced form parameters.[31] When γ_f is set such that the lag polynomial $1 - \gamma_f L^{-1} - \gamma_1 L - \gamma_2 L^2 - \gamma_3 L^3$ has exactly one explosive root, (4.33) is the unique solution to the rational expectations model (4.35).

Obviously, the generalized forward-looking specification (4.35), which is consistent with the GG data, implies that the baseline equation (4.7) is dynamically mis-specified due to the omission of $(\pi_{t-2}, \pi_{t-3}, s_{t-1}, s_{t-2})$. Moreover, the structural equation (4.35) is unidentified, since there is an infinity of observationally equivalent models of the form (4.35) for different arbitrary choices of γ_f. On the other hand, the restricted equation (4.7) is now strongly identified through the instruments $(\pi_{t-2}, \pi_{t-3}, s_{t-1}, s_{t-2})$, although this identification is achieved via mis-specification.

The questions that naturally arise are: first, how detectable is that mis-specification through a test of over-identifying restrictions; and second, what are the properties of the estimators and test statistics under this type of mis-specification?

To answer these questions, we extend our previous Monte Carlo setting to simulate this more general DGP (4.33) and $s_t \sim AR(1)$. We use the Hansen-Sargan statistic to test the validity of the over-identifying restrictions, and investigate its power to detect mis-specification (a) when too many irrelevant instruments are used and (b) under 'too generous' corrections for serial correlation. This last point relates to the 12-lag HAC GMM estimator that Galí and Gertler used to robustify their inference. We argue that, in fact, such practice may result in significant loss of power to detect mis-specification and that, instead, these corrections should be made under the Null hypothesis of MA(1) serial correlation.

4.3.2 The Hansen-Sargan test

Figures 4.5 (i) and (ii) present QQ plots of the simulated distribution of the Hansen-Sargan test under the null hypothesis of no mis-specification and a fixed alternative corresponding to (4.35), denoted H_0 and H_1, respectively. The mis-specification can also be characterized by the non-centrality parameter ν_T, which measures the distance of the moment conditions from zero, under the null hypothesis of first-order serial correlation.

In each case, four different estimation procedures are considered: using a small ($k = 8$) versus a large ($k = 24$) instrument set; a short (1-lag) versus a long (12-lag) HAC estimator for the covariance of the moment conditions.[32] As expected, the power of the tests is increasing in the

---

[31] Matching coefficients yields: $\lambda = \alpha[1 - \gamma_f(\delta_1 + \rho)]$, $\lambda_1 = -\alpha_2 \gamma_f$, $\lambda_2 = \alpha_2(1 - \gamma_f \delta_1)$, $\gamma_1 = \delta_1(1 - \gamma_f \delta_1)$, $\gamma_2 = -\delta_3 \gamma_f$, $\gamma_3 = \delta_3(1 - \gamma_f \delta_1)$ and $\epsilon_t = (1 - \gamma_f \delta_1) u_t$.

[32] The small set contains only four lags of the variables in the model s_t and π_t, whereas the larger set contains

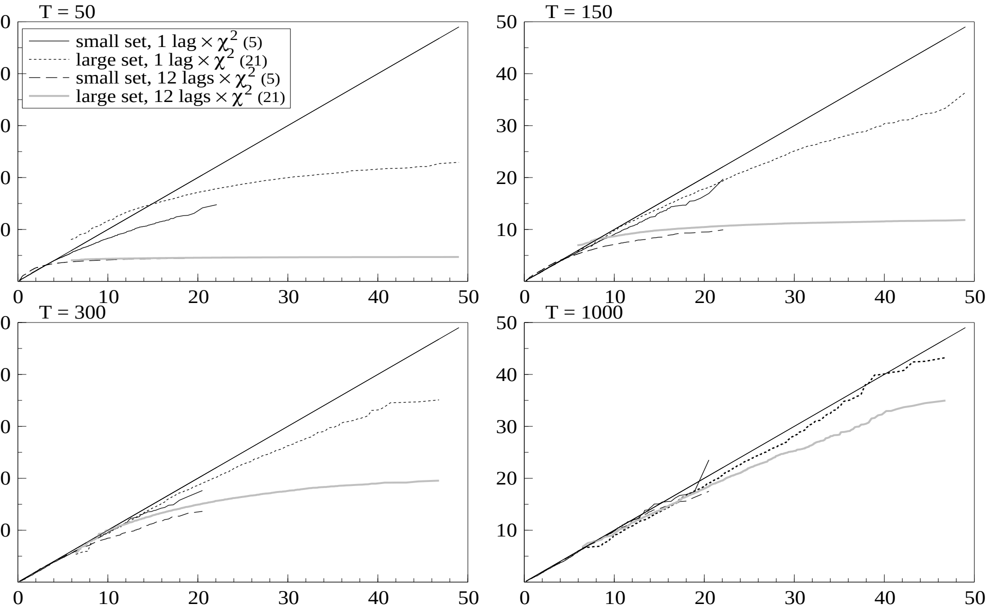


Figure 4.5: (i) QQ plots of the Hansen-Sargan test under the null ($\nu = 0$).

Estimated model: $\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t$

DGP: $\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t$

sample size, since we are testing against a fixed alternative. The QQ plots reveal significant departures of the Hansen-Sargan test from its assumed asymptotic $\chi^2(k-3)$ distribution in all four cases, but more so under HAC(12) covariance estimation. The statistic seems to be biased downwards relative to its asymptotic distribution under the null, but also under the alternative in small samples. Some intuition for these findings is given below.

Table 4.7: Rejection frequencies for the Hansen-Sargan test of over-identifying restrictions, 5% level.

	Under the null				Under the alternative			
	HAC(1)		HAC(12)		HAC(1)		HAC(12)	
T	$k=8$	$k=24$	$k=8$	$k=24$	$k=8$	$k=24$	$k=8$	$k=24$
50	0.018	0.000	0.000	0.000	0.031	0.000	0.000	0.000
150	0.020	0.003	0.000	0.000	0.252	0.009	0.000	0.000
300	0.039	0.009	0.012	0.000	0.540	0.133	0.220	0.000
1000	0.050	0.035	0.041	0.007	0.982	0.844	0.961	0.399

The bias of the test static is also demonstrated by the rejection frequencies at the 5% level under

---

all the instruments that GG used.

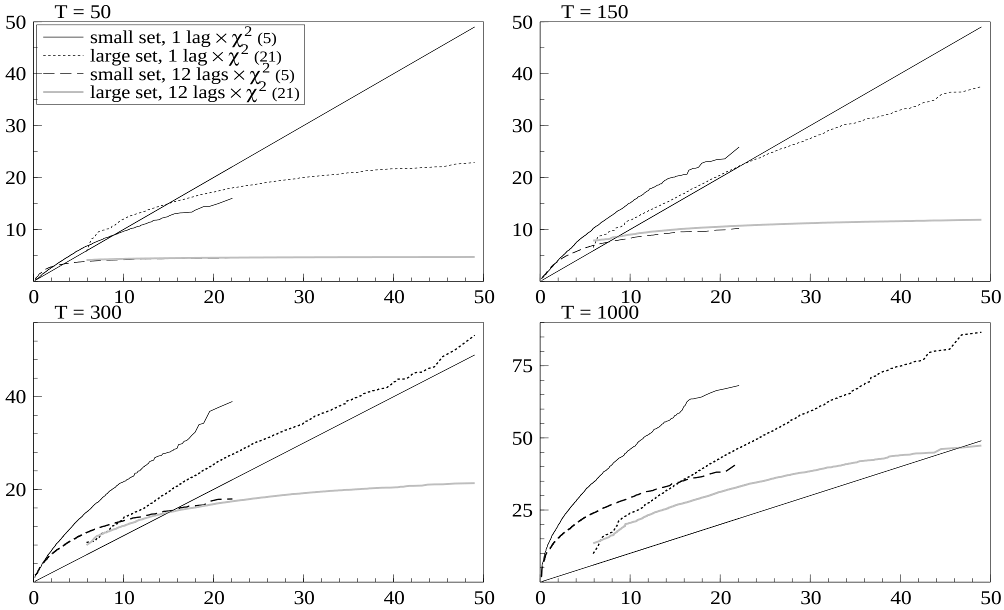


Figure 4.5: (ii) QQ plots of the Hansen-Sargan test under the alternative ($\nu_{100} = 0.14$).
Estimated model: $\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t$
DGP: $\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_1 \pi_{t-1} + \gamma_2 \pi_{t-2} + \gamma_3 \pi_{t-3} + \epsilon_t$

the null and the alternative, reported in table 4.7. The implications of using too many irrelevant instruments ('over-instrumenting'), and too much autocorrelation correction ('over-correcting') are in line with the theory and Monte Carlo evidence presented in the previous chapters. The test appears to seriously under-reject under the null. In view of the simulation study in section 3.4, this is due to the use of lagged instruments and the impact of HAC estimators. The under-rejection on the null is so dramatic (other than in case $k = 8$ and HAC(1)) that the test appears to be biased under the alternative. Moreover, over-correcting for serial correlation tends to cause further under-rejection. This can be explained by the fact that the autocorrelation corrections tend to overestimate the variance of the moment conditions under the null hypothesis, in small samples. Hence, the test statistic, which uses the inverse of that matrix, will tend to be biased downwards. This bias declines with the sample size, as expected.

Turning to the rejection frequencies under the alternative, we observe that power is falling in the number of over-identifying restrictions. The HAC(1)-based test seems to be more powerful under the alternative. This partly explained by the under-rejection on the null, but also by the fact that the HAC(1) covariance matrix estimator possibly under-estimates the true covariance when the model is mis-specified due to omitted dynamics. In other words, under the alternative of higher-order autocorrelation in the moment conditions, the difference between the variance matrix under the null and under the alternative is positive semi-definite. Hence, if a 1-step GMM estimator was used to perform the test, the test statistic would trivially be higher under the null, thus delivering higher power (and correct size). Nevertheless, the use of 2 (or more) step efficient estimators means that the statistic is no longer guaranteed to be more powerful, mathematically. Practically, however, the results of our present experiment show that testing using only first-order serial correlation correction performs significantly better.

The above results help explain why GG failed to detect any mis-specification with the Hansen-Sargan test using 24 instruments and correcting for 12th-order autocorrelation. The p-value of this test for the GG data is 0.97, well within the acceptance region of no mis-specification. Another example comes from the estimation of forward-looking monetary policy rules by Clarida, Galí, and Gertler (1998), where the authors estimate the same model for six different industrialized countries and report p-values for the Hansen-Sargan test of 0.999 in all cases. Even though the different data sets may be correlated, the likelihood of this event happening under the null of no mis-specification is implausibly small when the test is correctly sized, suggesting that it may actually be biased in that case too.[33]

4.3.3 Estimator distributions under mis-specification

GMM estimators under mis-specification can be seen to be biased away from their true value, in general, see corollary 2.2. As we show in that chapter, the bias in the parameters is not related to the extent to which mis-specification is detectable. In other words, mis-specification will manifest itself both as a bias in the estimated parameters, and as a non-centrality in a test of over-identifying restrictions.

We now turn to the simulated distributions of our estimators of interest. Given that we know the DGP in a Monte Carlo setting, we can derive $\boldsymbol{\theta}^*$ exactly, see appendix 3.B. This is reported in table 4.8, and the biases of the estimators are compared both against the hypothesized and the pseudo true values. We also report the concentration and mis-specification parameters, but this time these are computed not under the null, but under the pseudo-true values, revealing strong identification of the latter. Unsurprisingly, the bias relative to $\boldsymbol{\theta}^*$ is small and declines with T. In contrast, the bias relative to the hypothesized true parameter $\boldsymbol{\theta}_0$ is rising with T, even though it is incidentally small at $T = 50$.

---

[33] The probability that six independent test statistics will all have a p-value in excess of 0.9 is $0.1^6 = 0$.

Table 4.8: Monte Carlo experiment of GMM estimators of (γ_f, γ_b) in (4.15) under mis-specification.

	$\boldsymbol{\gamma_b}$				$\boldsymbol{\gamma_f}$			
T	mean	st. dev.	$\mathrm{E}(\widehat{\gamma}_b - \gamma_{b,0})$	$\mathrm{E}(\widehat{\gamma}_b - \gamma_b^*)$	mean	st. dev.	$\mathrm{E}(\widehat{\gamma}_f - \gamma_{f,0})$	$\mathrm{E}(\widehat{\gamma}_f - \gamma_f^*)$
50	0.420	0.145	-0.029	0.105	0.548	0.205	0.048	-0.152
150	0.367	0.097	-0.082	0.052	0.640	0.114	0.140	-0.060
300	0.340	0.071	-0.109	0.025	0.672	0.082	0.172	-0.028
1000	0.316	0.040	-0.132	0.001	0.699	0.045	0.199	-0.001

Estimated model: $\pi_t = \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t$
DGP: $\pi_t = 0.12\ s_t + 0.24\ s_{t-1} - 0.32\ s_{t-2} + 0.50\ \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + 0.45\ \pi_{t-1}$ $-0.12\ \pi_{t-2} + 0.16\ \pi_{t-3} + \epsilon_t, \quad \sigma_\epsilon = 0.17.$
Pseudo-true values: $\gamma_b^* = 0.32, \gamma_f^* = 0.70.$
Concentration: $\mu^2 = 0.18 \times T,$ Mis-specification: $\nu^2 = 0.0017 \times T.$
MCSE: $\gamma_b :$ $0.0045 - 0.0013, \quad \gamma_f : 0.0065 - 0.0014.$

4.3.4 Other omitted variables

Next, we consider the case in which some of the additional instruments in the information set (i.e., not just the lags of the variables already included in the structural equation, π_t and s_t), are not in fact orthogonal to the structural error. Unlike in a static structural model, it is not meaningful to selectively remove invalid instruments from the information set, because their invalidity carries important diagnostic information about the model. As we already mentioned, this invalidity may arise either due to dynamic mis-specification, e.g., because an exogenous variable or its lags have been implicitly removed from the model and used incorrectly as instruments; or, when the equation is assumed to be correctly specified, it may simply imply that expectations are not unbiased/rational relative to the available information.

As in a standard simultaneous equations model, we could answer this question by looking at the reduced form of the endogenous variables, here $(\pi_t, \pi_{t+1|t}, s_t)$, on the exogenous ones $\mathbf{z}_t$, containing four lags of inflation (π_t), the labour share (s_t), output gap (gap_t), yield spread (spr_t), wage growth (dw_t) and commodity price inflation (dc_t). The stylized reduced form would look like:

$$\begin{aligned} \widehat{y}_t &= a_1' Z_{1t} + a_2' Z_{2t} \\ \widehat{Y}_t &= \Pi_1' Z_{1t} + \Pi_2' Z_{2t} \end{aligned} \tag{4.36}$$

where $Z_{1t}(= \pi_{t-1}$, in this case) denotes variables used as exogenous regressors in the structural model:

$$\underbrace{\pi_t}_{y_t} = \underbrace{\gamma_b\ \pi_{t-1}}_{\gamma' Z_{1t}} + \underbrace{\gamma_f\ \pi_{t+1} + \lambda\ s_t}_{\beta' Y_t} + e_t$$

When that last model is correctly specified, it implies that the reduced form parameters a_2, Π_2

satisfy the restriction $a_2 = \mathrm{Col}\{\Pi_2\}$.

