ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

MODEL SELECTION IN EQUATIONS WITH MANY 'SMALL' EFFECTS

Jennifer L. Castle, Jurgen A. Doornik and David F. Hendry

Number 528
February 2011

Manor Road Building, Oxford OX1 3UQ

Model Selection in Equations with Many 'Small' Effects

JENNIFER L. CASTLE[†], JURGEN A. DOORNIK[♦] and DAVID F. HENDRY[*]
[†]Magdalen College and Institute for Economic Modelling,
Oxford Martin School, University of Oxford, UK (jennifer.castle@magd.ox.ac.uk)
[♦]Economics Department and Institute for Economic Modelling,
Oxford Martin School, University of Oxford, UK (jurgen.doornik@nuffield.ox.ac.uk)
[*]Economics Department and Institute for Economic Modelling,
Oxford Martin School, University of Oxford, UK (david.hendry@nuffield.ox.ac.uk)

February 3, 2011

Abstract

General unrestricted models (GUMs) may include important individual determinants, many small relevant effects, and irrelevant variables. Automatic model selection procedures can handle perfect collinearity and more candidate variables than observations, allowing substantial dimension reduction from GUMs with salient regressors, lags, non-linear transformations, and multiple location shifts, together with all the principal components representing 'factor' structures, which can also capture small influences that selection may not retain individually. High dimensional GUMs and even the final model can implicitly include more variables than observations entering via 'factors'. We simulate selection in several special cases to illustrate.

JEL classifications: C51, C22.
KEYWORDS: Model selection, high dimensionality, principal components, non-linearity, Monte Carlo.

1 Introduction

Macroeconomic time-series are highly complicated, with many potential intercorrelated explanatory variables, long dynamic interactions, various non-stationarities, non-linearities, and multiple structural breaks. Building econometric models of such phenomena from data measured with non-negligible errors requires that all aspects of the time-series be captured, as any omissions 'contaminate' the included effects. Very high-dimensional initial models are likely as the potential set of explanatory variables will usually include individual determinants with significant explanatory power, irrelevant variables, and many relevant variables that may have small effects individually that would not be significant at conventional levels, and hence not retained when selection is undertaken. This third group may be captured by combining variables with small relevant effects to increase their joint explanatory power. We propose doing so by capturing the otherwise unexplained co-movements of the observable time series using latent factors represented by their principal components. These would also allow for genuinely relevant common forces to be modelled explicitly. Automatic model selection can then be applied from a general unrestricted model (GUM). That GUM could include all the individual variables and their principal components, so will necessarily be perfectly collinear, but we exploit the ability of automatic model selection

[*]This research was supported in part by grants from the Open Society Institute and the Oxford Martin School.

to handle such a problem (see e.g., Hendry and Krolzig, 2005). Alternatively, significant individual variables can be selected first and then principal components computed for the remaining omitted variables to capture additional small effects. Both procedures are evaluated below.

Dimension reduction

As we anticipate very high-dimensional GUMs, reduction and selection take five distinct forms:
1. conventional selection, where variables with insignificant estimated coefficients are eliminated;
2. lag-length reduction;
3. reducing a saturating set of impulse indicators (i.e., one for every observation);
4. representing potentially very high-dimensional non-linear reactions in a low-dimensional form;
5. combinations of 'small effects' by their principal components.
The first three involve selection, perhaps after an extension of the basic information set, whereas the last two require transformations followed by selection. Importantly, in any realistic setting, all these dimension reductions need to be implemented together.

To formalize these reductions, let $\{\mathbf{x}_t\}$ denote the time series of n potential explanatory variables modelling y_t, where $\mathbf{z}_t$ is the complete set of their principal components, and $1_{\{i=t\}}$ are the saturating set of impulse indicators, then the GUM allowing for all the complications simultaneously is:

$$\begin{aligned} y_t = &\sum_{i=1}^{n}\sum_{j=0}^{s}\beta_{i,j}x_{i,t-j} + \sum_{i=1}^{n}\sum_{j=0}^{s}\kappa_{i,j}z_{i,t-j} + \sum_{i=1}^{n}\sum_{j=0}^{s}\theta_{i,j}z^2_{i,t-j} \\ &+ \sum_{i=1}^{n}\sum_{j=0}^{s}\gamma_{i,j}z^3_{i,t-j} + \sum_{j=1}^{s}\lambda_j y_{t-j} + \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \epsilon_t \end{aligned} \tag{1}$$

resulting in $N > T$ regressors, L of which are relevant as defined by non-zero population t-values.

1. Castle, Doornik and Hendry (2011a) discuss the general approach to model selection based on *Autometrics* (see Doornik, 2009a, embodied in *PcGive*, Hendry and Doornik, 2009) and establish its excellent properties when the GUM nests the data-generating process (DGP), even when there are more variables, N, than observations, T. Castle and Hendry (2010b) analyze the converse problem of selection in under-specified equations with breaks, and Hendry and Johansen (2010) propose a procedure for retaining theory-based specifications when there are more variables than observations. Section 3 briefly describes automatic model selection (see Castle and Hendry, 2011b, for a more extensive overview).

2. Castle *et al.* (2011a) and Hendry and Doornik (2009) also discuss lag-length reduction, which can use sequential F-tests from the longest lag jointly on all variables. We do not explicitly address this issue here, to focus on the reduction and representation of many small effects, but recognize that a similar analysis could apply to collecting many small dynamic influences, as in (say) common-factor dynamics (see Sargan, 1980, and Hendry and Mizon, 1978).

3. Castle, Doornik and Hendry (2011b) investigate impulse-indicator saturation (IIS) for handling multiple parameter shifts, outliers and data contamination, based on Hendry, Johansen and Santos (2008) and its extension to both stationary and unit-root autoregressions in Johansen and Nielsen (2009), with an empirical application in Hendry and Mizon (2011). Section 5 explains the key aspects of IIS. Banerjee, Marcellino and Masten (2008) and Stock and Watson (2009) consider 'factor' forecasts of macroeconomic variables facing structural change.

4. Castle and Hendry (2011a) describe an automatic algorithm for non-linearity which includes cubic and exponential functions of the principal components, following the test for non-linearity in Castle and Hendry (2010a). Section 4 describes our handling of non-linear model formulation and selection, as already reflected in (1) by the use of powers of individual principal components rather than cubic polynomials in the original variables.

5. is the subject of this paper.

In each setting, the aim is to select relevant variables and eliminate irrelevant effects. The converse trade-off between omitting relevant and retaining irrelevant depends on the chosen significance level, α. For $N-L$ irrelevant regressors in (1), $\alpha(N-L)$ variables would be retained adventitiously, so for $(N-L)=1000$ (say) and $\alpha=0.001$, then $\alpha(N-L)=1$, so on average 999 (i.e., almost all) irrelevant variables will be eliminated. Conversely, relevant variables would be retained when their t-statistics exceeded the corresponding critical value c_α, so $|\mathsf{t}| \geq c_\alpha$. For variables with small non-centralities, ψ, at the available sample size, again using $\alpha=0.001$:

$$\Pr\left(\left|\mathsf{t}_{\psi^2<4}\right| \geq c_\alpha\right)$$

would be small, and hence such variables, while relevant, would rarely be retained. Here we consider approximating the effects of many small influences. To focus on capturing such effects, we do not jointly address the other additional complications in (1).