When this restriction is not satisfied, the model can be written as:

$$\begin{aligned}\pi_t &= \gamma_b\pi_{t-1} + \gamma_f\pi_{t+1} + \lambda s_t + \underbrace{\psi' Z_{2t} + e_t}_{\text{error}} \\ &= \gamma_b^*\pi_{t-1} + \gamma_f^*\pi_{t+1} + \lambda^* s_t + \underbrace{\bar{\psi}' Z_t + e_t}_{\text{residual}} \qquad (4.37)\end{aligned}$$

where $\bar{\psi}'Z_t$ is orthogonal to the optimal instruments $(Z_{1t}', Z_{2t}'\Pi_2)'$, see (4.36), and $(\gamma_f^*, \gamma_b^*, \lambda^*)$ denote the pseudo-true values (to which the GMM estimators will converge, provided these are strongly identified). As explained in the previous chapters, these are the values for the parameters that minimize the limiting objective function, and they will be different for different types of GMM estimators (e.g. 1-step, 2-step, CUE) as well as for different autocorrelation and heteroscedasticity corrections.

Of course, $(\gamma_f^*, \gamma_b^*, \lambda^*)$ as well as $\bar{\psi}$ can be computed exactly upon knowledge of the reduced form. The problem is that this is not known, and attempts to estimate it from the data are subject to the curse of dimensionality.[34] However, careful general-to-specific modelling, with the help of powerful software such as PcGets, see Krolzig and Hendry (2001), may help to overcome this difficulty.

We estimated a parsimonious PVAR(4) for the entire system, using PcGets on each of the equations, as described in Krolzig (2001). The final FIML estimates of the restricted reduced form are given in the appendix. Treating the estimated system as the DGP, we derive the pseudo-true values for the parameters of the baseline model (4.7), $(\gamma_b, \gamma_f, \lambda)$, for the 2-step GMM estimator with first-order autocorrelation correction, see appendix 3.B for details and appendix A for the relevant code. These values are $\gamma_b^* = 0.308$, $\gamma_f^* = 0.61$ and $\lambda^* = 0.034$, all within the range of values reported by GG, see table 4.1.

Given the pseudo-true values, the implied $\bar{\psi}$ can be derived as follows. Given the restricted reduced form of the entire system in the appendix, we can derive the forecast of the endogenous variables (π_{t+1}, s_t) on all the variables dated $t-1$,

$$\begin{aligned}\pi_{t+1|t-1} &= \Pi_{11}\pi_{t-1} + \Pi_{12}Z_{2t} \\ s_{t|t-1} &= \rho\, s_{t-1}\end{aligned}$$

where Z_{2t} consists of lags 2 to 4 of inflation and lags 1 to 4 of s_t, gap_t, spr_t, dw_t and dc_t. Subtracting $0.61\pi_{t+1|t-1}$ and $0.034\, s_{t|t-1}$ from both sides of the reduced form equation for inflation (4.A.1)

---

[34] A six-equation VAR(4) requires estimation of about 28 parameters per equation, thus being very imprecise.

given in the appendix yields, after some simplification, the following approximate equation for the mis-specified GG model:

$$\begin{aligned} \widehat{\pi}_t \;=\;& 0.308\ \pi_{t-1} + 0.610\ \pi_{t+1} + 0.034\ s_t + \\ & \underbrace{0.17\ \pi_{t-4} + 0.75\ \Delta s_{t-1} - 0.45\ \Delta s_{t-2} + 0.09\ \Delta gap_{t-1} - 0.04\ \Delta gap_{t-2} - 0.08\ \Delta_2 dw_{t-1}}_{\bar{\psi}' Z_{2t}} \end{aligned} \tag{4.38}$$

Upon comparison with the baseline GG specification of the hybrid Phillips curve, several variables appear to be missing, although we cannot provide an estimate of their significance in that equation.

In addition, the implied concentration and mis-specification parameters are:

$$\mu_T^2 = 3.33 \times T; \qquad \nu_T^2 = 0.0107 \times T$$

Thus, for a sample of size 150, we find strong identification and weak mis-specifiability of the model. This situation combined with the use of many potentially irrelevant over-identifying instruments (21 in total) and over-correction for serial correlation leads to very low power in detecting any mis-specification involved. As we already mentioned, the p-value for the Hansen-Sargan test on the GG data, using a 2-step GMM with Newey West HAC(12) is 0.97. However, when we use a 1-lag correction, this falls to 0.81. Moreover, if we chose a subset of the over-identifying instruments carefully, it is even possible to get the test statistic significant at the 5% level.[35]

The above discussion demonstrated the dangers of 'over-instrumenting' and 'over-corrections' in the estimation of forward-looking models. When mis-specification is difficult to detect by standard tests of over-identifying restrictions, the estimators may become biased in unknown directions. In that case, structural models of this form may almost become just-identifying re-parameterizations of the reduced form. That in itself is not a criticism of structural econometric modelling. Indeed, such re-parameterizations are desirable whenever they add useful economic intuition to a model. However, the use of invalid inferential procedures to provide evidence in favour of a just-identified model seems to be methodologically unjustifiable.

4.3.5 Anything goes...

In this final section we show how the un-detectability of mis-specification of structural models can give rise to several, apparently over-identified and but almost observationally equivalent models,

---

[35] Searching over the directions of maximal mis-specification revealed that the use of instruments $(\pi_{t-1}, s_3, dw_1, dw_3, dc_4, ygap_2)$ yields a p-value for the Hansen-Sargan test of 0.036. This is a drastic reduction compared with 0.81 using all instruments. Of course, choosing the instruments so as to maximize the value of the Hansen-Sargan test relative to its degrees of freedom will lead to the test being significantly over-sized, so this cannot be used as evidence of mis-specification.

with significantly different estimates for the parameters of interest.

The most important parameter in this model is arguably the coefficient on the inflation lead in the structural equation, γ_f. This is interpreted as a measure of 'forward-lookingness' in the determination of inflation, as we discussed above. Table 4.9 reports the different point estimates and associated t-values obtained for γ_f using various generalizations of the baseline structural equation (4.7), in the spirit of the analysis in the previous section. In the third column we give a list of the variables that are included as exogenous regressors in the estimated structural equation. It is assumed that all remaining variables in the instrument set should be used as instruments. The last two columns give the results of identification (QMD) and mis-specification (Hansen-Sargan) tests. These results are based on a 2-step GMM estimator allowing only for first-order residual serial correlation (HAC(1)).

Table 4.9: Alternative estimates of γ_f using different models.

$\widehat{\gamma}_f$	t-prob	Exogenous regressors	Identif.	Mis-spec.
0.19	[0.244]	$\pi_{t-1}, \pi_{t-4}, s_{t-1} \dots s_{t-3}, gap_{t-1} \dots gap_{t-3},$ $dw_{t-2}, dw_{t-3}, dc_{t-1} \dots dc_{t-4}$	[0.002]	[0.764]
0.25	[0.047]	$\pi_{t-1}, \pi_{t-4}, s_{t-1} \dots s_{t-3}, gap_{t-1} \dots gap_{t-3},$ $dw_{t-2}, dw_{t-3}, dc_{t-2}, dc_{t-4}$	[0.087]	[0.823]
0.31	[0.005]	$\pi_{t-1}, \pi_{t-4}, s_{t-1}, s_{t-2}, gap_{t-1} \dots gap_{t-3},$ $dw_{t-2}, dw_{t-3}, dc_{t-2}, dc_{t-4}$	[0.000]	[0.889]
0.41	[0.000]	$\pi_{t-1}, \pi_{t-4}, s_{t-1}, s_{t-2}, gap_{t-1}, gap_{t-2}, dw_{t-3}, dc_{t-4}$	[0.000]	[0.771]
0.56	[0.000]	$\pi_{t-1}, \pi_{t-4}, dw_{t-3}, dc_{t-4}$	[0.000]	[0.860]
0.61	[0.000]	π_{t-1}	[0.000]	[0.884]

As it is shown in the table, by choosing our identifying restrictions appropriately, we can get almost any estimate for the parameter of interest within the unit interval (and beyond). Also, in none of those models does the Hansen-Sargan test reject the validity of the instruments, which are also found to be strongly identifying.

4.4 Conclusions

In this study, we have been concerned with the problem of weak or partial identification of the parameters of a single equation structural econometric model with forward-looking rational expectations. In particular, the first part of the analysis was based on the assumption that the econometric model is *correctly* specified, but poorly identified. In the second part, we analyzed cases in which the model is strongly identified but mis-specified in ways that are hard to detect. In this setting, we examined the properties of single-equation estimation and inference based on some popular GMM estimators and test statistics. Our conclusions point to potential dangers of this type of econometric practice.

First, when the model is partially identified, the use of apparently valid moment restrictions may obscure the lack of identification of the parameters. Namely, adding irrelevant instruments on the basis of rational expectations restrictions, say, would result in an apparent 'over-identification' when the model may in fact be un-identified.

Second, the lack of identification will lead to inconsistent estimates of the (un-identifiable) parameters, and the addition of extra instruments will induce bias in the direction of OLS estimates. This is related to the well-known trade-off between efficiency and bias, although in this case, there is clearly no efficiency gain in adding instruments (when they are irrelevant). Moreover, the inconsistency in the estimates of the endogenous parameters will contaminate the estimates of any exogenous parameters.

Third, modelling the 'exogenous' driving process in the rational expectation model may prove highly informative in uncovering pathological situations in which the parameters of interest are poorly identified. In the case of the New-Keynesian Phillips curve, these pathological situations appear to be relevant empirically. However, identification could be achieved indirectly through mis-specification.

Fourth, when the model is mis-specified due to omitted dynamics, the use of too many lagged instruments as well as too general autocorrelation corrections is likely to reduce the power of mis-specification tests, and obscure the inadequacies of the structural model. In that sense, looking at the reduced form of the complete system may prove a valuable alternative to the single-equation approach, as it may help uncover that mis-specification.

Fifth, the presence of undetectable mis-specification may result in a multiplicity of almost observationally equivalent models, giving rise to the possibility of choosing a desirable model within this class. As a result, tests of over-identification lose their strength as evidence in favor of the proposed structure, (and therefore as evidence of forward-looking behavior).

In practice, careful selection of instruments may help avoid the problem of over-instrumenting that we analyzed here. Instead of that, a popular alternative is to include a large number of potentially relevant instruments, so as to maximize the chances of getting identification. We hope that the evidence presented here will deter researchers from following that route.

Finally, with regards to the estimation of the New Phillips curve of Galí and Gertler (1999), our analysis reveals that it is either weakly identified, or more probably mis-specified, casting doubts about its utility as a model of inflation dynamics.[36]

4.A Appendix

4.A.1 Description of the GG data

The empirical results of this chapter are based on the original data set of Galí and Gertler (1999).[37] The data is quarterly, and the sample size is from 1960:Q1 to 1997:Q4. The variable definitions and measurement are given in table 4.10. Variables gap, spr, dw and dc are plotted in figure 4.6.

Table 4.10: GG data description.

Mnemonic	Description	Definition
π_t	Quarterly inflation rate	100 Δ log(GDP deflator).
s_t	Labour share (in deviation from steady state)	$c \times 100 \log \frac{\text{unit labour cost}}{\text{unit price}}$. [a]
gap_t	Output gap	Quadratically detrended real GDP.
spr_t	Long-short interest rate spread	1y bond rate – 3m Fed funds rate.
dw_t	Quarterly wage growth	100 Δ log (unit labour costs).
dc_t	Quarterly commodity price inflation	100 Δ log(commodity prices).

[a] c is a correction factor due to Sbordone (2002): $c = \frac{n(m-1)}{m^2-n}$, where n is the share of labour in the Cobb Douglas production function $Y = A\,K^{1-n}\,L^n$, and m is the average markup of prices over unit costs. GG set $n = 2/3$ and $m = 1.1$, so $c = 0.12$.

4.A.2 Determination of simulation parameters

For the first leading case in section 4.2.1, the nuisance parameters, (ρ, σ_v), are determined by estimating (4.12):

$$\begin{array}{ccccl} s_t & = & \underset{(0.008)}{0.005} & + & \underset{(0.04)}{0.88} s_{t-1} + \widehat{v}_t \\ & & & & \qquad \widehat{\sigma}_v = 0.10 \end{array}$$

whereas σ_ϵ can be estimated either by estimating the variance of u_t in (4.17) and noting that ϵ_t is proportional to u_t, i.e. $\epsilon_t = (1 - \delta\gamma_f)u_t$ see (4.17); or using the IV residuals e_t defined in (4.9),

---

[36] These conclusions are consistent with results found independently by other researchers. Rudd and Whelan (2001) find similar evidence of mis-specification of the Galí-Gertler model through omitted variables, and give further intuition as to why this mis-specification may be undetectable.

[37] I thank Prof. Adrian Pagan for kindly sharing this data with me.

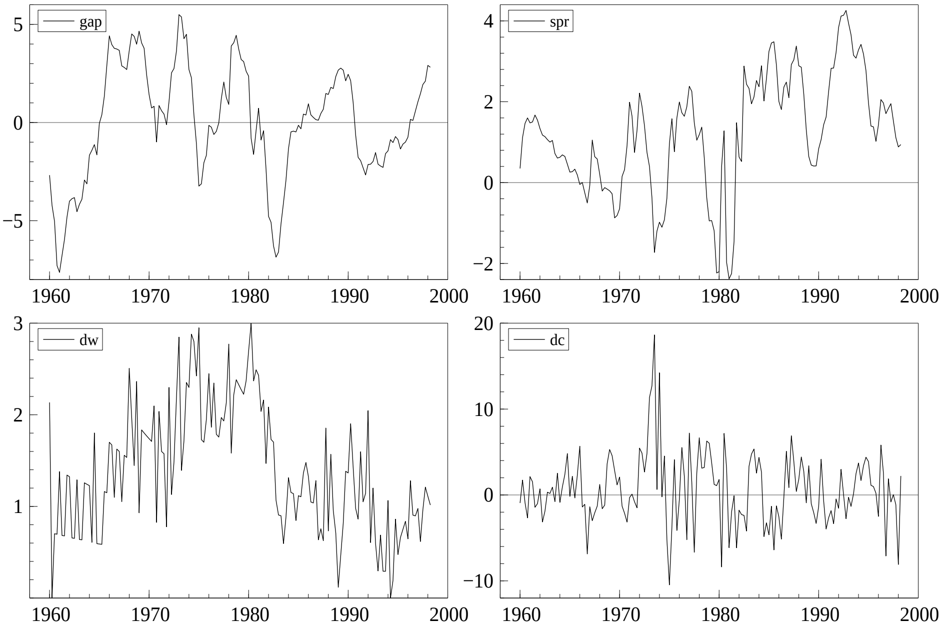


Figure 4.6: Plots of the variables in the Galí-Gertler data set.

given the parameter values set above and in the top line of table 4.1. From equation (4.19) we see that $e_t = (1 - \delta\,\gamma_f)\,u_t - \gamma_f\,u_{t+1} - \gamma_f\,\alpha\,v_{t+1}$. Thus, using either approach we would get:

$$\widehat{\sigma}_\epsilon^2 = (1-\delta\gamma_f)^2\widehat{\sigma}_u^2 \qquad \text{or} \qquad \tilde{\sigma}_\epsilon^2 = \left[1 + \frac{\gamma_f^2}{(1-\gamma_f\delta)^2}\right]^{-1}(\tilde{\sigma}_e^2 - \alpha^2\gamma_f^2\tilde{\sigma}_v^2)$$

respectively, where $\gamma_f = 0.591$, the value reported by GG (top line of table 4.1) and $\alpha = 0.11$, $\delta = 0.57$ are derived from the solution to the model given the AR(1) specification for s_t, see (4.17). The intermediate estimates for σ_e and σ_u on GG data were $\tilde{\sigma}_e = 0.22$ and $\widehat{\sigma}_u = 0.33$, yielding $\widehat{\sigma}_\epsilon = 0.18109$ or $\tilde{\sigma}_\epsilon = 0.17991$ for either approach, which are remarkably similar. Also, we set $\tilde{\rho} = 0.9$ and $\tilde{\sigma}_v = 0.10$.

For the third leading case (section 4.2.2), we need to estimate an AR(2) model for s_t. In fact, our estimated AR(2) model is parsimoniously encompassed by the previous AR(1) specification (F(1, 149) = 0.36 with p-value 0.54) and both the estimates of σ_v and ρ_1 remain virtually unchanged. The insignificant value of ρ_2 is set to the point estimate -0.05.

4.A.3 PVAR on inflation and instruments using PcGets

The data set is that used in Galí and Gertler (1999), including (π_t), the labour share (s_t), output gap (gap_t), yield spread (spr_t), wage growth (dw_t) and commodity price inflation (dc_t). Each series is plotted in figure 4.6.[38]

The estimated equations are:

$$\begin{aligned}\widehat{\pi}_t \;=\;& \underset{(0.0685)}{0.393}\ \pi_{t-1} + \underset{(0.0585)}{0.224}\ \pi_{t-4} + \underset{(0.275)}{0.771}\ s_{t-1} - \underset{(0.408)}{1.28}\ s_{t-2} + \underset{(0.271)}{0.694}\ s_{t-3} \\ & + \underset{(0.0314)}{0.0801}\ gap_{t-1} - \underset{(0.0519)}{0.123}\ gap_{t-2} + \underset{(0.0343)}{0.0731}\ gap_{t-3} + \underset{(0.00539)}{0.0175}\ dc_{t-2} \\ & + \underset{(0.00491)}{0.0149}\ dc_{t-4} + \underset{(0.0375)}{0.146}\ dw_{t-2} + \underset{(0.0383)}{0.117}\ dw_{t-3}, \qquad \widehat{\sigma}_u = 0.22\end{aligned} \tag{4.A.1}$$

$$\widehat{s}_t = \underset{(0.0397)}{0.873}\ s_{t-1}, \qquad \widehat{\sigma}_v = 0.10$$

$$\widehat{gap}_t = \underset{(0.0271)}{0.908}\ gap_{t-1} + \underset{(0.0173)}{0.0539}\ \mathrm{dc}_{t-1} + \underset{(0.0379)}{0.103}\ \mathrm{spr}_{t-2} - \underset{(0.347)}{1.37}\ s_{t-2}, \qquad \widehat{\sigma} = 0.85$$

$$\begin{aligned}\widehat{\mathrm{spr}}_t \;=\;& \underset{(0.0438)}{0.143}\ \pi_{t-4} - \underset{(0.0185)}{0.0557}\ gap_{t-1} - \underset{(0.0123)}{0.0313}\ \mathrm{dc}_{t-1} + \underset{(0.0795)}{0.984}\ \mathrm{spr}_{t-1} \\ & - \underset{(0.111)}{0.376}\ \mathrm{spr}_{t-2} + \underset{(0.0751)}{0.283}\ \mathrm{spr}_{t-3}, \qquad \widehat{\sigma} = 0.58\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{dw}}_t \;=\;& \underset{(0.105)}{0.926} + \underset{(0.115)}{0.386}\ \pi_{t-1} + \underset{(0.106)}{0.215}\ \pi_{t-4} + \underset{(0.0137)}{0.0498}\ gap_{t-1} \\ & - \underset{(0.0293)}{0.125}\ \mathrm{spr}_{t-4}, \qquad \widehat{\sigma} = 0.45\end{aligned}$$

$$\widehat{\mathrm{dc}}_t = \underset{(0.070)}{0.231}\ \mathrm{dc}_{t-3} + \underset{(2.06)}{13.27}\ \mathrm{OIL73}, \qquad \widehat{\sigma}_v = 3.53$$

---

[38] Prof. Muellbauer pointed out that the signs of the interest rate and commodity inflation effects in the output gap equation are implausible. For instance, the long-short interest rate spread could be seen as an expectation of future short rate movements, so that this equation seems to suggest that an expected increase in interest rates raises the output gap. This is consistent with omitted demand side variables typical of many VAR models. However, the objective here is to find a parsimonious representation of the reduced form given the *available* set of instruments, not to provide a congruent model for the complete system. Though undoubtedly useful, this would be beyond the scope of the current discussion.

Conclusions

The contributions of this thesis were threefold. First, we studied the asymptotic theory needed to analyze estimation and inference of forward-looking rational expectations models with the Generalized Method of Moments. Known results were extended to situations in which partial identification affects all of the estimated structural parameters. In that case, we showed that the GMM estimators of all the parameters become $O_p(1)$ and are centered away from the true values, and in the direction of the respective Least Squares estimators. In linear models, the limiting distributions of 2-step and Continuously updated GMM estimators are shown to be random mixtures of normals and to depend on unknown nuisance parameters. Moreover, those two estimators are no longer asymptotically equivalent under weak identification.

Weak identification also renders conventional test statistics, such as the Wald, Likelihood ratio and score-type tests, non-pivotal. Similar test, and confidence sets that are robust to weak identification proposed in the literature, are discussed. We also derive bounds to the distribution of conventional statistics, which can be used to derive valid but asymptotically conservative inference.

Next, we characterized the power of the Hansen-Sargan test of over-identifying restrictions, and showed that it is driven by the distance of only the *over-identifying* instruments from orthogonality, measured in the metric of the GMM objective function. Thus, we showed how mis-specification will manifest itself partly as a bias of the GMM estimators, which is undetectable, and partly as a non-centrality in the distribution of the test statistic. The power of the test crucially depends on the number of instruments, and, in finite samples, also on the type of HAC estimator of the weighting matrix. We also developed two simple tests of identification and excess serial correlation, which are applicable to dynamic forward-looking models.

The second contribution of the thesis is in providing simulation evidence on the quality of asymptotic approximations and finite-sample properties of estimators and inference, in situations that are relevant for forward-looking models, namely, where instruments are not strongly exogenous and structural errors exhibit strong negative serial correlation. We found that the local-to-zero asymptotic approximations to finite sampling distributions were good when instruments were strongly exogenous, but considerably worse for dynamic forward-looking models. In those models, robust confidence sets based on inverting the Anderson-Rubin statistic, such as the S-sets of Stock and Wright (2000), lead to conservative inference and large type II errors. The Hansen-Sargan test appears to under-reject under the null, especially when evaluated at the CUE. Hence the suggestion of an asymptotically boundedly pivotal version of the test based on the CUE doesn't appear to be of practical use. Much work is needed in that area, where contributions are still scant.[39]

The finite-sample size of the proposed test of excess serial was evaluated and was found to be close to its nominal size, even when serial correlation is present under the null hypothesis. However, the test appears to have little power in situations that are particularly relevant for forward-looking models, casting doubts on its utility as an alternative mis-specification test. The Quasi Minimum Distance identification test that we developed also appears to have problems, exhibiting size distortions that are increasing with the number of instruments. Furthermore, its main weakness is that it tests the null hypothesis of partial identification, and appears to be powerful in cases where conventional asymptotic theory is still inappropriate.

Moreover, with respect to the bias of the 2-step estimator versus the CUE, we found little evidence in favour of the latter in a situation akin to a forward-looking model with negative residual autocorrelation. This is in contrast to static models, where the CUE is preferable because it exhibits lower bias and is less sensitive to over-instrumenting than the 2-step estimator.

The third contribution is a critique of the single-equation GMM approach to the estimation of forward-looking monetary models. We showed why this approach involves information loss, and that the variables left un-modelled cannot be weakly exogenous for the structural parameters of interest. We also demonstrated the need to undertake identification analysis prior to estimation in order to identify pathological situations in which the model becomes poorly identified. When such pathological situations arise, the 'local-to-zero' first-order asymptotic theory provided above sheds light to the ensuing problems for GMM estimators and inference procedures, and these results can be further corroborated by Monte Carlo simulations tailored to the specifics of the model at hand.

---

[39] [40]ome promising work in the area of mis-specification testing under weak identification is currently undertaken by Kleibergen (2002), who showed how an asymptotically pivotal version of the J-statistic can be derived based on the Anderson-Rubin and K-statistic.

The analysis of mis-specification also demonstrated the dangers of over-instrumenting, i.e., using many possibly irrelevant instruments, as well as over-correction for serial correlation. Such practices have detrimental effects on the power of the Hansen-Sargan test to detect mis-specification in the form of omitted variables or dynamics. Alternative tests, such as tests of excess serial correlation, may sometimes be useful in detecting dynamic mis-specification, but more information can be gained by looking at the reduced form of the whole system of variables under analysis.

All in all, this thesis casts doubts on the utility of GMM for the estimation of dynamic forward-looking models, and recommends against the use of single-equation methods, and in favour of Full Information analysis of the system of variables under study.

Appendix A

Appendix: Computer programs

In this appendix, we give details of the main computer programs that we developed to implement the analysis in the present study. The programs are of interest beyond the scope of this thesis, because they have been designed to have wider applicability. They are based on Ox version 3.2, see Doornik (2001). The two main programs are:

1. `PcLiRE`: an Ox class designed to solve, estimate and simulate multivariate Linear Rational Expectations models (LiRE).
2. `Gmm`: an Ox class that performs most known GMM estimators, and implements new tests.

Additionally, we give details of the code developed to perform the asymptotic analysis and the simulations of chapters 3 and 4:

3. `LtZasymptotics`: asymptotic analysis for linear GMM models, and derivation of local-to-zero approximating distributions.
4. `IVGMM_simulator`: simulation of GMM estimators and test statistics for linear Rational Expectations models.

Each of the programs is explained in turn, and some short sample code for their implementation is provided so as to exemplify their use. The source code and documentation are available from the author. Although consigned to an appendix, writing and debugging the code was a major input to the thesis.

1.1 Solving multivariate Linear RE models

Solution methods for rational expectations models have been the subject of considerable research since the seminal work of Blanchard and Kahn (1980), see Binder and Pesaran (1995) for a review. One of the most commonly used solution algorithms in monetary economics is that of Anderson and Moore (1985). Here, we implement a recent approach proposed by Klein (2000), that is based on the generalized Schur decomposition. This approach is simpler and relatively more intuitive than the others.

In this appendix, we provide an exposition of the issues involved in solving multivariate linear Rational Expectations models, such as conditions for existence and uniqueness of solutions, and explain the computational procedure in `PcLiRE`.

Generic LiRE model A multivariate (infinite-horizon) Linear Rational Expectations model (LiRE) has the generic form:

$$\sum_{i=0}^{k}\sum_{j=0}^{l} M_{ij}\,\mathrm{E}(y_{t+j-i}|\mathcal{F}_{t-i}) = Q\,z_t, \qquad t = 0, 1, \ldots \tag{1.1}$$

where y_t is a n-dimensional vector of decision or 'endogenous' variables, z_t is a m-dimensional vector of 'forcing variables' or 'driving processes', M_{ij} and Q are fixed coefficient matrices of dimension $n \times n$ and $n \times m$ respectively, and $\mathcal{F}_t$ denotes the (non-decreasing) information set at time t, such that y_t, z_t are adapted to it. This implies that $\mathcal{F}_t$ contains current and lagged values of (y_t, z_t). Using the notation $Y_T^1 = \{y_t; t = 1, 2, \ldots, T\}$ and $Z_T^1 = \{z_t; t = 1, 2, \ldots, T\}$ for the history of (y_t, z_t) with Y_0 and Z_0 denoting initial conditions, $\mathcal{F}_t = \{Y_t^1, Y_0; Z_t^1, Z_0\}$. Also, $\mathrm{E}(y_t|\mathcal{F}_s)$ denotes the conditional expectation of y_t given $\mathcal{F}_s$, which is a (possibly non-linear) function of the variables in the information set.

A stylized example could be:

$$M_{00}\,y_t + M_{10}\,y_{t-1} + M_{20}\,y_{t-2} + M_{01}\,\mathrm{E}(y_{t+1}|\mathcal{F}_t) + M_{11}\,\mathrm{E}(y_t|\mathcal{F}_{t-1}) + M_{21}\,\mathrm{E}(y_{t+1}|\mathcal{F}_{t-1}) = Q\,z_t$$

Solution Next, we explain what is meant by a solution to a Rational Expectations model. Since the terms $\mathrm{E}(y_{t+j-i}|\mathcal{F}_{t-i})$ in (1.1) are functions of $(Z_t^1, Y_{t-1}^1, Z_0, Y_0)$, it follows that y_t is an implicit function of z_t and the rest of the variables in the information set $\mathcal{F}_{t-1}$, parameterized by the coefficients $\{M_{ij}\}_{i,j}$ and Q, say

$$y_t = f(z_t, Z_{t-1}^1, Y_{t-1}^1, Z_0, Y_0; \theta),$$

where θ is a vector containing all the elements of $\{M_{ij}\}_{i,j}$ and Q. Thus, conditional on the history $(Y_{t-1}^1, Z_{t-1}^1, Y_0, Z_0)$, the distribution of y_t can be determined by a change of variables given the conditional distribution of z_t.