The structure of the paper is as follows. Section 2 considers the case where the model is an over-specification of the DGP, allowing for many small relevant effects, as well as substantively relevant variables. §2 considers equations with factor structures. Section 3 briefly describes automatic model selection and how we evaluate its performance, including handling non-linear model selection in §4, and three different situations of perfect collinearity in §5, one when there are more variables than observations. Section 6 examines Monte Carlo evidence, including the properties of selection (i) under the null when no variables or factors are relevant (§6), (ii) when principal components are used to parsimoniously approximate many small effects (§6), and (iii) when there are both individually relevant variables and small effects, as well as irrelevant variables (§6). Finally, we examine the case when the DGP has a factor structure in §6, and §7 concludes.

2 Factors

Factor structures have been used extensively in economics, ranging from risk measures, distilling of disaggregated data (e.g. Mayo and Espasa, 2009), construction of economic indicators and forecasting (Stock and Watson, 2002). There are two common approaches to 'factor forecasting'. Stock and Watson (1998) propose static principal components analysis to estimate the factors, whereas Forni, Hallin, Lippi and Reichlin (2000) use dynamic principal components. Favero, Marcellino and Neglia (2005) find evidence that these methods deliver similar results, based on goodness of fit, when a common information set is used. As we propose to use the principal components within a regression framework, we abstract from many of the issues of standard factor analysis, namely determining the number of factors, the factor loadings, and the idiosyncratic components. Instead, we use principal components analysis to ensure all the variability in the data is accounted for, which delivers perfect collinearity when all of these are included together with the individual regressors.[1] Approximate factor models relax the assumptions of serially uncorrelated and homoskedastic idiosyncratic errors made in the static factor model case by assuming $N \to \infty$, (Chamberlain and Rothschild, 1983, and Stock and Watson, 2002) which still ensures the principal components estimator is consistent and asymptotically normal, established by Bai (2003).

Let $\widehat{\mathbf{\Omega}}$ denote the $n \times n$ sample correlation matrix of the set of $(T \times n)$ potential variables $\mathbf{X}$ which have been transformed to non-integrated by appropriate differencing. The eigenvalue decomposition is:

$$\widehat{\mathbf{\Omega}} = \widehat{\mathbf{H}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{H}}' \tag{2}$$

[1] Factor analysis and principal components analysis are equivalent when the variances of the idiosyncratic components are identical.

where $\widehat{\mathbf{\Lambda}}$ is the diagonal matrix of ordered eigenvalues $(\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_n \geq 0)$ and $\widehat{\mathbf{H}} = (\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_n)$ is the corresponding matrix of eigenvectors, with $\widehat{\mathbf{H}}'\widehat{\mathbf{H}} = \mathbf{I}_n$. The sample principal components are computed as:

$$\widehat{\mathbf{Z}} = \widehat{\mathbf{H}}'\widetilde{\mathbf{X}} \tag{3}$$

where $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_T)'$ is the standardized data, $\widetilde{x}_{j,t} = (x_{j,t} - \overline{x}_j) / \widetilde{\sigma}_{x_j}$ $\forall j = 1, \ldots, n$ where $\overline{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_{j,t}$ and $\widetilde{\sigma}_{x_j} = \left[\frac{1}{T}\sum_{t=1}^{T}(x_{j,t} - \overline{x}_j)^2\right]^{1/2}$.

Abadir, Distaso and Žikeš (2010) propose a method of estimating the eigenvalues based on using sub-sample estimates of the eigenvectors to approximately orthogonalize the data. As $\widehat{\mathbf{H}}$ is a consistent estimator and is well conditioned for any T and n here, estimates of (3) are computed by the standard decomposition in (2). The two-step estimator could be used to compute the eigenvalues, $\widetilde{\mathbf{\Lambda}}$, but these are not explicitly used in the selection algorithm.

Factor models are used as a solution to dimensionality constraints. Automatic model selection techniques mean that degrees of freedom constraints do not bind; selection can be applied with $N \gg T$, where N is the full set of regressors (e.g., n variables, n principal components, possibly further non-linear transformations of the principal components, lags and T impulse indicators). However, increasing the dimension of N will require selection to be undertaken at a tighter significance level α to avoid over-fitting, and this will reduce the probability of detecting the individually small relevant effects.

Equations with 'factor' structures

There are two distinct states of nature for a 'factor structure' DGP:
(i) the DGP contains many small relevant effects that can be approximated by latent factors, and
(ii) the DGP is a function of 'common trends' that are modelled as latent variables.
The former is the interesting case here, and we consider two alternative selection procedures.

The simplest DGP over $t = 1, \ldots, T$ that focuses on case (i) is given by:

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \epsilon_t \tag{4}$$

where $\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$. The $n \ll T$ valid conditioning variables $\mathbf{x}_t$ are not perfectly collinear, and are independent of $\{\epsilon_t\}$. $\boldsymbol{\beta}$ contains non-zero elements (relevant variables) and zero elements (irrelevant variables). Of the non-zero elements, some βs result in low but non-zero non-centralities, which are defined as absolute t-statistics with non-centralities that are less than the critical value, c_α, for the chosen significance level.

In the first procedure, denoted the 'joint' procedure, the GUM will include all the individual regressors and their principal components computed from (3), leading to $2n$ regressors (abstracting from non-linear transformations of the principal components and impulse indicators):

$$y_t = \boldsymbol{\gamma}'\mathbf{x}_t + \boldsymbol{\delta}'\widehat{\mathbf{z}}_t + \nu_t. \tag{5}$$

so the joint set $\mathbf{w}_t = (\mathbf{x}_t : \widehat{\mathbf{z}}_t)$ must be perfectly collinear, with n collinearities. We have re-labelled the coefficients of (4) from $\boldsymbol{\beta}$ to $\boldsymbol{\theta} = (\boldsymbol{\gamma} : \boldsymbol{\delta})$ because different subsets of the augmented (collinear) set may be retained in different replications.