However, the definition of f appears to be circular: on the one hand, we need to know $\mathrm{E}(y_{t+j-i}|\mathcal{F}_{t-i})$ before we can get f, but on the other hand, we need to know f in order to derive the conditional expectations. A solution to the LiRE model (1.1) involves finding a function f such that the resulting stochastic process y_t satisfies equation (1.1).

Such a stochastic process will always exist, and so the problem is trivial without further restrictions. Hence, we also impose the additional requirement that the process y_t should be non-explosive. This implies that $\mathrm{E}(y_{t+\tau}|\mathcal{F}_t)$ is bounded for all $\tau > 0$ and hence rules out speculative bubbles ('no-bubbles condition').

The above discussion is summarized in the following definition:

Definition A.1 (Solution to LiRE model). *Given a law for the stochastic process z_t in (1.1), a solution to the multivariate linear Rational Expectations model (1.1) is a stochastic process y_t that: (i) satisfies the difference equation (1.1); and (ii) is non-explosive.*

The forcing variables For most of the models in the monetary economics literature (and elsewhere), the forcing variables z_t follow a stable linear process.[1] Thus, we restrict attention to such processes in our implementation. Without loss of generality, we assume that z_t can be represented by a first-order Vector Autoregression:

$$z_t = c_z + \Phi\, z_{t-1} + u_t \tag{1.2}$$

where $u_t \sim (0, \Sigma_u)$ is a vector mean innovation process w.r.t. $\mathcal{F}_{t-1}$. This can be seen as a companion form of a more general VARMA(p,q) specification. For instance, let e_t be a vector white noise process, $e_t \sim iid(0, I)$, and suppose that the driving process is x_t and follows a VARMA(2,1):

$$C_0\, x_t - C_1\, x_{t-1} - C_2\, x_{t-2} = D_0\, e_t + D_1\, e_{t-1}$$

One way of writing this in companion form is:

$$z_t = \Phi_0^{-1} \Phi_1\, z_{t-1} + \Phi_0^{-1}\, \mathbf{e}_t,$$

---

[1] There is no loss of generality in ruling out unit roots in z_t. If there was an integrated variable in z_t, it can be categorized as an endogenous variable by including it in y_t, treating only its stationary part as a forcing variable. This will also enable a more thorough cointegration analysis of the system.

$$z_t = \begin{pmatrix} e_t \\ x_t \\ x_{t-1} \end{pmatrix} \quad \mathbf{e}_t = \begin{pmatrix} e_t \\ 0 \\ 0 \end{pmatrix} \quad \Phi_0 = \begin{pmatrix} I & 0 & 0 \\ -D_0 & C_0 & 0 \\ 0 & 0 & I \end{pmatrix} \quad \Phi_1 = \begin{pmatrix} 0 & 0 & 0 \\ D_1 & C_1 & C_2 \\ 0 & I & 0 \end{pmatrix}$$

such that $\Phi = \Phi_0^{-1}\Phi_1$ and $u_t = \Phi_0^{-1}\mathbf{e}_t$.

Notice that we have also assumed that z_t is *not Granger-caused* by y_t. There is no loss of generality in that either, since if there is an element of z_t that doesn't fulfil this requirement, it can be absorbed into the endogenous variables y_t. The Granger-non-causality of y_t for z_t is used to simplify the solution, as we see below, because it implies that $\mathrm{E}(z_{t+j}|\mathcal{F}_t)$ is only function of the history of z_t.

We summarize the above discussion in the following statement.

Assumption 1. *The forcing variable z_t in the LiRE model (1.1) follows a stationary VAR(1) process, and is not Granger-caused by y_t.*

Predetermined variables For the solution, it is useful to define the concept of a predetermined variable.

Definition A.2 (Predetermined variable). *A variable x_t is* predetermined *if $\mathrm{E}(x_t|\mathcal{F}_s) = x_t$ for some $s < t$, where $\mathcal{F}_s$ is an information set containing at least current and past values of x_s. Setting $\tau = \min\{s \,:\, \mathrm{E}(x_t|\mathcal{F}_s) = x_t\}$, we say that x_t is predetermined for $t - \tau$ periods.*

An example of a predetermined variable is one that lies in the $(t-1)$-dated information set or earlier, e.g.,, $x_t = \mathrm{E}(y_{t+2}|\mathcal{F}_{t-1})$ or $x_t = y_{t-i}$ for $i > 0$. In contrast, all variables dated t, such as y_t, $\mathrm{E}(y_{t+j}|\mathcal{F}_t)$, $j > 0$, are non-predetermined.

The PcLiRE model Even though the above setup allows the treatment of models with back-dated expectational terms such as $\mathrm{E}(y_{t+j-i}|\mathcal{F}_{t-i})$ for $i > 0$ and $j > 0$, as claimed by Binder and Pesaran (1995), the solution of these models presents some difficulties that require a different treatment and somewhat stronger assumptions for existence of a solution, as we explain below. Thus, our implementation focuses only on contemporaneous expectations, i.e., we consider only models with $M_{ij} = 0$ for $i, j > 0$ in (1.1):[2]

$$\sum_{i=0}^{k} A_{-i}\, y_{t-i} + \sum_{j=1}^{l} A_j\, \mathrm{E}(y_{t+j}|\mathcal{F}_t) = Q\, z_t, \qquad t = 0, 1, \ldots \tag{1.3}$$

---

[2] We adopted the simpler notation A_i for the difference equation coefficients, where a positive subscript corresponds to a lead and a negative to a lag.

Example: Consider the Galí-Gertler model of section 4.2:

$$\begin{aligned} \pi_t &= \lambda s_t + \gamma_f \mathrm{E}(\pi_{t+1}|\mathcal{F}_t) + \gamma_b \pi_{t-1} + \epsilon_t \\ s_t &= \rho s_{t-1} + \varphi \pi_{t-1} + \epsilon_{2t}, \end{aligned} \tag{1.4}$$

with $(\epsilon_t, \epsilon_{2t})' \sim NID(0, I)$. This can be cast in the form (1.3) as follows:

$$\underbrace{\begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix}}_{A_0} \underbrace{\begin{pmatrix} \pi_t \\ s_t \end{pmatrix}}_{y_t} + \underbrace{\begin{pmatrix} -\gamma_f & 0 \\ 0 & 0 \end{pmatrix}}_{A_1} \underbrace{\begin{pmatrix} \pi_{t+1|t} \\ s_{t+1|t} \end{pmatrix}}_{y_{t+1|t}} + \underbrace{\begin{pmatrix} -\gamma_b & 0 \\ -\varphi & -\rho \end{pmatrix}}_{A_{-1}} \underbrace{\begin{pmatrix} \pi_{t-1} \\ s_{t-1} \end{pmatrix}}_{y_{t-1}} = \underbrace{\begin{pmatrix} \epsilon_t \\ \epsilon_{2t} \end{pmatrix}}_{z_t = u_t}.$$

The corresponding `PcLiRE` code is given below.

-------------------------- Sample Code ---------------------------

```
#include "PcLiRE1.ox"
main()
{
    decl cY, model, mA1, mA0, mAb1;
    cY = 2;                         // number of endogenous variables.
    model = new PcLiRE(cY);         // create the PcLiRE object.
    mA1 = <-0.591, 0; 0, 0>;        // specify coefficients.
    mA0 = <1, -0.05; 0, 1>;
    mAb1 = <-.378, 0; 0.1, -0.9>;
    model.SetYParameter({mA1, 1, 0}, {mAb1, -1, 0}, {mA0, 0, 0});
                                    // set LiRE equation parameters.
/*  Defaults:
    model.SetZParameter(unit(2), zeros(2,2), zeros(2,1));
    model.SetUParameter(unit(2));
*/
}
```

The function `SetYParameter(...)` takes an unspecified number of arguments, thus allowing the solution of models with arbitrary number of lags/leads.[3] Each argument is a three-dimensional array: the first argument is the coefficient matrix, the second is an integer specifying whether it refers to a lead (positive) or lag (negative). The third argument refers to the information set (zero for contemporaneous expectations $\mathrm{E}(\cdot\ |\mathcal{F}_t)$). The last two functions specify Q, Φ, c_z and Σ_u. They are redundant in this example, because their values coincide with the program defaults: $Q = I$, $\Phi = 0$ and $c_z = 0$ and $\Sigma_u = I$.

1.1.1 The LiRE model in canonical form

Equation (1.3) is a high-order expectational difference equation, but it can be written as a first-order system. This is referred to as the **canonical form** of model (1.3):

$$A_c\, \mathrm{E}(\mathbf{x}_{t+1}|\mathcal{F}_t) = B_c\, \mathbf{x}_t + Q_c\, z_t \tag{1.5}$$

---

[3] However, computational problems arise when `cY` is large.

This can be done in many ways. In our implementation, we have chosen to stack all the non-predetermined variables at the bottom of $\mathbf{x}_t$, as follows:

$$\underset{n(k+l)\times 1}{\mathbf{x}_t} = \begin{pmatrix} y_{t-k} \\ \vdots \\ y_t \\ \vdots \\ y_{t+l-1|t} \end{pmatrix} \qquad Q_c = \begin{pmatrix} Q \\ \mathbf{0}_{(k+l)(n-1)\times m} \end{pmatrix} \qquad A_c = \begin{pmatrix} A_{-k+1} & \dots & A_0 & \dots & A_l \\ I_n & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \dots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & I_n & 0 \end{pmatrix}$$

$$\underset{pk\times pk}{B_c} = \begin{pmatrix} -A_{-k} & 0 & \dots & 0 \\ 0 & I_n & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & I_n \end{pmatrix}$$

Let $N = n(k+l)$ denote the dimension of canonical vector $\mathbf{x}_t$. This canonical form is derived in `PcLiRE` by the function `GetCanonicalForm()`, which casts A_c, B_c and Q_c into the class members `m_mA`, `m_mB` and `m_mQ`.

It is also noteworthy to point out that the non-zero latent roots of the LiRE matrix polynomial $A(\lambda) = \sum_{i=-k}^{l} A_i\,\lambda^i$ are the same as the generalized eigenvalues of the canonical matrix pair (A_c, B_c). Those latent roots are the λs that make $A(\lambda)$ singular, i.e., they satisfy the equation:

$$\left(\sum_{i=-k}^{l} A_i\,\lambda^i\right) v = 0$$

for some $v \neq 0$. For $\lambda \neq 0$, this can be re-written equivalently as

$$\left(\sum_{i=-k}^{l} A_i\,\lambda^{i+k}\right) v = 0.$$

Now, consider a (right) generalized eigenvector $v = (v_1', \dots, v_{k+l}')'$ of the pair (A_c, B_c) which satisfies the equation $\lambda\,A_c\,v = B_c\,v$, or

$$\begin{aligned} \sum_{i=-k+1}^{l} A_i\,\lambda\,v_{i+k} &= -A_{-k}\,v_1 \qquad (1.6) \\ \lambda v_i &= v_{i+1}, \qquad i = 1, \dots, k+l-1. \end{aligned}$$

Thus, $v_{i+1} = \lambda^i\,v_1$ and inserting this in (1.6) we get:

$$\sum_{i=-k+1}^{l} A_i\,\lambda^{i+k} v_1 + A_{-k}\,v_1 = \left(\sum_{i=0}^{k+l} A_{i-k}\lambda^i\right) v_1 = 0,$$

thus proving that $\lambda \neq 0$ is also a latent root of $A(z)$.

The roots of $A(z)$ may be printed out in `PcLiRE`, by calling the function

```
model.PrintRoots(TRUE);   //default is FALSE.
```

see the examples below.

1.1.2 Solving the Canonical LiRE model

By construction, $\mathbf{x}_t$ is split into predetermined and non-predetermined variables:

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{x}_t^p \\ \mathbf{x}_t^f \end{pmatrix} \quad \mathbf{x}_t^p = \begin{pmatrix} y_{t-k} \\ \vdots \\ y_{t-1} \end{pmatrix} \quad \mathbf{x}_t^f = \begin{pmatrix} y_t \\ \vdots \\ y_{t+l|t} \end{pmatrix} \tag{1.7}$$

Let $N_p = nk$ denote the number of predetermined variables in the system and $N_f = nl$ the number of non-predetermined ones. Since the predetermined variables are lags of y_t, which are known, the solution to the model will involve expressing y_t as a function of the predetermined variables $\mathbf{x}_t^p$ and the forcing variables z_t. Evidently, the only variables that need to be 'solved' are the non-predetermined ones (also referred to as forward-looking), $\mathbf{x}_t^f$.

We do this by computing the Generalized (real) Schur decomposition of the matrices A_c and B_c in (1.5). This involves solving the generalized eigenproblem $|A_c\, x - B_c| = 0$ (where A_c may be singular, otherwise the problem reduces to a simple one) and finding a block upper triangular matrix S and an upper triangular matrix T with associated orthogonal matrices U and V such that $A_c = U\, S\, V'$ and $B_c = U\, T\, V'$.[4] A generalized eigenvalue for a pair of matrices (A_c, B_c) is a scalar λ or a ratio $\alpha/\beta = \lambda$, such that $|\lambda\, A_c - B_c|$ is singular. It is usually represented as the pair (α, β), where α could be complex and β is real, as there is a reasonable interpretation for $\beta = 0$ ('infinite' eigenvalue) or both being zero (ill-conditioned case). Moreover, when λ_i is real, it is equal to t_{ii}/s_{ii}, whereas when it is complex, λ_i and it's conjugate λ_{i+1} are the generalized eigenvalues of the (2×2) submatrix pair $\big(S_{[i:j][i:j]}, T_{[i:j][i:j]}\big)$, $j = i + 1$.

The above decomposition is non-unique. In particular, we may re-order the generalized eigenvalues λ_i and partition S and T conformably so that the blocks at the bottom S_{22} and T_{22} correspond to the explosive roots ($|\lambda_i| > 1$), see Klein (2000, Theorem 3.1):

$$\begin{aligned} A_c = USV' &= \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} V_1' \\ V_2' \end{pmatrix} \\ B_c = UTV' &= \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} V_1' \\ V_2' \end{pmatrix} \end{aligned} \tag{1.8}$$

---

[4] The diagonal of S contains (1×1) or (2×2) blocks, corresponding to real and complex eigenvalues respectively.

Similarly, we define

$$\mathbf{w}_t = V'\mathbf{x}_t = \begin{pmatrix} V_1'\mathbf{x}_t \\ V_2'\mathbf{x}_t \end{pmatrix} = \begin{pmatrix} \mathbf{w}_t^s \\ \mathbf{w}_t^e \end{pmatrix},$$

where $\mathbf{w}_t^s$ are the linear combinations of $\mathbf{x}_t$ corresponding to the non-explosive roots, whereas $\mathbf{w}_t^e$ correspond to the explosive roots.

Note that, unlike Klein (2000), we did not rule out the presence of unit roots, which is why we distinguished between non-explosive/explosive, as opposed to stable/unstable eigenvalues. Indeed, the presence of unit roots may generate some additional considerations, which we discuss later on. However, it is highly restrictive to rule them out at the outset, since they are common both in empirical and in theoretical macroeconomic models.

We note the dimensions of those two different sets, N_s and N_e, which are crucial for the existence and uniqueness of the solution. It is also useful to partition the columns of V conformably with the classification of eigenvalues and its rows conformably with the distinction into predetermined and non-predetermined variables as follows:

$$V = \begin{pmatrix} \underset{N_p \times N_s}{V_{11}} & \underset{N_p \times N_e}{V_{12}} \\ \underset{N_f \times N_s}{V_{21}} & \underset{N_f \times N_e}{V_{22}} \end{pmatrix} \tag{1.9}$$

Hence, the following relationships hold:

$$\mathbf{w}_t^s = V_{11}'\,\mathbf{x}_t^p + V_{21}'\,\mathbf{x}_t^f \tag{1.10}$$

$$\mathbf{w}_t^e = V_{12}'\,\mathbf{x}_t^p + V_{22}'\,\mathbf{x}_t^f \tag{1.11}$$

$$\mathbf{x}_t^f = V_{21}\,\mathbf{w}_t^s + V_{22}\,\mathbf{w}_t^e \tag{1.12}$$

The canonical model (1.5) can then be equivalently re-written as the (quasi) triangular system:

$$\begin{pmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{pmatrix} E\left[\begin{pmatrix} \mathbf{w}_{t+1}^s \\ \mathbf{w}_{t+1}^e \end{pmatrix} \middle| \mathcal{F}_t\right] = \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_t^s \\ \mathbf{w}_t^e \end{pmatrix} + \begin{pmatrix} U_1' \\ U_2' \end{pmatrix} Q_c\, z_t. \tag{1.13}$$

For convenience, we gather the notation in the table below:

Mnemonic	Description
n	number of endogenous variables y_t.
m	number of forcing variables z_t.
k	number of lags in LiRE model (1.3).
l	number of leads in LiRE model (1.3).
N	dimension of canonical vector $\mathbf{x}_t$, $N = n(k+l)$.
N_p, N_f	number of pred./nonpred. variables in $\mathbf{x}_t$: $N_p + N_f = N$.
N_s, N_e	number of non-explosive/explosive eigenvalues in (1.5): $N_s + N_e = N$.

Conditions for existence and uniqueness of a solution

The following regularity conditions are needed.

Assumption 2. *The matrix polynomial* $P(x) = A_c\, x - B_c$ *is* regular*, i.e., there exists* $x \in \mathbb{C}$ *such that* $|A_c\, x - B_c| \neq 0$.

This condition rules out singularities in the matrix polynomial. Its necessity is exemplified by the discussion in Klein (2000, section 2.1). Suppose it doesn't hold, so that there exists a matrix polynomial $a(x)$ such that $a(x)'(A_c\, x - B_c) = 0$ for all $x \in \mathbb{C}$. Letting $x = L^{-1}$ be the forward shift operator, equation (1.5) would imply $a(L^{-1})'Q_c\, z_t = 0$, which is not true for a generic process z_t and matrix Q_c.

An additional condition required for our solution to work is the following.

Assumption 3. *The matrix of right generalized eigenvectors* V *in (1.9) is such that* V_{22} *is of full rank* N_f.

We should emphasize that this condition is necessary for our proposed solution method to work, see below, but it is not clear whether it is necessary for the existence of a solution in general. As noted also in the discussion of Klein (2000, Section 5.3.1), the Schur decomposition of a matrix pair is non-unique, and we cannot rule out the possibility that the above condition would be satisfied by a different choice of basis V for the generalized right eigenspace of (A_c, B_c).

Next, drawing on Klein (2000, section 5.2), we find the solution to the transformed model (1.13). Since the generalized eigenvalues of the matrix pair (S_{22}, T_{22}) are all explosive, and assuming this eigenproblem is regular (ruling out $S_{ii} = T_{ii} = 0$), T_{22} is invertible (no zeros on the diagonal) and the infinite sum $\sum_{j=0}^{\infty}(T_{22}^{-1}S_{22})^j$ converges. The solution for the explosive part $\mathbf{w}_t^e$ then follows by forward substitution:

$$\begin{aligned} \mathbf{w}_t^e &= -T_{22}^{-1}\sum_{j=0}^{\infty}\left(T_{22}^{-1}S_{22}\right)^j U_2'\, Q_c \mathrm{E}(z_{t+j}|\mathcal{F}_t) + \lim_{\tau\to\infty}\left(T_{22}^{-1}S_{22}\right)^{\tau}\mathrm{E}(\mathbf{w}_{t+\tau}^e|\mathcal{F}_t) \\ &= -T_{22}^{-1}\sum_{j=0}^{\infty}\left(T_{22}^{-1}S_{22}\right)^j U_2'\, Q_c\, \left[\Phi^j\,(z_t - \mu_z) + \mu_z\right] \\ &= M\, z_t + M_c\, c_z, \end{aligned} \tag{1.14}$$

where

$$\begin{aligned} \mu_z &= (I - \Phi)^{-1} c_z \\ \mathrm{vec}\,(M) &= \left[(\Phi' \otimes S_{22}) - (I_m \otimes T_{22})\right]^{-1} \mathrm{vec}\,(U_2' Q_c)\,, \quad \text{and} \\ M_c &= \left[-T_{22}^{-1}\,(I_{N_e} - T_{22}^{-1}\, S_{22})^{-1}\, U_2'\, Q_c - M\right](I - \Phi)^{-1}. \end{aligned}$$

The derivation of M follows from the fact that

$$\sum_{j=0}^{\infty} \mathrm{vec}\left(A^j B C^j\right) = \sum_{j=0}^{\infty} (C' \otimes A)^j \mathrm{vec}\,(B) = \left[I - (C' \otimes A)\right]^{-1} \mathrm{vec}\,(B)\,.$$

The only other condition required for the above result is a bound on the growth of $\mathrm{E}(z_{t+\tau}|\mathcal{F}_t)$ so that

$$\lim_{\tau\to\infty} \left(T_{22}^{-1} S_{22}\right)^{\tau} \mathrm{E}(\mathbf{w}^e_{t+\tau}|\mathcal{F}_t) = \lim_{\tau\to\infty} \left(T_{22}^{-1} S_{22}\right)^{\tau} M\, \mathrm{E}(z_{t+\tau}|\mathcal{F}_t) = 0$$

This is clearly satisfied under assumption 1 even when the process z_t contains deterministic polynomial trends (but it is satisfied by some explosive processes as well).

On the other hand, the non-explosive part can be solved by backward substitution:

$$\begin{aligned} \mathbf{w}^s_{t+1|t} &= S_{11}^{-1}\, T_{11}\, \mathbf{w}^s_t + S_{11}^{-1}\, T_{12}\, \mathbf{w}^e_t - S_{11}^{-1}\, S_{12}\, \mathbf{w}^e_{t+1|t} + S_{11}^{-1}\, U_1'\, Q_c \\ &= S_{11}^{-1}\, T_{11}\, \mathbf{w}^s_t + S_{11}^{-1}\, \left(T_{12}\, M\, - S_{12}\, M\, \Phi + U_1'\, Q_c\right)\, z_t \end{aligned}$$

Using (1.10), defining the expectational error $\eta_{t+1} = \mathbf{x}^f_{t+1} - \mathbf{x}^f_{t+1|t}$ and noting that the $\mathbf{x}^p_{t+1} = \mathbf{x}^p_{t+1|t}$ by predeterminedness, the above yields:

$$\mathbf{w}^s_{t+1} = S_{11}^{-1}\, T_{11}\, \mathbf{w}^s_t + S_{11}^{-1}\, \left(T_{12}\, M\, - S_{12}\, M\, \Phi + U_1'\, Q_c\right)\, z_t + V_{21}'\, \eta_{t+1} \tag{1.15}$$

The solution can then be found by backward substitution, given N_s initial conditions $\mathbf{w}^s_0$. However, when there is a unique solution to the LiRE model, the solution to the explosive part (1.14) will suffice to derive it, as we show below, so there will be no need to solve the non-explosive part.

Now, we turn to the solution to the canonical LiRE model (1.5). There are three cases, which are outlined in the following result.

Proposition A.1. *Under assumptions 1, 2 and 3, we distinguish three cases for the solution to the canonical LiRE model (1.3):*

1. *If $N_e > N_f$, there exists no solution.*
2. *If $N_e = N_f$, there exists a unique solution.*
3. *If $N_e < N_f$, there exist infinitely many solutions.*

Proof. Equation (1.11) implies:

$$V_{22}'\, \mathbf{x}^f_t = -V_{12}'\, \mathbf{x}^p_t + M\, z_t + M_c\, c_z. \tag{1.16}$$

Since $\mathbf{x}^p_t$ and z_t are given, these are N_e equations in N_f unknowns. When $N_f < N_e$ no solution exists for all t, unless the square N_e-dimensional matrix $(V_{22}',\ -V_{12}'\, \mathbf{x}^p_t + M\, z_t)$ is of reduced rank

N_f for all t, which implies that $N_e - N_f$ equations are redundant. But this in turn requires that $\mathbf{x}_t^p, z_t$ are constant across time, which is not true.

When $N_f = N_e$, there exists a unique solution, which, by assumption 3, is:

$$\mathbf{x}_t^f = -(V_{22}')^{-1}\, V_{12}'\, \mathbf{x}_t^p + (V_{22}')^{-1}\, M\, z_t + (V_{22}')^{-1}\, M_c\, c_z. \tag{1.17}$$

When $N_f > N_e$, there are infinite solutions, since there are more unknowns than equations in (1.16). It may seem that the extra $N_f - N_e$ non-predetermined variables can be pinned down by appealing to the solution of the non-explosive part (1.15) and using (1.10). However, $N_f - N_e$ linear combinations of the expectational error η_{t+1} remain indeterminate up to a martingale difference sequence, giving rise to a multiplicity of solutions. □

1.1.3 Solving the original model

Given a solution for $\mathbf{x}_t^f$ found above and using the definition of $\mathbf{x}_t$ in (1.7), we can infer the solution for the endogenous variables y_t as a function of its lags and z_t:

$$y_t = \sum_{i=1}^{k} \Pi_i\, y_{t-i} + Q_s\, z_t + c \tag{1.18}$$

where

$$\begin{pmatrix} \Pi_k & \ldots & \Pi_1 \end{pmatrix} = \left(-(V_{22}')^{-1}\, V_{12}'\right)_{[:n][\cdot]}, \quad Q_s = \left((V_{22}')^{-1}\, M\right)_{[:n][\cdot]} \quad \text{and} \quad c = \left((V_{22}')^{-1}\, M_c\right)_{[:n][\cdot]}\, c_z.$$

The notation $(X)_{[:n][\cdot]}$ denotes selecting the first n rows of a matrix X. In `PcLiRE`, the solution is implemented by the function `PcLiRE::Solve()`, or `PcLiRE::Print()`, as exemplified below.

-------------------------- Sample Code ----------------------------

```
#include "PcLiRE1.ox"
main()
{
    decl cY, model, mA1, mA0, mAb1;
    cY = 2;                         // number of endogenous variables.
    model = new PcLiRE(cY);         // create the PcLiRE object.
    mA1 = <-0.591, 0; 0, 0>;        // specify coefficients.
    mA0 = <1, -0.05; 0, 1>;
    mAb1 = <-.378, 0; 0.1, -0.9>;
    model.SetYParameter({mA1, 1, 0}, {mAb1, -1, 0}, {mA0, 0, 0});
                                    // set LiRE equation parameters.
    model.Print();
}
```

- Output -

```
------------------ PcLiRE (1.0) -------------------
y is (2 x 1).

The structural model is:

A0 y[t] + A1 y[t+1|t] + Ab1 y[t-1] = z[t]

z[t] = u[t]

A0 =
       1.0000    -0.050000
      0.00000       1.0000
A1 =
     -0.59100      0.00000
      0.00000      0.00000
Ab1 =
     -0.37800      0.00000
      0.10000     -0.90000
u ~ N(0,1)

The solved model is a VAR(1):

y[t] = Pi1 y[t-1] + Qs z[t] Pi1 =
      0.51169      0.27159
     -0.10000      0.90000
Qs =
       1.4335      0.30176
      0.00000       1.0000
```

- -

Verifying the solution

The solution coefficients Π_i and Q_s, satisfy some restrictions, the validity of which could provide a verification of the proposed solution. In particular, abstracting from the deterministic term c, we can write (1.18) in companion form, using $Y_t = (y_t', \ldots, y_{t-k+1}')'$:

$$Y_t = \Pi\, Y_{t-1} + \begin{pmatrix} Q_s \\ 0 \end{pmatrix} z_t,$$

where Π is the usual square companion matrix of dimension equal to nk. Define $\Pi^j_{i_1,i_2} = (\Pi^j)_{i_1,i_2}$, namely the (i_1, i_2) $n \times n$ block of the $nk \times nk$ matrix Π^j, where $i_1, i_2 = 1, \ldots, k$. Thus,

$$y_{t+j|t} = \sum_{i=1}^{k} \Pi^j_{1i}\, y_{t-i+1} + Q_s\, \Phi^j\, z_t$$

Substituting into the LiRE model (1.3) we have:

$$\sum_{i=0}^{k} A_{-i}\, y_{t-i} + \sum_{j=1}^{l} A_j \sum_{i=1}^{k} \Pi^j_{1i}\, y_{t-i+1} = \left(Q - \sum_{j=1}^{l} A_j\, Q_s\, \Phi^j \right) z_t.$$

By defining

$$\begin{aligned} B_0 &= A_0 + \textstyle\sum_{j=1}^{l} A_j \Pi_{11}^{j}, \\ B_{i-1} &= A_{-i+1} + \textstyle\sum_{j=1}^{l} A_j \Pi_{1i}^{j}, \quad i = 2, \ldots, k \\ B_k &= A_{-k} \end{aligned}$$

the above can be re-written as:

$$\sum_{i=0}^{k} B_i \, y_{t-i} = \left(Q - \sum_{j=1}^{l} A_j \, Q_s \, \Phi^j \right) z_t \tag{1.19}$$

Hence, the restrictions implied by Rational Expectations can be found by matching the coefficients in (1.18) and (1.19):

$$\begin{aligned} B_0 \, \Pi_i + B_i &= 0, \qquad i = 1, \ldots, k, \qquad \text{and} \\ B_0 \, Q_s + \textstyle\sum_{j=1}^{l} A_j \, Q_s \, \Phi^j &= Q \end{aligned}$$

Other issues

Unit roots The presence of unit roots in the LiRE polynomial $A(\lambda)$ (or equivalently, the canonical polynomial $A_c \, \lambda - B_c$) does not pose any problem for the unique solution given in (1.18) insofar as there are as many explosive roots as there are non-predetermined variables. Any additional unit roots in the non-explosive part will typically be present in the solution for y_t, which will be integrated. Although we give no proof of this conjecture, it can be easily checked in any given application by looking at the eigenvalues of the companion matrix Π. These can be reported by the program at the user's request, using the function `PcLiRE::CointAnalysis()`, as shown in the following example, whereby we also print out the roots of the LiRE polynomial $A(L)$.