When the principal components capture combinations of small relevant effects in the DGP, they are not directly 'relevant' variables. Several principal components may be needed to capture these effects, or the individual regressors could be retained in conjunction with the principal components. To return to the original specification as a function of $\mathbf{x}_t$ to evaluate biases and root mean squared errors (RMSEs), we solve out the retained principal components. After selection, denote retained $\mathbf{x}_t$ and $\widehat{\mathbf{z}}_t$ by $\mathbf{x}_{r,t}$ and $\widehat{\mathbf{z}}_{r,t}$ respectively, with estimated coefficients $\widetilde{\boldsymbol{\gamma}}_r$ and $\widetilde{\boldsymbol{\delta}}_r$:

$$y_t = \widetilde{\boldsymbol{\gamma}}_r'\mathbf{x}_{r,t} + \widetilde{\boldsymbol{\delta}}_r'\widehat{\mathbf{z}}_{r,t} + \widehat{v}_t \tag{6}$$

where from (3) the retained principal components are:

$$\widehat{\mathbf{z}}_{r,t} = \widehat{\mathbf{H}}_r \widetilde{\mathbf{x}}_t$$

Solving out from the principal components for the $x_{i,t}$ results in:

$$\begin{aligned} y_t &= \widetilde{\boldsymbol{\gamma}}_r' \mathbf{x}_{r,t} + \widetilde{\boldsymbol{\delta}}_r' \widehat{\mathbf{H}}_r \widetilde{\mathbf{x}}_t + \widehat{v}_t \\ &= \phi_r + \widetilde{\boldsymbol{\gamma}}_s' \mathbf{x}_t + \widehat{v}_t \end{aligned} \tag{7}$$

An alternative procedure, denoted the 'sequential' procedure, pins down the coefficient estimates of the retained variables first, $\overline{\boldsymbol{\gamma}}_r'$, before selecting the principal components to see if a combination of the remaining omitted variables yields further explanatory power. In this procedure, selection is applied to (4) and the retained set of regressors is denoted $\mathbf{x}_{r,t}$. Define the complement of the retained set as $\mathbf{x}_{o,t}$, i.e. the variables omitted after selection ($\mathbf{x}_{o,t} = \{S : S \in \mathbf{x}_t, S \notin \mathbf{x}_{r,t}\}$). Principal components are then computed from the omitted set:

$$\widetilde{\mathbf{z}}_{o,t} = \widehat{\mathbf{H}}_o \widetilde{\mathbf{x}}_{o,t} \tag{8}$$

Selection is then undertaken on the principal components, forcing the retention of the previously selected variables (i.e., not selecting over $\mathbf{x}_{r,t}$) with the retained set of factors denoted $\widetilde{\mathbf{z}}_{o,r,t}$:

$$y_t = \overline{\boldsymbol{\gamma}}_r' \mathbf{x}_{r,t} + \overline{\boldsymbol{\delta}}' \widetilde{\mathbf{z}}_{o,r,t} + \widehat{\eta}_t. \tag{9}$$

Solving out for the principal components as in (7) yields:

$$y_t = \phi_{o,r} + \overline{\boldsymbol{\gamma}}_{o,s}' \mathbf{x}_t + \widehat{\eta}_t. \tag{10}$$

This method ensures that retained 'large effects' are not contaminated by additional 'small effects', as they are excluded when the principal components decomposition is calculated. We now describe how the variables and/or factors are selected.

3 Automatic model selection

Autometrics within *PcGive* (see Doornik, 2007, 2009a, Hendry and Doornik, 2009) is an automatic model selection algorithm that seeks to locate the local data generating process (LDGP: the DGP for the set of variables under consideration, see e.g. Hendry, 2009). A multi-path general-to-specific search is undertaken, eliminating irrelevant variables while ensuring congruency, with the resulting selected model a valid restriction of the general model that should encompass all other models that are also valid restrictions. The full GUM is not identified under perfect singularity (see Belsley and Klema, 1974, for detection of multicollinearity using singular value decompositions), but as shown in Hendry and Krolzig (2005) and Doornik (2009b), the multi-path search procedures in automatic model selection algorithms can handle perfect collinearity. One of the perfectly-collinear variables would be initially excluded from the model (determined by the rotation matrix), but the multi-path search allows that excluded variable to be included in an different path search, with another perfectly-singular variable being dropped. §5 discusses the use of expanding as well as contracting searches when $N \gg T$.

Evaluating the selected model

Even when the DGP includes latent factors as part of the explanatory regressors, selection can be evaluated by its success at retaining 'relevant' and excluding 'irrelevant' variables. Let the first L regressors in (1) be relevant, with the remaining $N - L$ irrelevant. Let $\widetilde{\beta}_{j,i}^*$ denote the OLS coefficient estimate

after selection for the coefficient on the jth regressor in replication i, with M replications. When $1(\cdot)$ is the indicator variable, *potency* and *gauge* respectively calculate the retention frequencies of relevant and irrelevant variables as:

$$\begin{array}{lcl} \text{retention rate: } \widetilde{\mathsf{p}}_j & = & \frac{1}{M}\sum_{i=1}^{M} 1(\widetilde{\beta}^*_{j,i} \neq 0), \quad j = 1, \ldots, N, \\ \text{potency} & = & \frac{1}{L}\sum_{j=1}^{L} \widetilde{\mathsf{p}}_j, \\ \text{gauge} & = & \frac{1}{N-L}\sum_{j=L+1}^{N} \widetilde{\mathsf{p}}_j. \end{array} \tag{11}$$

As discussed in Castle *et al.* (2011a), when selecting from the linear model:

$$y_t = \sum_{j=1}^{N} \beta_j x_{j,t} + \varepsilon_t \tag{12}$$

under the null that $\beta_j = 0 \; \forall j$, using a significance level α, then *Autometrics* will retain approximately $\alpha(N - L)$ irrelevant regressors. When mis-specification tests are undertaken to check the congruence of (4), the gauge, g, can exceed the nominal significance level α because irrelevant variables can be retained to offset chance significant diagnostic tests, but usually only by a small amount, so $g \simeq \alpha$.

If the latent factors do not directly enter the DGP, but such factors approximate variables in the DGP, then potency and gauge are not useful concepts because retained factors would be counted in the gauge but are contributing to potency by approximating relevant variables. Instead, we evaluate selection based on the conditional 'solved out' estimated coefficient biases:

$$\mathsf{bias}_j = \frac{\sum_{i=1}^{M}\left[\left(\widetilde{\beta}^*_{j,i} - \beta_j\right) 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)\right]}{\sum_{i=1}^{M} 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)}, \quad j = 1, \ldots, n \tag{13}$$

where the mean conditional estimated coefficients are:

$$\beta^*_j = \frac{\sum_{i=1}^{M}\left[\widetilde{\beta}^*_{j,i} 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)\right]}{\sum_{i=1}^{M} 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)}, \quad j = 1, \ldots, n, \tag{14}$$

and conditional RMSEs:

$$\mathsf{RMSE}_j = \left[\frac{\left[\sum_{i=1}^{M}\left(\widetilde{\beta}^*_{j,i} - \beta_j\right)^2 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)\right]}{\sum_{i=1}^{M} 1\left(\widetilde{\beta}^*_{j,i} \neq 0\right)}\right]^{1/2}, \quad j = 1, \ldots, n \tag{15}$$

or β_j^2 when $\sum_{i=1}^{M} 1(\widetilde{\beta}^*_{j,i} \neq 0) = 0$. We also recorded retention rates as calculated by (11).

4 Non-linear model formulation and selection

Factor models are commonly used in a regression framework, for example, in the estimation of composite indicators (OECD, 2008). However, this implies the assumption of linear behaviour. The LDGP could be a non-linear entity for which a linear model is a poor approximation, resulting in a non-congruent model. Yalcin and Amemiya (2001) propose a general parametric non-linear factor analysis model that retains linearity in the parameters, and Jones (2006) provides an application that uses non-linear functions (orthogonal polynomials) of factors to estimate futures options.

Testing for evidence of non-linearity prior to selection can point towards whether an alternative functional form is required. Castle and Hendry (2010a) propose a parsimonious portmanteau test based on principal components, which is applicable prior to undertaking estimation, or can form part of a model selection procedure. The test uses non-linear functions of the principal components, namely squares, cubes and expnentials, to capture general third-order polynomials and exponential functions, and yields an F-test with $3n$ degrees of freedom for n linear variables. If the non-linearity test does not reject, then a linear specification given by dropping the non-linear terms in (1) forms the GUM from which selection is undertaken. Alternatively, rejection leads to specifying the general non-linear approximation in (1).

The dimensionality reduction from these functions of principal components is essential when there are large numbers of potential regressors, as a third-order polynomial in n variables will have $K_n = n(n+1)(n+5)/6$ terms, leading to an explosion as n increases ($K_n = 12,300$ for $n = 40$). For n initial linear regressors, adding the squares, $z^2_{j,t}$, cubes $z^3_{j,t}$, and exponential functions of their principal components, $z_{j,t}$, only adds $3n$ variables, yet potentially includes many squares, cubes, second and third order interactions and exponential functions of the original $x_{j,t}$. This results in a flexible parsimonious non-linear approximation, yet jointly resolves the potential problems of high dimensionality, identification, and the restriction to second-order departures intrinsic to a quadratic formulation. The class of non-linearity that we consider retains linearity in the parameters, by redefining non-linear functions of the principal components as new variables (e.g., $z^2_{j,t} = u_{j,t}$ say), so the model becomes linear but larger, and standard selection theory applies. Non-linear transformations of the individual regressors could be included instead of, or as well as, the non-linear principal components. We propose just the principal components in order to pick up non-linearities in common trends which are likely to be slowly evolving, but further terms could be included as long as the significance level is tightly controlled.

Next, the multi-path search procedure seeks a parsimonious, congruent, non-linear simplification. The use of non-linear principal components does imply that interpretation of the regression coefficients in terms of economic theory is more complicated, but the non-linear model can be solved back to a model that is non-linear in the original variables, so can approximate models such as the smooth transition model, or the Markov-switching model: see Castle and Hendry (2011a). In turn, the approximation can be tested against such specific non-linear functional forms using encompassing tests (see, e.g., Mizon and Richard, 1986, Hendry and Richard, 1989, and Bontemps and Mizon, 2008), and further simplified to them if appropriate.

Including both the linear and non-linear transformations of many variables can generate substantial 'collinearity' additional to the perfect collinearity between variables discussed below. A simple operational de-meaning rule eliminates most of the collinearity between the linear and non-linear transforms of each regressor, so selection proceeds as in the linear case. Selection will keep a 'representative' of the relevant effect if collinearity is high.

Non-linear functions of variables can also generate extreme outcomes, and the resulting 'fat tails' can be problematic for inference and model selection, as the assumption of normality is in-built into most procedures' critical values. Non-linear functions can also 'align' with outliers, causing the functions to be retained spuriously, which can be detrimental for forecasting and policy. Thus, data contamination, outliers and non-linearity interact, so need to be treated together. We propose using impulse-indicator saturation to handle such outliers, as well as data contamination and breaks, see §5.

The retention of linear and non-linear functions and indicators due to a highly over-parameterized GUM must be controlled by implementing a 'super-conservative' strategy for the non-linear functions, where selection is undertaken at stringent significance levels to control the null rejection frequency. With n variables there will be a further n principal components, n second-order polynomials of the principal components, n third-order polynomials, possibly further non-linear transformations, and T indicators, resulting in $N \gg T$. Setting $\alpha \approx 1/N$ will control chance significant effects at a small cost of increasing the probability of eliminating relevant effects, noting that IIS helps ensure normality as

well as congruence.

5 Perfect collinearity

In this section, we consider 3 cases that generate perfect collinearity in the GUM specification, all of which are handled by *Autometrics*. Two occur because $N > T$, and the third because linear combinations of variables and the variables themselves are included in the GUM.

Impulse-indicator saturation

Impulse-indicator saturation (IIS) is used to detect measurement errors, outliers and structural breaks. The theory of IIS is derived under the null of no outliers or location shifts. In the simplest analysis (the 'split-half' approach), the first $T/2$ impulse-indicators ($1_{\{i=t\}}, i = 1, \ldots, T/2$) are included in a regression, effectively dummying out the first half of the observations. Estimates are based on the remaining data, and any observations in the first half that are discrepant will result in significant indicators, which are recorded (similar to Salkever, 1976). Those indicators are dropped, and the second half of the indicators are included and the procedure repeated. The two sets of significant indicators are then added to the general model for selection of those that remain significant. Hendry *et al.* (2008) and Johansen and Nielsen (2009) show that αT impulse indicators will be retained on average at significance level α. Setting $\alpha \leq r/T$ maintains the average false null retention at r outliers, equivalent to 'losing' r observations, which is a small efficiency loss. Under the alternative, IIS can detect outliers and and multiple location shifts, including breaks close to the start and end of the sample (see Castle *et al.*, 2011b, for Monte Carlo evidence on the detectability of such breaks).

More variables than observations and sparcity

A straightforward generalization of IIS demonstrates that selection can be made with $N > T$. Rather than splitting into just two halves, when $N \gg T$ more blocks are needed. The search algorithm used allows for learning about the relevance of variables as the blocks are formed by the use of expanding as well as contracting searches. The expanding searches check for omitted variables, which allows the inclusion of perfectly collinear variables in the GUM so long as they are in separate blocks. Then a reduction stage eliminates variables for each block that is augmented by those omitted variables. Iterations of these procedures are undertaken until convergence is reached, where the retained set of variables is unchanged from the previous iteration. Doornik (2009b) outlines the algorithm and provides Monte Carlo evidence on its performance.

Selection from a GUM with more variables than observations, so $N > T$ in (12) say, appears to hinge on the sparcity of the LDGP parameter vector. Indeed, we require sparcity of the population vector $(\boldsymbol{\beta} : \boldsymbol{\kappa} : \boldsymbol{\theta} : \boldsymbol{\gamma} : \boldsymbol{\lambda} : \boldsymbol{\delta})$ in (1), otherwise the LDGP is not identified. Moreover, the final selection must have fewer unrestricted variables than observations, otherwise the model is not estimable. However, the unrestricted version of the final model may be derived from a 'factor' formulation where the number of free parameters is less than T, but has more non-zero elements than T, as in (7) where $dim(\boldsymbol{\gamma}_r) + dim(\boldsymbol{\delta}_r) < T$ but nothing precludes $dim(\boldsymbol{\gamma}_s) > T$. Similarly, for a parsimionious selection from (1) after the retained $z^k_{i,t}$ are expanded in terms of $x^k_{j,t}$. Thus, when selecting over principal components in conjunction with variables, sparcity in the final model is no longer required: fewer than T effects can be separately estimated but the other variables could all be 'small' relevant effects captured by principal components.