---------------------------- Sample Code ----------------------------

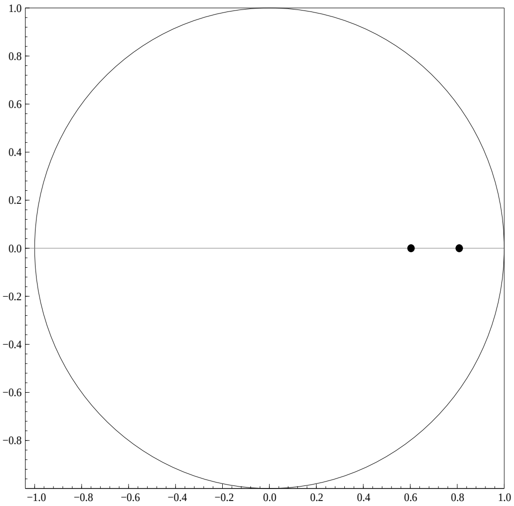
```
#include <oxstd.h>
#include "PcLiRE1.ox"

main()
{
    decl cY, model, mA1, mA0, mAb1;
    model = new PcLiRE(2);          // create the PcLiRE object.
    mA1 = <-0.4, 0; 0, 0>;          // specify coefficients.
    mA0 = <1, -0.05; 0, 1>;
    mAb1 = <-.6, 0; 0.1, -0.9>;
    model.SetYParameter({mA1, 1, 0}, {mAb1, -1, 0},
        {mA0, 0, 0});   // set LiRE equation parameters.

    model.PrintRoots(TRUE);
    model.Solve();
    model.CointAnalysis();
}
```

------------------------------ Output ------------------------------

```
Roots of PcLiRE polynomial:
      alpha_r      alpha_i         beta      modulus
       1.7546      0.24542       1.9006      0.93216
       1.7546     -0.24542       1.9006      0.93216
     -0.95586      0.00000      0.00000  1.0797e+032
      0.72868      0.00000      0.46901       1.5537
Eigenvalues of companion matrix Pi
         Real    Imaginary      Modulus
      0.92317     -0.12913      0.93216
      0.92317      0.12913      0.93216
```



Plot of the eigenvalues of the companion matrix.

- -

Notice the role of the φ parameter in the example (1.4). When $\varphi = 0$, together with the restriction $\gamma_f + \gamma_b = 1, \gamma_f < \gamma_b$, the system (inflation) has a unit root. This was encountered also in section 4.2.1. In that case, the output of PcLiRE is:

- Output -

```
Roots of PcLiRE polynomial:
        alpha         beta      modulus
      0.82462      0.82462       1.0000
      0.94420       1.0491      0.90000
      0.72904      0.48603       1.5000
      0.95131      0.00000  1.2809e+016
Eigenvalues of companion matrix Pi
       1.0000
      0.90000
There are 1 unit roots. Long-run matrix
      0.00000      0.18750
      0.00000     -0.10000
alpha =
     -0.18750
      0.10000
beta' =
      0.00000      -1.0000
```

- -

Deterministic terms In the above exposition, we abstracted from deterministic terms in the LiRE equation (1.3). Such terms pose no additional considerations, since they can be incorporated in the driving process z_t, provided the coefficient matrix Q is of rank n. In our implementation, we allow for a constant by noting that:

$$\begin{aligned} \mathrm{E}(A(L)y_t|\mathcal{F}_t) &= c_y + Q\, z_t = Q\, z_t^* \\ z_t^* = z_t + \mu. \end{aligned}$$

When $m = n$, $\mu = Q^{-1}\, c_y$, whereas when $m > n$, $\mu = (c_y'(Q_1^{-1})', 0')'$, where Q_1 is a non-singular $n \times n$ submatrix of the $n \times m$ matrix Q. A constant vector can be added using the command `PcLiRE::SetDeterministic(vc)`, where `vc` is a $n \times 1$ matrix.

However, the role of the constant must be interpreted carefully when there are unit roots in the solution polynomial $\Pi(L)$. It will contribute partly a non-zero mean in the co-integrated combinations of y_t and partly a linear trend in y_t, see Johansen (1995).

Back-dated expectations When the LiRE model contains back-dated expectational terms, such as $\mathrm{E}(y_{t+1}|\mathcal{F}_{t-1})$, the solution provided by proposition A.1 is no longer sufficient. The reason is that it is not adequate to express the non-predetermined variables $\mathbf{x}_t^f$ as a function of predetermined ones $\mathbf{x}_t^p$, since the latter will contain expectational terms such as $\mathrm{E}(y_{t+1}|\mathcal{F}_{t-1})$, which have not been specified. In that case, we have to look at the backward-solution of the non-explosive part, $\mathbf{w}_t^s$, given in (1.15). However, as we noted earlier, that solution relies on the presence of a sufficient number N_s of initial conditions, which requires the specification of initial conditions for expectational terms such as $y_{2|0}$.

This difficulty can be resolved when the difference equation (1.15) is stable, by finding a distribution for the initial conditions $\mathbf{w}_0^s$ such that the y_t process becomes stationary. However, it is non-trivial to transform the final solution, which is initially expressed as an infinite distributed lag of z_t, into a finite-order VAR. This is work-in-progress, and is left for a future implementation.

Simulation

For simulation studies, we need to specify a distribution for the forcing variables z_t, and then draw samples from the resulting distribution of y_t. This can be done in `PcLiRE` directly using using the commands:

`PcLiRE::SetUDistribution(iDistType,vDf)`

`PcLiRE::SetUParameter(mSigma)`

`PcLiRE::SetUParameter(mSigma), and`

`PcLiRE::Generate(cT).`

The first one determines the type of distribution of the innovations u_t. Currently, the program allows multivariate normal (`iDistType = MVNORMAL`) or Student's t (`iDistType = T_DIST` and `vDf` is a vector of positive integers, carrying the degrees of freedom for each of the variables in u_t). The second function is used to set the covariance matrix of the innovations. Importantly, this can be singular (can contain zeros on the diagonal) so that we can simulate models where the stochastic variation is of lower dimension than the system. The last function returns a matrix Y of dimension $T \times n$. Concerning the initial conditions, the program draws Z_0 randomly from a distribution that would make z_t stationary. We do not use the same approach in the case of Y_0, since it is unjustified when unit roots are present in the y_t equation. Following the approach in PcNaive, Y_0 is set to zero by default.

Alternative simulation using PcNaive The above simulations can be performed also using the `PcNaiveDgp` class, which comes with the standard Ox 3.2 release, see Doornik and Hendry (1998) for details. For that purpose, the user can recover the coefficients of the solved model using the function `PcLiRE::GetSolutionPar()` in order to use them in PcNaive. The example below shows how 100 observations can be drawn from the previous model directly using `PcLiRE`, or indirectly using `PcNaiveDgp`.

-------------------------- Sample Code ---------------------------

```
#include <oxstd.h>
#include "PcLiRE1.ox"
#import <pcnaive>

main()
{
    decl model, mA1, mA0, mAb1;
    model = new PcLiRE(2);                          // create the PcLiRE object.
    mA1 = <-0.591, 0; 0, 0>;                            // specify coefficients.
    mA0 = <1, -0.05; 0, 1>;
    mAb1 = <-.378, 0; 0.1, -0.9>;
    model.SetYParameter({mA1, 1, 0}, {mAb1, -1, 0},
        {mA0, 0, 0});                           // set LiRE equation parameters.
    model.Solve();

/*  direct simulation:  */
    model.SetUParameter(unit(2));                                // variance of u.
    model.SetUDistribution(MVNORMAL, 0);                     // distribution of u.
    decl mdata;
    mdata = model.Generate(100);
```

```
/*  indirect simulation: */
    decl mQ, mPi;
    [mQ, mPi] = model.GetSolutionPar();                         // in array {Q, Pi}.
    decl dgp = new PcNaiveDgp(2,0);                       //create PcNaiveDgp object.

    dgp.SetYParameter(zeros(2,2), mPi, 0, zeros(2,1));
    dgp.SetDistribution(U_DGP, MVNORMAL, zeros(2,1), mQ*mQ');
    mdata = dgp.Generate(100);
}
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

1.2 GMM estimation using Ox

The `Gmm` class is derived from the Modelbase class in Ox, and is designed to perform Generalized Method of Moments estimation, given a set of data and specified moment conditions. It supports multi-equation estimation and performs a number of recently developed diagnostic tests for model identification and mis-specification. The current exposition is brief and it is not intended to be a detailed user manual for the program. Description of the underlying Modelbase commands for loading data and setting up models for estimation can be found in the standard Ox documentation. We comment briefly on the main functions of the program and give an example of its use.

The program works as follows. The user first has to load the relevant data file. Then, the empirical model must be set up by selecting variables into one of the following three groups:

Group name	Description	Class data member
`Y_VAR`	regressand(s) – multiple equations allowed	`m_mY`
`X_VAR`	regressors – endogenous and exogenous	`m_mX`
`Z_VAR`	instruments	`m_mZ`

When the same variable is included in both `X_VAR` and `Z_VAR` groups, it is an ‘exogenous’ regressor. The selection is done using the function `Gmm::Select(iGroup, aVar)`, as shown in the sample code below. The program casts the variables of each group into the class data members `m_mY`, `m_mX` and `m_mZ`, respectively.

An important part of the modelling procedure is setting up the orthogonality conditions to be used in the estimation. As explained at the beginning of chapter 2, these moment conditions are derived by the requirement that certain G ‘moment functions’ $h(\mathrm{y}_t, \theta)$ are orthogonal (in expectation) to the instruments Z_t, at some parameter value θ_0. These moment functions can sometimes admitted the interpretation of structural residuals, e.g., $h(\mathrm{y}_t, \theta) = y_t - \theta' Y_t$. The empirical moment conditions are then derived by averaging these moment conditions per observation

$f_t(\theta) = h(\mathrm{y}_t, \theta) \otimes Z_t$, i.e.,

$$g_T(\theta) = \sum_{t=1}^{T} f_t(\theta)$$

The corresponding class member functions for these are:

$h(\mathrm{y}_t, \theta)$	`Gmm::MomentFunctions(mData, vPar)`
$g_T(\theta)$	`Gmm::MomentConditions(vPar)`

By default, the $h_t(\theta)$ is linear, but `Gmm::MomentFunctions(vPar)` is "virtual", which, in programming jargon, means that it can be overridden by the user to setup non-linear moment conditions.

Then, the estimation settings must be specified, namely the type of estimator, the optimization method (for non-linear optimization), the covariance matrix estimator. Optionally, parameter names and initial parameter values can be supplied to aid presentation and improve the chance of convergence. We present each one of these in turn.

Estimators The type of estimator is determined by the function `Gmm::SetMethod(iType)`. The following estimators are implemented:

Code name	Estimator
`M_1STEP`	1-step (inefficient)
`M_2STEP`	2-step
`M_MSTEP`	Iterated
`M_CUE`	Continuously updated

One-step efficient Generalized Empirical Likelihood estimators, such as exponential tilting, are not implemented in the current version.

Covariance estimators The user must specify the type of robust covariance estimator, the number of lags for the autocorrelation correction, whether a finite-sample degrees-of-freedom correction is required, and the type of centering as explained in section 2.3. This is done by the function `Gmm::SetVariance(bSmallSample, iCovar, clags, iCenter)`. The following HAC estimators of the GMM weighting matrix are available:

HAC type (`iCovar`)	Description
`COV_NONE`	homoscedastic, serially uncorrelated
`COV_HC_W`	heteroscedasticity consistent, White
`COV_AC_NW`	autocorrelation consistent, Newey-West
`COV_HAC_NW`	heteroscedasticity and autocorrelation consistent, Newey-West
`COV_HAC_W`	heteroscedasticity and autocorrelation consistent, West (MA-l)

Centering (`iCenter`)	Description
`CRN_NONE`	un-centered (default)
`CRN_CONDITIONAL`	conditionally centered
`CRN_UNCONDITIONAL`	unconditionally centered

If `clags = -1`, the lags for the Newey-West estimator are estimated automatically from the data, otherwise they are fixed to `clags`.[5]

Inference The following is a list of tests implemented by the program:

Function	Description
`AndersonRubinTest(vPar0)`	Anderson-Rubin statistic $\mathcal{Q}_T(\theta_0, \theta_0)$.
`LikelihoodRatioTest(vPar0)`	Likelihood ratio test of $\mathrm{H}_0 :\ \theta = \theta_0$.
`HansenTest()`	Hansen-Sargan test of over-identifying restrictions.
`SerialCorrelationTest(q,s)`	Excess serial correlation test, $\mathrm{SC}(q, s)$.
`IdentificationTest()`	QMD and RR identification tests.
`IdentificationTest(crank)`	Returns QMD test only for specific reduced-rank hypothesis.

The identification test is only available for linear single-equation models (though it can be done equation by equation for multivariate models).[6]

Example The relevant sample Ox code for the estimation of the Galí and Gertler (1999) model is given below.

-------------------------- Sample Code ---------------------------

```
#include "gmm.ox"

main()
{
    decl GGmodel = new Gmm();                                   // create GMM object.
    GGmodel.LoadIn7("GGdata");                                     // load the data.
    GGmodel.Deterministic(FALSE);                          // to create constant term.

/*  Set model  */
    GGmodel.Select(Y_VAR, {"INFL", 0, 0});                         // dependent var.
```

---

[5] West's MA-l estimator requires that the moment functions be passed through an MA filter. This is currently only available for single-equation models, G= 1, and is done using the Arfima package, developed by Jurgen Doornik and Marius Ooms, see `http://www.nuff.ox.ac.uk/Users/Doornik/papers/arfima.pdf`.

[6] The development of an identification test for non-linear models, based on the Jacobian of the moment conditions, is still work-in-progress.