Multi-path selection with perfect collinearity

To illustrate the analysis, consider the simple DGP:

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t \tag{16}$$

where in fact $\beta_1 = -\beta_2$, but that is not known. When the GUM is specified as

$$y_t = \gamma_1 x_{1,t} + \gamma_2 x_{2,t} + \gamma_3 (x_{1,t} - x_{2,t}) + v_t \tag{17}$$

using a comprehensive multi-path search, then one path will delete $(x_{1,t} - x_{2,t})$ and (for sufficiently large test non-centralities) retain $x_{1,t}$ and $x_{2,t}$; a second path will eliminate $x_{2,t}$ and should retain $(x_{1,t} - x_{2,t})$ but also drop $x_{1,t}$ as now insignificant; and similarly for the third path commencing from first dropping $x_{1,t}$. Thus, an advantage of such procedures in dynamic specification searches is that they allow many forms of possible lag response to be included, such as $x_{1,t}$, $x_{1,t-1}$, $\Delta x_{1,t} (= x_{1,t} - x_{1,t-1})$, $(x_{1,t} + x_{1,t-1})$ and so on, where the relevant subset is retained.

Campos and Ericsson (1999) discuss the importance of the choice of the initial linear transformation of the regressors in the GUM for the final selection in methods which test null hypotheses to select. One possible solution is to include all the theoretically viable combinations from the outset. A potential cost of (say) doubling the number of variables, n, by also including n perfectly collinear combinations is that the procedure may now retain approximately $2\alpha n$ irrelevant regressors. On the one hand, it could be argued that such should not occur because there are still only n 'independent' regressors, so that null retention frequencies should be unchanged. On the other hand, many linear combinations of variables are being included, and in a t-test based approach, such combinations could be significant (still under the null) even when their components would not be. For example, $x_{1,t}$ and $x_{1,t-1}$ might have t-test values less than c_α, yet $(x_{1,t} + x_{1,t-1})$ have a significant coefficient by chance and so be retained.

Given the difficult analytic nature of trying to establish which of the two arguments holds, we have undertaken a number of Monte Carlo simulation studies. The simplest is close to the model in (17). The DGP is:

$$y_t = \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2] \tag{18}$$

for $t = 1, \ldots, T = 100$ where:

$$\mathbf{x}_t \sim \mathsf{IN}_2[\mathbf{0}, \mathbf{I}] \tag{19}$$

We first consider a case with no collinearity. The GUM has the form in (20) as a baseline to check that $g \simeq \alpha$ with 5 irrelevant regressors:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t-1} + \beta_4 x_{2,t-1} + \epsilon_t \tag{20}$$

Setting $\alpha = 0.01$, with diagnostic tests for congruence also at 1%, $M = 10000$ replications delivered $g = 0.014$ which is a little 'over-gauged' as anticipated. Thus, on average $5g = 5 \times 0.014 = 0.07$ irrelevant variables were retained per replication. Without diagnostic checking $g = 0.0098$, so the algorithm is calibrated to deliver the correct significance level under the null when there are no diagnostic checks.

We now include perfectly collinear variables, repeating these simulations with the GUM specified as:

$$\begin{aligned} y_t = {} & \gamma_0 + \gamma_1 x_{1,t} + \gamma_2 x_{2,t} + \gamma_3 (x_{1,t} - x_{2,t}) + \gamma_4 x_{1,t-1} + \gamma_5 x_{2,t-1} \\ & + \gamma_6 (x_{1,t-1} - x_{2,t-1}) + \gamma_7 \Delta x_{1,t} + \gamma_8 \Delta x_{2,t} + v_t \end{aligned} \tag{21}$$

where (21) has 9 irrelevant regressors with four collinearities. Again at $\alpha = 0.01$, now $M = 10000$ replications delivered $g = 0.0057$, so that $9g = 9 \times 0.0057 = 0.05$ irrelevant variables were retained

per replication. This is slightly smaller than anticipated, but is consistent with the argument that adding perfectly collinear variables does not increase the null retention rate.

Finally we examine the case where $N \gg T$ and there is perfect collinearity. We augment (21) with $\sum_{j=1}^{100} \delta_j u_{j,t}$ where:

$$\mathbf{u}_t \sim \mathsf{IN}_{100}[\mathbf{0}, \mathbf{I}] \tag{22}$$

resulting in 109 regressors for 100 observations. Selection at $\alpha = 0.01$ with $M = 10000$ replications delivered $g = 0.0105$, which is only slightly larger than α, with $109g = 1.14$ irrelevant variables retained per replication, thus controlling gauge despite including $109 > T$ irrelevant variables. Undertaking selection at $\alpha = 0.005$ yielded $109g = 0.44$ variables retained per replication, which is now smaller than the expectation of $109\alpha = 0.545$, demonstrating that inclusion of many irrelevant variables does not lead to over-fitting, but not resolving whether or not the 4 perfect collinearities added to gauge.

6 Monte Carlo evidence on principal component combinations

We undertook a range of simulations of (4) to evaluate the performance of model selection for principal components formulations. First, §6 confirms the properties of selection under the null that no principal components are relevant. Secondly, §6 evaluates the ability of principal components to capture many small effects. Finally, §6 considers the merits of selecting the variables and principal components jointly or sequentially.

The DGP is (4), with regressors generated by:

$$\mathbf{x}_t \sim \mathsf{IN}_n \left[\mathbf{0}, \mathbf{\Sigma}\right], \quad t = 1, \ldots, T, \tag{23}$$

where $\sigma_{i,i} = 1$ for $i = 1, \ldots, n$ and $\sigma_{i,j} = \rho$, $\forall i \neq j$. In all reported simulations, the sample size $T = 100$ and $M = 10,000$ replications were undertaken, with regressors drawn independently for each replication (i.e., not fixed in repeated samples), with $n = 2, 10, 50 < T$ regressors so $2n = 4, 20, 100 = T$. The model selection algorithm used is *Autometrics*, with no diagnostic checking, and selection from assuming (4) is known provides an upper bound on performance.

Null retention frequency

The first setting we consider is when $\boldsymbol{\beta} = \mathbf{0}$ in (4) to establish the null retention frequency (gauge) when estimated factors $\widehat{z}_{i,t}$ are entered, with or without variables. The three alternative GUMs are given by (5) where (i) $\boldsymbol{\delta} = \mathbf{0}$; (ii) $\boldsymbol{\gamma} = \mathbf{0}$ and; (iii) $\boldsymbol{\delta} \neq \mathbf{0}$ and $\boldsymbol{\gamma} \neq \mathbf{0}$. We set $\rho = 0.5$ and 0.9.

Table 1 records the gauge, where α denotes the significance level at which selection is undertaken. The inclusion of factors instead of variables has no effect on the gauge which is identical to selection over the individual variables. The gauge is close to the nominal significance level for small n, but is slightly under for $n = 50$, although with many regressors or factors, a tighter significance level might be used. Inclusion of both regressors and factors results in perfect collinearity for the joint procedure. For small n, only reduction paths are searched and this reduces the null retention frequency relative to the nominal significance level. When $n = 50$, there are as many regressors as observations so expanding as well as contracting block searches are used, and this results in a gauge close to the nominal significance level. Thus, the second argument above seems correct: doubling the number of irrelevant variables to search over by adding perfectly collinear combinations, doubles the number of adventitious retentions. Nevertheless, at tight significance levels this will be a small cost.

Table 1: Gauge for selection with (i) variables, (ii) principal components, and (iii) variables and principal components.

	$\alpha = 0.05$			$\alpha = 0.01$		
	$n = 2$	$n = 10$	$n = 50$	$n = 2$	$n = 10$	$n = 50$
$\rho = 0.5$						
(i)	0.0521	0.0482	0.0307	0.0116	0.0087	0.0047
(ii)	0.0521	0.0482	0.0307	0.0116	0.0087	0.0047
(iii)	0.0244	0.0016	0.0468	0.0038	0.0002	0.0105
$\rho = 0.9$						
(i)	0.0518	0.0567	0.0312	0.0111	0.0128	0.0056
(ii)	0.0518	0.0567	0.0312	0.0111	0.0128	0.0056
(iii)	0.0259	0.0019	0.0484	0.0041	0.0002	0.0130

Using factors to capture small effects

We next consider whether principal components can capture many small effects. The DGP is given by (4) where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ is calibrated to give non-centralities, $\psi_i = 1, \forall i$, so these are small effects (the probability of retaining a variable with a non-centrality of 1 at $\alpha = 0.05$ is 16% and at $\alpha = 0.01$ is 5% for a single t-test). The models considered include:

[A] estimation of (4), with estimated parameters denoted $\widehat{\beta}_i$;

[B] selection commencing from (4), with parameter estimates denoted $\widetilde{\beta}_i$;

[C] bias correction applied to the estimates obtained by selection commencing from (4), with parameter estimates denoted $\widetilde{\beta}_i^{BC}$;

[D] estimation of the principal components model:

$$y_t = \sum_{i=1}^{n} \delta_i \widehat{z}_{i,t} + \eta_t \tag{24}$$

with the estimated coefficients computed by $\widehat{\overline{\boldsymbol{\beta}}} = \widehat{\mathbf{H}}\widehat{\boldsymbol{\delta}}$. Thus, when all principal components are included, there will be no difference between estimating the original regression model or the model after applying linear transformations (other than calculation of the eigenvectors).

[E] selection commencing from the principal components model (24), with the parameters denoted $\widetilde{\overline{\boldsymbol{\beta}}} = \widehat{\mathbf{H}}_r \widetilde{\boldsymbol{\delta}}_r$;

[F] bias correrection applied to the parameter estimates from the selected principal components model, denoted $\widetilde{\overline{\beta}}_i^{BC}$.

As the principal components are orthogonal, it is sufficient to use the '1-cut rule' to select the number of factors for [E] (see Castle *et al.*, 2011a). Coefficient estimates of the individual regressors and principal components will both be biased after selection. Hendry and Krolzig (2005) propose an approximate bias correction, and we use their 2-step correction in [C] and [F]. For [F], the retained coefficient estimates of the principal components are bias corrected, $\widetilde{\delta}_{r,i}^{BC}$, and the original coefficients are backed out using $\widetilde{\overline{\boldsymbol{\beta}}}^{BC} = \widehat{\mathbf{H}}_r \widetilde{\boldsymbol{\delta}}_r^{BC}$. Evaluation criteria are given in §3, where $\beta_i^* = \widehat{\beta}_i, \widetilde{\beta}_i, \widetilde{\beta}_i^{BC}, \widehat{\overline{\beta}}_i, \widetilde{\overline{\beta}}_i$ and $\widetilde{\overline{\beta}}_i^{BC}$.

First consider the case where the factor structure of the $\mathbf{x}_t$ matches that of the relation between y_t and $\mathbf{x}_t$ in (4). The simplest setting, which is then amenable to analysis as well as simulation, is when:

$$\mathbf{x}_t \sim \mathsf{IN}_n \left[\mathbf{0}, \sigma_w^2 \mathbf{\Omega}\right] \tag{25}$$

where $\mathbf{\Omega} = (1-\rho)\,\mathbf{I}_n + \rho \boldsymbol{\iota}\boldsymbol{\iota}'$, so all the variables have a common correlation ρ, and $\boldsymbol{\beta} = \phi\boldsymbol{\iota}$, so all the coefficients have a common value. For example, when $n = 3$:

$$\mathbf{\Omega} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \tag{26}$$

which has an eigenvector of $\mathbf{h}_1 = \boldsymbol{\iota} = (1 \dots 1)'$ so there is a 'factor' $f_{1,t}$ such that:

$$\phi f_{1,t} = \phi \boldsymbol{\iota}' \mathbf{x}_t = \boldsymbol{\beta}' \mathbf{x}_t. \tag{27}$$

This must be one of the most favourable cases, having a 'factor DGP', since (4) becomes:

$$y_t = \phi f_{1,t} + \epsilon_t \tag{28}$$

Let the non-centrality of the t-test in (4) of the null hypothesis that $\beta_i = 0$ be:

$$\psi_i = \mathsf{E}\left[\mathsf{t}_{\beta_i=0} \mid \beta_i \neq 0\right] = \mathsf{E}\left[\frac{\widehat{\beta}_i}{\mathsf{SE}\left[\widehat{\beta}_i\right]}\right] \simeq \phi \frac{\sigma_w \sqrt{1 + (n-2)\,\rho - (n-1)\,\rho^2}}{\sigma_\epsilon \sqrt{1 + (n-2)\,\rho}}$$

then the non-centrality τ of the t-test of $\phi = 0$ in (28) is based on:

$$\tau = \frac{\widehat{\phi}}{\mathsf{SE}\left[\widehat{\phi}\right]} \simeq \phi \frac{\sigma_w \sqrt{n\,(1 + (n-1)\,\rho)}}{\sigma_\epsilon}$$

as:

$$\widehat{\phi} = \left(\sum_{t=1}^{T} f_{1,t}^2\right)^{-1} \sum_{t=1}^{T} f_{1,t} y_t = \phi + \left(\sum_{t=1}^{T} f_{1,t}^2\right)^{-1} \sum_{t=1}^{T} f_{1,t} \epsilon_t$$

where:

$$\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_{1,t}^2\right] = \boldsymbol{\iota}' \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right] \boldsymbol{\iota} = \sigma_w^2 \boldsymbol{\iota}' \mathbf{\Omega} \boldsymbol{\iota} = n\sigma_w^2\,(1 + (n-1)\,\rho)\,.$$