```
    GGmodel.Select(X_VAR, {"Constant", 0, 0, "INFL", 1, 1,            // regressors, exog.
                             "INFL1" , 0, 0, "share", 0, 0});      // ... and endogenous.
    GGmodel.Select(Z_VAR, {"Constant", 0, 0, "INFL", 1, 4,                 // instruments.
                           "share" , 1, 4, "dw", 1, 4,
                           "dc", 1, 4, "ygap", 1, 4, "spr", 1, 4});

/*  Estimation settings  */
    GGmodel.SetSample(1960, 1, 1997, 4);                             // estimation sample.
    GGmodel.SetMethod(/*estimator*/ M_2STEP);                              // 2-step GMM.
    GGmodel.SetVariance(/*Small Sample?*/ FALSE,            // no small sample correction.
                        /*HAC type*/ COV_HAC_W,                            // West's MA-l.
                        /*lags*/ 1,                           // with 1-lag AC correction.
                        /*centering*/ CRN_NONE);                           // un-centered.
    GGmodel.SetParNames({"Constant", "lambda", "gamma_f", "gamma_b"});

    GGmodel.Estimate();

/*  Do additional tests */
    GGmodel.HansenTest();
    GGmodel.IdentificationTest();

    delete GGmodel;                                      // delete GMM object at the end.
}
```

```
---------------------------------------  Output  ---------------------------------------
------------------ Gmm --------------------
The estimation sample is:  1960 (1) - 1997 (4)
The dependent variable is: INFL
Linear 2-step GMM estimation with 1-lag HAC correction (West's MA-l).

                  Coefficient  Std.Error  t-value  t-prob
Const              0.00500599   0.007728    0.648   0.518
lambda              0.0902515    0.02845     3.17   0.002
gamma_f              0.644994    0.03521     18.3   0.000
gamma_b              0.345349    0.03357     10.3   0.000

GMM objective        -0.11082  no. of instruments         25
no. of observations       152  no. of parameters           4
mean(INFL)            1.04107  var(INFL)            0.382333

Instruments used: Constant, INFL_1, INFL_2, INFL_3, INFL_4,
share_1, share_2, share_3, share_4, dw_1, dw_2, dw_3, dw_4, dc_1,
dc_2, dc_3, dc_4, ygap_1, ygap_2, ygap_3, ygap_4, spr_1, spr_2,
spr_3, spr_4

Hansen test of overidentifying restrictions: Chi^2(21)= 16.1175 [0.763013]

Identification rank test

        rank eigenvalues        DM        LR [pval-DM] [pval-LR]
     0.00000     0.78908  645.3474  365.7980    0.0000    0.0000
      1.0000     0.57271  170.2192  129.2430    0.0000    0.0000

-----------------------------------------------------------------------------------------
```

1.3 Local-to-Zero asymptotic approximations

We developed a class to compute the local-to-zero approximations given by theorem 2.1 and used in the Monte Carlo experiments of chapter 3, called `LtZasymptotics`. The code is the implementation of the computational procedures outlined in appendix 3.B.

First, we need to specify the parameters of the DGP, which is a slight generalization of (3.1), i.e.,

$$\begin{aligned} \mathrm{y}_t &= A_1\,\mathrm{y}_{t-1} + A_2\,\mathrm{y}_{t-1} + A_z\mathbf{z}_t + \mathbf{e}_t \\ \mathbf{z}_t &= C\,\mathbf{z}_{t-1} + \mathbf{v}_t \end{aligned}$$

where $\mathbf{e}_t \perp v_t$ and $(\mathbf{e}_t, \mathbf{v}_t)$ is a vector mean innovation process w.r.t. the history of y_t and z_t. The distinction of the strongly exogenous variables $\mathbf{z}_t$ is done mainly for convenience, since the model could have been equivalently written as a closed VAR. (the program re-writes it this way in order to compute the unconditional second moment of all the variables). The parameters of the DGP are $\{A_1, A_2, C, \Sigma_e, \Sigma_v\}$. These must be specified both at their 'limiting' values and also at their values depending on the sample size. This is done by the functions:

```
SetDgp(mA1, mA2, mAz, mC, mSigmaE, mSigmaV)
SetDgpT(mA1_T, mA2_T, mAz_T, mC_T, mSigmaE_T, mSigmaV_T, cT)
```

where at least one of the finite-sample coefficient matrices in the second function (e.g. `mA1_T`) is different from its corresponding limiting value (`mA1`) in the first function above.

Then, we must specify a (single-equation) structural model to be estimated by GMM of the form

$$y_t = \theta'\,\bar{Y}_t + u_t$$

where $\bar{Y}_t = (Y_t', Z_{1t}')'$ are all the endogenous and exogenous regressors, and a set of instruments Z_t taken from $(\mathrm{y}_t, \mathbf{z}_t)$ and their history. This is done with the following function:[7]

`SetModel(vYindex, vXindex, vZ1index, vZindex)`, or

`SetModel(vYindex, vXindex, vZ1index, vZindex, mSelection)`.

This works as follows. All of the variables are stacked in a vector of dimension (`cy+cz`) $*$ `cMaxLag`. For example, suppose `cy` $= 2$ and `cz` $= 1$, and that we need up to the fourth lag of some of the variables as an instrument, i.e., `cMaxLag` $= 4$. Then,

$$\begin{array}{rcccccccccccl} \mathbf{w}_t' & = & \big(\mathrm{y}_{1,t} & \mathrm{y}_{2,t} & \mathbf{z}_t & \mathrm{y}_{1,t-1} & \mathrm{y}_{2,t-1} & \mathbf{z}_{t-1} & \ldots & \mathrm{y}_{1,t-4} & \mathrm{y}_{2,t-1} & \mathbf{z}_{t-4}\big) \\ \text{index} & : & [0 & 1 & 2 & 3 & 4 & 5 & \ldots & 12 & 13 & 14] \end{array}$$

---

[7] For the definition and role of the (optional) selection matrix `mSelection`, see appendix 3.B.

Based on this, we need to specify an index vector for each variable group, picking out the relevant entries of $\mathbf{w}_t$. For example, $y_t = \mathrm{y}_{1t}$ would correspond to `vYindex = <0>`, $Y_t = \mathrm{y}_{2t}$ would be `vXindex = <1>`, no exogenous regressors `vZ1index = <>`, and the instruments $Z_t = (\mathrm{y}_{1,t-1}, \ldots \mathrm{y}_{1,t-4}, \mathbf{z}_{t-1}, \ldots \mathbf{z}_{t-4})$ would be `vZindex = <3,6,9,12,5,8,11,14>`.

Finally, we need to specify the lags of any serial correlation correction, the sample size, and the type of GMM estimator, 1-step, 2-step or CUE.

Then, the program will return the implied probability limits of the particular GMM estimators, together with the corresponding plims of the OLS estimators, the endogeneity parameter λ, the eigenvalues of the concentration parameter at the particular sample size T, the mis-specification parameter ζ_T, the well-identified and weakly identified parameter combinations (d and $d_\perp$ in the terminology of appendix 3.B), and the autocovariance function (ACF) of the structural residual u_t.

Additionally, the program can plot the concentrated limiting objective function (as a function of the weakly identified parameters only), to check for multiple optima or non-existence of a global minimum. This is particularly useful for the CUE, see figure 3.5 in chapter 3.

The following sample code does the asymptotic analysis for the GG hybrid new Phillips curve model under case 3 in section 4.2.2. The DGP is such that the model is correctly specified (the parameters are derived after having solved the original LiRE model, with s_t being AR(2)).

-------------------------- Sample Code ----------------------------

```
#include <oxstd.h>
#include "LtZasymptotics.ox"

main()
{
    decl cy = 2, cz = 0, cT = 150;

    decl mSigma, mSigmaT, mA1, mA1T, mA2, mA2T;

    mA1 = <0.57005, 0.10290;
           0.00000, 0.90000>;
    mA1T = <0.57005, 0.081450;
            0.00000, 0.90000>;
    mA2 = zeros(2,2);
    mA2T = <0.00000, -0.0047608;
            0.00000, -0.050000>;

    mSigma =  <1.5081, 0.11433;
              0.00000, 1.0000>;
    mSigma = mSigma*mSigma';

    mSigmaT =  <1.5081, 0.095215;
                0.00000, 1.0000>;
    mSigmaT = mSigmaT*mSigmaT';
```

```
    decl asy = new LtZasymptotics(/*MaxLag*/ 3,/*Precision*/ 1e+5);

    asy.SetDgp(mA1, mA2, /*mAz*/<>, /*mC*/<>, mSigma, /*mSigmaZ*/<>);
    asy.SetDgpT(mA1T, mA2T, <>, <>, mSigmaT, <>, cT);
    asy.SetModel(/*vYindex*/<2>, /*vXindex*/<3,0>, /*vZ1index*/<4>, /*vZindex*/<4:11>);

    asy.GetAllParam();
    asy.EstiMethod(M_2S);
    asy.Print();
}
```

```
------------------------------ Output ------------------------------
Implied param    Ols plims
     0.015000     0.042300
      0.59110      0.42840
      0.37800      0.43080
Endogeneity
      0.21107
Eigenvalues of Concentration
       5790.2  1.4619e-005        +.Inf
Mis-specification: 0 Well-identified parameter combinations
      0.14677      0.32793      0.93323
      0.97650      0.10252     -0.18959
Weakly identified parameter combinations
      0.15784     -0.93912      0.30517
Weakly ident. true
     -0.43742
Error ACF
       1.7991     -0.89140      0.00000      0.00000
--------------------------------------------------------------------
```

As expected, the implied GMM parameter estimates are the same as the true values, see top line of table 4.1, since there is no mis-specification. One of the eigenvalues of the concentration matrix is 'infinite' corresponding to the exogenous regressor π_{t-1}, and one of the eigenvalues is very small ($\ll 0.1$), revealing weak identification. Evidently, the well-identified parameter space

$$d' = \begin{pmatrix} 0.53235 & 0.51624 & 0.67089 \\ -0.70392 & -0.17027 & 0.68957 \end{pmatrix}$$

doesn't contain any of the unit vectors (1,0,0) (0,1,0) or (0,0,1), thus showing that none of the parameters is well-identified on its own, consistent with the discussion of section 4.2.1. Finally, the structural error exhibits negative autocorrelation at lag 1 only, as anticipated.

1.4 Simulation

The code for the simulations in this thesis was developed based on the `Simulation` class in Ox, which is described in the standard Ox 3 documentation. The main simulation class that we developed, `IVGMM_simulator`, is 'derived' from the `Simulation` class, which, in programming jargon, means that it inherits all of the functionality and data members of the latter, but it is customized to the problem of interest.

The simulation code also uses the above three classes, `PcLiRE`, `Gmm` and `LtZasymptotics` as follows: `PcLiRE` is used to solve the structural model if necessary and the `PcNaiveDgp` class to sample from the resulting DGP; then, the `Gmm` class is used to get GMM estimators and test statistics; and, finally, the `LtZasymptotics` class is used to perform asymptotic analysis of the model, and draw the local-to-zero asymptotic approximations of chapter 3.

As an example, we give the sample code for the Monte Carlo simulation of the GG hybrid New Phillips curve model under case 3, in section 4.2.2. The code, which is written in detail so as to be more legible, spans several pages, and will therefore be presented in fragments:

1. All the relevant header files and auxiliary code must be included at the beginning.

```
...................................................................................
#include <oxfloat.h>
#include "PcLiRE1.ox"
#include "LtZasymptotics.ox"
#include "IVGMM simulator.ox"
...................................................................................
```

2. In the main function, we first set the different sample sizes for the simulation experiment.

```
...................................................................................
main()
{
    decl time = timer();

    decl mT = <100;200;500>;                             // sample sizes for simulation.
    decl it;                                                 // counter for each sample.

    for(it = 0; it < 3; ++it)                              // do experiment for every T.
    {
        decl lire, exp, asy;                        // class objects to be used below.

...................................................................................
```

3. Next, we specify the RE model (to be solved in order to get DGP). This is the hybrid GG model of chapter 4, with parameter values set as in the top line of table 4.1.

```
...................................................................................

    ///////////////////////////////////////////////////////////////////////////
    //  RE model to solve:

        decl mAf1, mA1, mA2T, mA0, mSigmaE;      // PcLiRE coefficient matrices.
        decl mPi1, mPi1T, mPi2T, mSigma, mSigmaT;  // Solution (DGP) parameters.
```

```
decl gammaf, gammab, lambda, rho1, rho2;         // GG model parameters.
gammaf = 0.591;
gammab = 0.378;
lambda = 0.05;
rho1 = 0.9;
rho2 = -0.05;
mSigmaE = diag(0.18^2 ~ 0.1^2);

mAf1 = -gammaf ~ 0 | <0, 0>;                    // specify coefficients.
mA0 = 1 ~ -lambda  | <0, 1>;
mA1 = -gammab ~ 0  | 0 ~ -rho1;
mA2T = <0, 0> | 0 ~ -rho2;                // second-order dynamics at T.

lire = new PcLiRE(2);         // create the PcLiRE object (2 variables).

lire.SetYParameter({mAf1, 1, 0}, {mA0, 0, 0},
                   {mA1, -1, 0});           // solve first-order system.

decl mQ;                                         // see PcLiRE solution.
lire.Solve();
[mQ, mPi1] = lire.GetSolutionPar();                 // store parameters.
mSigma = mQ*mSigmaE*mQ';

lire = new PcLiRE(2);
lire.SetYParameter({mAf1, 1, 0}, {mA0, 0, 0},
     {mA1, -1, 0}, {mA2T, -2, 0});   // next, solve second-order system.

lire.Solve();
[mQ, mPi1T, mPi2T] = lire.GetSolutionPar();         // store parameters.
mSigmaT = mQ*mSigmaE*mQ';
```

....

Notice that we have specified two completing models for s_t, the benchmark un-identified case 1, in which s_t follows an AR(1), and the identified case 3, in which s_t is an AR(2), with 'local-to-zero' second-order dynamics.

4. Next, we create the simulation object, and specify some initial settings.

....

```
///////////////////////////////////////////////////////////////////////////
//  Simulation Settings:

    decl cTdiscard = 10;        // number of initial observation to discard.
    decl cleads = 1;                  // number of leads in estimated model.

    decl vPar0 = gammab | gammaf | lambda;           // set true parameters.

    exp = new IVGMM_simulator();

    exp.Simulation(
            /*Sample Size*/ mT[it],
            /*Max sample to draw*/ max(mT) + cTdiscard + cleads,
            /*No of Replic.*/ 20000,
```

```
            /*Common Seed?*/ TRUE,
            /*What seed?*/ -1,
            /*P-values*/ <0.2,0.1,0.05,0.01>,
            /*True parameters */ vPar0);
```

....

5. We then need to specify the parameters of the `PcNaiveDgp` DGP.

....