Thus, at $\sigma_w^2 = \sigma_\epsilon^2 = 1$, $\rho = 0.5$ and $n = 3$, $\psi_i = \sqrt{0.67}\phi$, whereas $\tau = \sqrt{6}\phi$; and at $\rho = 0.9$, $\psi_i = \sqrt{0.15}\phi$, whereas $\tau = \sqrt{8.4}\phi$. The ratios τ/ψ are $\sqrt{8.95} \simeq 3$ for $\rho = 0.5$ and $\sqrt{56.0} \simeq 7.5$ for $\rho = 0.9$ for $n = 3$ (as against the bias-corrected simulation values below of 3.6 and 7.6). Similarly, we can derive ratios of 10 and 28.8 at $n = 10$ (versus MC outcomes of 10.5 and 29); and 55 and 162 at $n = 50$ (MC of 60 and 124), so when there are many small effects, in this setting the principal components will be retained with a high probability relative to retention of individual regressors.

Simulation evidence is presented in table 2. For $n = 10$, when $\rho = 0.5$ and $\rho = 0.9$, $\beta_i = 0.141$ and $\beta_i = 0.315$ respectively, which deliver population t-statistics equal to one. $\overline{n}$ denotes the average number of principal components retained, 'Bias' averages (13) over $i = 1 \dots, n$, and is reported as $\times 100$, and 'RMSE' averages (15) over $i = 1 \dots, n$. The table confirms the huge advantages of representing the individually insignificant effects from $\mathbf{x}_t$ by principal components in a 'factor DGP', with a vast reduction in RMSEs relative to just estimating the correct model. Bias correction further downweights the retained

irrelevant principal components yielding a smaller RMSE, since the first principal component is highly significant and captures all of the variation in y, as demonstrated by (27).

Figure 1a reports the retention rates for the variables [B] and factors [E]. One factor is retained, with the retention rates of other factors equal to the nominal significance level. The retention of variables is slightly higher than the theoretical retention rate for a single t-test because the regressors are highly correlated. We comment on the remaining panels below.

Table 2: Approximating small effects by principal components in a 'factor DGP'

	[A]	[B] 5%	[C] 5%	[B] 1%	[C] 1%	[D]	[E] 5%	[F] 5%	[E] 1%	[F] 1%
$n = 10$; $\psi_i = 1$; $\rho = 0.5$										
$\widehat{\sigma}$	0.998	0.995	0.995	1.012	1.012	1.004	0.993	0.993	1.001	1.001
$\overline{n}$							1.443	1.443	1.091	1.091
Bias	-0.022	23.54	17.13	31.56	25.50	-0.023	-0.069	-0.075	-0.082	-0.083
RMSE	0.143	0.253	0.217	0.330	0.290	0.144	0.076	0.052	0.044	0.032
$n = 10$; $\psi_i = 1$; $\rho = 0.9$										
$\widehat{\sigma}$	0.998	0.994	0.994	1.010	1.010	1.046	1.033	1.033	1.039	1.039
$\overline{n}$							1.375	1.375	1.070	1.070
Bias	-0.018	53.15	40.26	71.86	61.05	-0.018	-0.023	-0.023	-0.024	-0.024
RMSE	0.318	0.574	0.503	0.752	0.682	0.320	0.157	0.102	0.083	0.056
$n = 10$; $\psi_i = -(-1)^i$; $\rho = 0.5$										
$\widehat{\sigma}$	0.998	1.000	1.000	1.021	1.021	0.999	0.993	0.993	0.999	0.999
$\overline{n}$							1.592	1.592	0.636	0.636
Bias $(+)$	-0.020	17.63	10.01	22.02	12.63	-0.016	-4.817	-7.093	-5.470	-7.411
Bias $(-)$	-0.025	-17.73	-10.06	-21.92	-12.56	-0.029	4.795	7.085	5.464	7.406
RMSE	0.143	0.208	0.162	0.243	0.182	0.144	0.145	0.130	0.145	0.130
$n = 10$; $\psi_i = -(-1)^i$; $\rho = 0.9$										
$\widehat{\sigma}$	0.998	0.999	0.999	1.016	1.016	0.999	0.993	0.993	1.000	1.000
$\overline{n}$							1.607	1.607	0.650	0.650
Bias $(+)$	0.011	38.22	25.17	46.83	33.21	0.019	-10.71	-15.89	-12.38	-16.78
Bias $(-)$	-0.047	-38.15	-24.96	-47.09	-33.37	-0.056	10.70	15.89	12.38	16.78
RMSE	0.318	0.456	0.373	0.522	0.420	0.320	0.322	0.289	0.325	0.289

Next, consider the case when the link of y_t to $\mathbf{x}_t$ in (4) does not match the correlation structure in (25), say the β_i alternate in sign with the same magnitude, so are $\pm\phi$. The other two eigenvectors of (26) are:

$$\mathbf{h}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{h}_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

whereas one needs $\boldsymbol{\beta}'\mathbf{x}_t = \phi \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \mathbf{x}_t$ and hence $\mathbf{h}_1 + 2\mathbf{h}_2 - 4\mathbf{h}_3$ which requires all three factors.

Table 2 reports simulation results for the case where $n = 10$ with $\beta_j = -a \times -1^j$, for $j = 1, \ldots, n$ where a is calibrated to deliver $|\mathrm{t}| = 1$ ($a = 0.141, 0.315$ for $\rho = 0.5, 0.9$). 'Bias $(+)$' and 'Bias $(-)$' average over positive and negative parameters respectively. Despite the factor structure not capturing the correlations with y, the RMSEs of the bias corrected principal components model still yields a small improvement over estimating the DGP, although a cost is incurred in the bias. Few principal components