```
////////////////////////////////////////////////////////////////////////////
//  Set DGP:

    exp->SetDGP(
            /*Y param.*/ {zeros(2,2), mPi1T, mPi2T,/*Exog*/ 0},
            /*Error param.*/ {0, 0, mSigmaT},
            /*Z param.*/ {0, 0},
            /*Fixed Z*/ TRUE);               // PcNaiveDgp-style parameters.
```

....

6. Next, we specify the model and the estimation settings.

....

```
////////////////////////////////////////////////////////////////////////////
//  Set Model:

    decl asParNames = {"gammab", "gammaf", "lambda"};   // parameter names.

    exp->SetModel(/*Var Names*/ {"INFL", "s"},
                  /*Y_VAR*/ {"INFL", 0, 0},
                  /*X_VAR*/ {"INFL", 1, 1, "INFL", -1, -1, "s", 0, 0},
                  /*Z_VAR*/ {"INFL", 1, 4, "s", 1, 4},
                  /*Par Names*/ asParNames);

////////////////////////////////////////////////////////////////////////////
//  Set Estimation settings (Gmm class parameters):

    exp.SetEstimation(/*iMethod*/ M_2STEP,                    //2-step GMM.
                      /*iOptiMethod*/ OM_BFGS,      // irrelevant (linear).
                      /*iCovar*/ COV_HAC_NW,             // Newey-West HAC.
                      /*clag*/ 12,                    // 12 lag-correction.
                      /*iCentering*/ CRN_NONE );           // no centering.
```

....

7. We then specify which statistics we want to simulate. In this example, we've chosen the Anderson-Rubin, Hansen-Sargan and Likelihood ratio statistics. The null hypothesis value θ_0 for these tests is set equal to the true parameter values, but could be different (e.g. zeros). We could also simulate t-tests, Wald tests for linear restrictions, the QMD identification test

and the excess SC test for which extra settings are required. Parameter estimators are simulated by default.

....

```
////////////////////////////////////////////////////////////////////////////
//  Statistics to simulate:

    exp.WhatToSimulate(/*vtTestIdx*/<0;0;0>,        // no t-tests for  pars.
        /*fARtest*/ TRUE,                                       // AR test.
        /*fHStest*/ TRUE,                           // Hansen-Sargan test.
        /*fLRtest*/ TRUE,                                       // LR test.
        /*fWtest*/ FALSE,                                // not a Wald test.
        /*fIdent*/ FALSE,                               // not the QMD test.
        /*fSCtest*/ FALSE);                              // not the SC test.

    exp.SetTestNames({"AR", "J", "LR"});            // set respective names.
    exp.SetCoefNames(asParNames);                 // coefficient names, too.

    exp.SetNullParam(vPar0);     // set parameter for H0 of the above tests.
                                           // could be different from vPar0.
```

....

8. We may also wish to perform asymptotic analysis, and derive local-to-zero asymptotic approximations. This can be done with the `LtZasymptotics` class, outlined above.

....

```
////////////////////////////////////////////////////////////////////////////
//  Asymptotics:

    asy = new LtZasymptotics(/*MaxLag*/ 3,/*Precision*/ 10000);
    asy.SetDgp(/*mA1*/ mPi1,
               /*mA2*/ zeros(2,2),
               /*mAz*/ <>,
               /*mC*/ <>,
               /*mSigmaU*/ mSigma,
               /*mSigmaZ*/ <>);

    asy.SetDgpT(/*mA1*/ mPi1T,
                /*mA2*/ mPi2T,
                /*mAz*/ <>,
                /*mC*/ <>,
                /*mSigmaU*/ mSigmaT,
                /*mSigmaZ*/ <>,
                /*cT*/ mT[it]);

    asy.SetModel(/*vYindex*/ <2>,
                 /*vXindex*/ <0,3>,
                 /*vZ1index*/ <4>,
                 /*vZindex*/ <4:11>);

    asy.GetAllParam();
    asy.EstiMethod(M_2S);
//  asy.GetSimulDistn();              // get simulated l-t-z approximation.
```

```
    //  asy.Store("asy");                      // save simulated distributions.
        asy.Print();
```

……

9. Finally, we specify storage and printing settings for the simulation results, and proceed with the simulation.

……

```
    ///////////////////////////////////////////////////////////////////////////
    //  Store and print settings:

        exp.SetStore(TRUE);                        // store recursive results.
    //  exp.SetPrintRep(1000);              // print message every 1000th repl.
        exp.SetPlotRep(1000);                 //plot results every 1000th repl.
        exp.PrintSettings();

    //////////////////////////////////////////////////////////////////////////
    //  Do simulation:

        exp->Simulate();
    //  exp.SaveIn7("filename");        // save simul. results in Givewin file.
        delete exp;                                    // delete simulator.
    }
    println("\nlapsed time: ", timespan(time),"\n");
}
```

……

Bibliography

Abel, A. B. and F. S. Mishkin (1983). An integrated view of tests of rationality, market efficiency and the short-run neutrality of monetary policy. *Journal of Monetary Economics 11*, 3–24.

Anderson, G. A. and G. R. Moore (1985). A linear algebraic procedure for solving perfect foresight models. *Economics Letters 17*, 247–252.

Anderson, T. W. (1977). Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequences. *Econometrica 45*, 509–518.

Anderson, T. W. (1984). *An introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons Inc. 2nd Edition.

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statistics 20*, 46–63.

Anderson, T. W. and T. Sawa (1979). Evaluation of the distribution function of the two stage least squares estimate. *Econometrica 47*(1), 163–182.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica 67*(3), 543–564.

Andrews, D. W. K. and J. C. Monahan (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica 60*(4), 953–966.

Bårdsen, G., E. S. Jansen, and R. Nymoen (2002, May). Testing the New Keynesian Phillips Curve. Working paper, Norges Bank, Oslo.

Basmann, R. L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *J. Amer. Statist. Assoc. 55*, 650–659.

Batini, N., B. Jackson, and S. Nickell (2000). Inflaction dynamics and the labour share in the UK. Discussion paper 2, External MPC Unit, Bank of England, UK.

Binder, M. and H. M. Pesaran (1995). Multivariate rational expectations models: A review and some new results. In Pesaran and Wickens (Eds.), *Handbook of applied econometrics*, Volume Macroeconomics, pp. 139–187. Basil-Blackwell.

Blanchard, O. J. and C. M. Kahn (1980). The solution of linear difference models under rational expectations. *Econometrica 48*, 1305–11.

Brown, B. W. and S. Maital (1981). What do economists know? An empirical study of experts' expectations. *Econometrica 49*, 491–504.

Buiter, W. and I. Jewitt (1985). Staggered wage setting with real wage relativities: Variations on a theme of Taylor. In A. Arbor (Ed.), *Macroeconomic Theory and Stabilization Policy*, pp. 183–199. University of Michigan Press.

Burnside, C. and M. Eichenbaum (1996). Small sample properties of GMM-based Wald tests. *Journal of Business and Economic Statistics 14*, 294–308.

Calvo, G. A. (1983). Staggered prices in a utility maximizing framework. *Journal of Monetary Economics 12*, 383–398.

Choi, I. and P. C. B. Phillips (1992). Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations. *Journal of Econometrics 51*, 113–150.

Clarida, R., J. Galí, and M. Gertler (1998). Monetary policy rules in practice: Some international evidence. *European Economic Review 42*, 1033–67.

Cragg, J. G. and S. G. Donald (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory 9*(2), 222–240.

Cragg, J. G. and S. G. Donald (1997). Inferring the rank of a matrix. *J. Econometrics 76*(1-2), 223–250.

Cumby, R. E. and J. Huizinga (1992). Testing the autocorrelation structure of disturbances in Ordinary Least Squares and Instrumental Variables regressions. *Econometrica 60*(1), 185–195.

Cumby, R. E., J. Huizinga, and M. Obstfeld (1983). Two-step two-stage least squares estimation in models with rational expectations. *J. Econometrics 21*(3), 333–355.

Davidson, R. and J. G. MacKinnon (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica 49*(3), 781–793.

Davidson, R. and J. G. MacKinnon (1994). *Estimation and inference in econometrics.* Oxford: Oxford University Press.

den Haan, W. J. and A. T. Levin (1997). A practitioner's guide to robust covariance matrix estimation. In *Robust inference*, pp. 299–342. Amsterdam: North-Holland.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*(5), 1161–1191.

Doornik, J. A. (2001). *Ox: An Object-Oriented Matrix Language* (4th ed.). London: Timberlake Consultants Press.

Doornik, J. A. and D. F. Hendry (1998). Monte Carlo simulation using PcNaive for Windows. Unpublished typescript, Nuffield College, University of Oxford.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica 65*(6), 1365–1387.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*(2), 277–304.

Ericsson, N. R. and D. F. Hendry (1999). Encompassing and rational expectations: How sequential corroboration can imply refutation. *Empirical Economics 24*, 1–21.

Fuhrer, J. C. and G. R. Moore (1995). Inflation persistence. *Quarterly Journal of Economics 110*, 127–159.

Fuhrer, J. C., G. R. Moore, and S. Schuh (1995). Estimating the Linear-Quadratic inventory model: Maximum Likelihood versus Generalized Method of Moments. *Journal of Monetary Economics 35*, 115–158.

Galí, J. and M. Gertler (1999). Inflation dynamics: a structural econometric analysis. *Journal of Monetary Economics 44*, 195–222.

Galí, J., M. Gertler, and J. D. López-Salido (2001). European inflation dynamics. *European Economic Review 45*, 1237–1270.

Hahn, J. and J. Hausman (2002). A new specification test for the validity of instrumental variables. *Econometrica 70*, 1517–1527.

Hall, A. R. (2000). Covariance matrix estimation and the power of the overidentifying restrictions test. *Econometrica 68*(6), 1517–1527.

Hall, A. R. and A. Inoue (2001). A Canonical Correlations interpretation of Generalized Method of Moments Estimation with applications to moment selection. Mimeo, North Carolina State University, USA.

Hall, A. R. and F. P. M. Peixe (2001). A consistent method for the selection of relevant instruments. Mimeo, North Carolina State University, USA.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics 14*, 262–280.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica 76*, 1251–1271.

Hayashi, F. and C. Sims (1983). Nearly efficient estimation of time series models with predetermined, but not exogenous, instruments. *Econometrica 51*(3), 783–798.

Hendry, D. F. (1983). Monte Carlo Experimentation. In Z. Grilliches (Ed.), *The Handbook of Econometrics*, Volume 2, pp. 939–976. North-Holland.

Hendry, D. F. (1988). The encompassing implications of feedback versus feedforward mechanisms in econometrics. *Oxford Economic Papers 40*, 132–149.

Hendry, D. F. (2000). Forecast failure, expectations formation, and the Lucas Critique. Mimeo, Nuffield College, Oxford.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Kleibergen, F. (2001). Testing parameters in GMM without assuming that they are identified. Discussion paper, Tinbergen Institute, Amsterdam.

Kleibergen, F. (2002). Two Independent Pivotal Statistics that test Location and Misspecification and add-up to the Anderson-Rubin Statistic. Discussion paper 064/4, Tinbergen Institute, Amsterdam.

Klein, P. (2000). Using the generalized Schur form to solve a multivariate linear rational expectations model. *Journal of Economic Dynamics and Control 0*, 0–0.

Krolzig, H. M. (2001). General-to-specific reductions of vector autoregressive processes. Economics discussion paper 2000-w34, Nuffield College, Oxford.

Krolzig, H. M. and D. F. Hendry (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control 25*, 831–866.

Lucas, R. E. J. (1976). Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (Eds.), *The Philips Curve and Labor Markets.*, Carnegie-Rochester Conference Series on Public Policy. Amsterdam: North-Holland.

Moreira, S. (2001). Tests with correct size when instruments can be arbitrarily weak. Manuscript, Department of Economics, UC Berkeley.

Moreira, S. (2002). A Conditional Likelihood Ratio Test for Structural Models. Manuscript, Department of Economics, UC Berkeley.

Nelson, C. R. and R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica 58*(4), 967–976.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics*, Volume 4, pp. 2111–2245. North-Holland.

Newey, W. K. and K. D. West (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Newey, W. K. and K. D. West (1994). Automatic lag selections in covariance matrix estimation. *Review of Economic Studies 61*(3), 631–653.

Pagan, A. and J. Robertson (1997). GMM and its problems. Mimeo, Australian National University, Canbera.

Phillips, P. C. B. (1977). Approximations to some finite sample distributions associated with a first-order stochastic difference equation. *Econometrica 45*, 463–485.

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In S. Griliches and M. D. Ingriligator (Eds.), *The Handbook of Econometrics*, Volume 1, pp. 449–516. North-Holland.

Phillips, P. C. B. (1989). Partially identified econometric models. *Econometric Theory 5*, 181–240.

Robin, J.-M. and R. J. Smith (2000). Tests of rank. *Econometric Theory 16*, 151–175.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In S. Griliches and M. D. Ingriligator (Eds.), *The Handbook of Econometrics*, Volume 2, pp. 881–935. North-Holland.

Rudd, J. and K. Whelan (2001). New tests of the New-Keynesian Phillips Curve. Discussion Paper, Federal Reserve Board, USA.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica 26*, 393–415.

Sargan, J. D. (1983). Identification and lack of identification. *Econometrica 51*, 1605–1633.

Sbordone, A. M. (2002). Prices and unit labor costs: a new test of price stickiness. *Journal of Monetary Economics 49*, 265–292.

Shea, J. (1997). Instrument relevance in multivariate linear models: a simple measure. *The Review of Economics and Statistics 79*, 348–352.

Staiger, D. and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–586.

Stock, J., J. Wright, and M. Yogo (2002). GMM, weak instruments, and weak identification. *Journal of Business and Economic Statistics 20*, 518–530.

Stock, J. and M. Yogo (2001). Testing for weak instruments in linear IV regression. manuscript, Kennedy School of Government, USA.

Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica 68*(5), 1055–1096.

Taylor, J. B. (1999). *Monetary Policy Rules.* Chicago: University of Chicago Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes.* New York: Springer-Verlag. With applications to statistics.

Wang, J. and E. Zivot (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica 66*(6), 1389–1404.

West, K. D. (1986). Full versus limited information estimation of a rational expectations model: Some numerical comparisons. *Journal of Econometrics 33*, 367–386.

West, K. D. (1997). Another heteroskedasticity- and autocorrelation-consistent covariance matrix estimator. *Journal of Econometrics 76*, 171–191.

West, K. D. and D. W. Wilcox (1996). A comparison of alternative instrumental variables estimators of a dynamic linear model. *J. Bus. Econom. Statist. 14*(3), 281–293.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–26.

Wright, J. H. (2000). Detecting lack of identification in GMM. *Board of Governors of the Federal Reserve System, International Finance Discussion Paper 674*, 1–1.