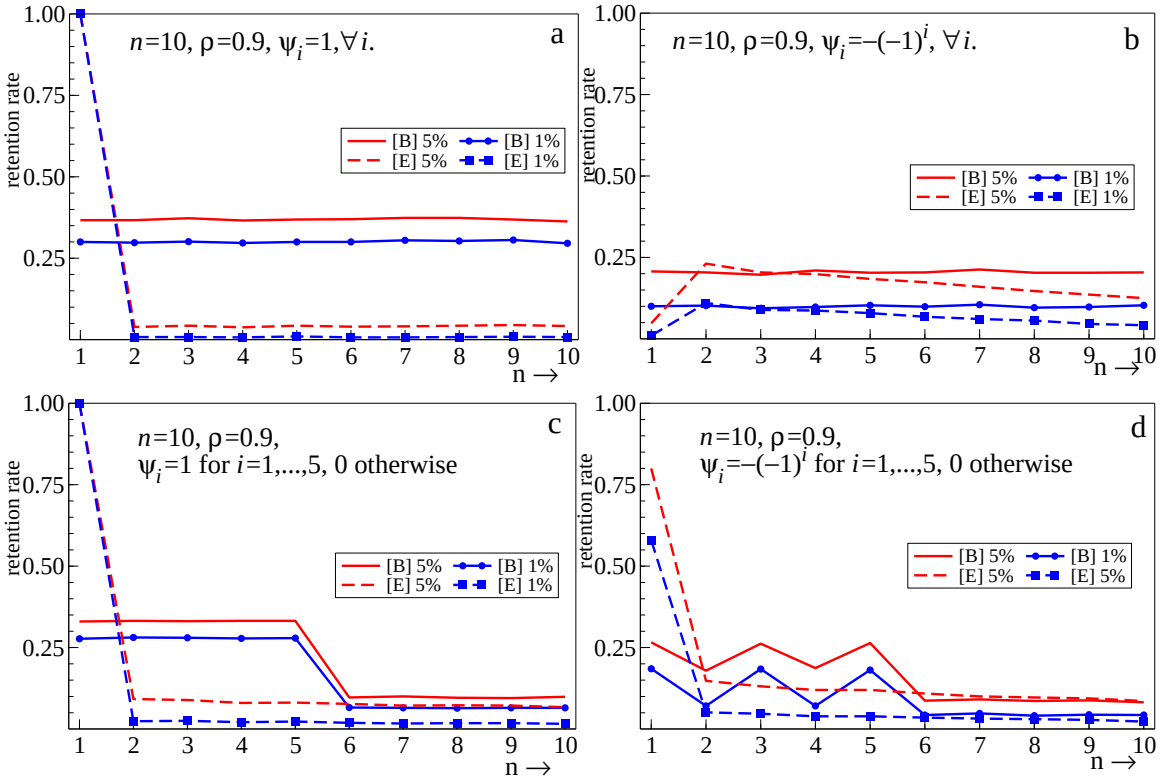

Figure 1: Retention rates of variables and factors. [B] variables; [E] factors

are retained on average, as can be seen from figure 1b, despite requiring all of them to capture the factor structure.

A final case is where a subset of the $\mathbf{x}_t$ are irrelevant so $\boldsymbol{\beta}' = \phi \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$ or $\phi \begin{pmatrix} 1 & -1 & 0 \end{pmatrix}$. The former, for example, is $\phi(2\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3)$ (again needing all three factors) whereas the latter is $\phi\mathbf{h}_2$. Table 3 reports result for the cases:

$$\boldsymbol{\beta} = (a, a, a, a, a, 0, 0, 0, 0, 0)'$$
$$\boldsymbol{\beta} = (a, -a, a, -a, a, 0, 0, 0, 0, 0)'$$

where a is calibrated to deliver $|\mathrm{t}| = 1$, for $\rho = 0.9$.

When the correlation structure between regressors matches that of the correlation structure with y, the first principal component picks up most of the variation. RMSEs are much larger than when all variables were relevant, suggesting that the factors cannot detect the irrelevant variables, so non-zero weight is placed on the irrelevant variables. There is a reduction in RMSE relative to estimating the DGP though, so despite an increase in bias, the reduction in variance is advantageous. The retention rate of factors $2-10$ has doubled, suggesting that factors are attempting to cancel out otherwise non-zero effects of the irrelevant variables.

When the correlation structure differs between the regressors and y, estimates of the relevant variables from the principal component model are significantly biased, and this is reflected in the RMSE of the relevant variables. However, the bias in the irrelevant variables is much smaller than for the previous case, due to the differing correlation structure. A high correlation results in the first principal component capturing most of the variation, whereas lower correlations imply that there is no dominant principal component. Figure 1c and d record the retention rates of variables and principal components.

To summarize, when the correlation structure of the regressors matches that between the dependent variable and the regressors, the first principal component captures most of the variation from 'small

Table 3: Approximating small effects by principal components, with irrelevant variables.

	[A]	[B] 5%	[C] 5%	[B] 1%	[C] 1%	[D]	[E] 5%	[F] 5%	[E] 1%	[F] 1%
$n = 10$; $\psi_i = 1$ for $i = 1, \ldots, 5$; $\psi_i = 0$ for $i = 6, \ldots, 10$; $\rho = 0.9$										
$\widehat{\sigma}$	0.998	0.989	0.989	0.999	0.999	1.011	1.004	1.004	1.015	1.015
$\overline{n}$							1.703	1.703	1.181	1.181
Bias $(+)$	-0.054	46.89	38.72	59.95	49.86	-0.066	-10.22	-12.11	-13.71	-14.41
Bias (0)	0.018	52.58	47.15	80.85	70.07	0.029	10.16	12.06	13.66	14.35
RMSE $(+)$	0.318	0.510	0.466	0.662	0.610	0.320	0.231	0.188	0.189	0.171
RMSE (0)	0.319	0.734	0.640	0.880	0.788	0.321	0.231	0.187	0.188	0.170
$n = 10$; $\psi_i = -(-1)^i$ for $i = 1, \ldots, 5$; $\psi_i = 0$ for $i = 6, \ldots, 10$; $\rho = 0.9$										
$\widehat{\sigma}$	0.998	0.992	0.992	1.008	1.008	0.999	0.994	0.994	1.002	1.002
$\overline{n}$							1.804	1.804	0.904	0.904
Bias $(+)$	-0.033	29.30	22.41	23.12	18.62	-0.041	-15.14	-19.46	-20.29	-22.97
Bias $(-)$	-0.084	-30.67	-15.06	-34.77	-16.26	-0.103	17.98	22.60	24.77	27.16
Bias (0)	0.018	11.67	13.58	24.96	26.12	0.029	1.604	1.681	2.345	2.229
RMSE $(\pm)$	0.318	0.403	0.349	0.394	0.332	0.320	0.307	0.284	0.308	0.295
RMSE (0)	0.319	0.613	0.510	0.585	0.511	0.321	0.240	0.176	0.193	0.143

Notes: $(+)$ denotes averaging over coefficients where $\psi_i = 1$; (0) for $\psi_i = 0$; $(-)$ for $\psi_i = -1$; and $(\pm)$ denotes averaging over coefficients where $\psi_i = \pm 1$. Results for $\rho = 0.5$ available on request.

effects' and is highly significant. This is the optimal case for a 'factor structure'. When the correlation structure differs, more principal components are needed to capture the variation due to small effects, but the additional ones have smaller non-centralities and are harder to detect. Figure 2 plots the average number of factors retained as the analytic number of principal components required increases from 1 to 3 on the horizonal axis, for $n = 10$. After a small increase in the number of principal components retained at 2, the retention rate flattens out, as does the average RMSE. Selection of principal components outperforms selection of variables when the DGP contains 'small relevant effects', evaluated on RMSE, as these variables will frequently be omitted when undertaking variable selection. The converse, that principal components selection will place a non-zero weight on all variables by construction, applies when irrelevant variables are included in the GUM, as investigated below.

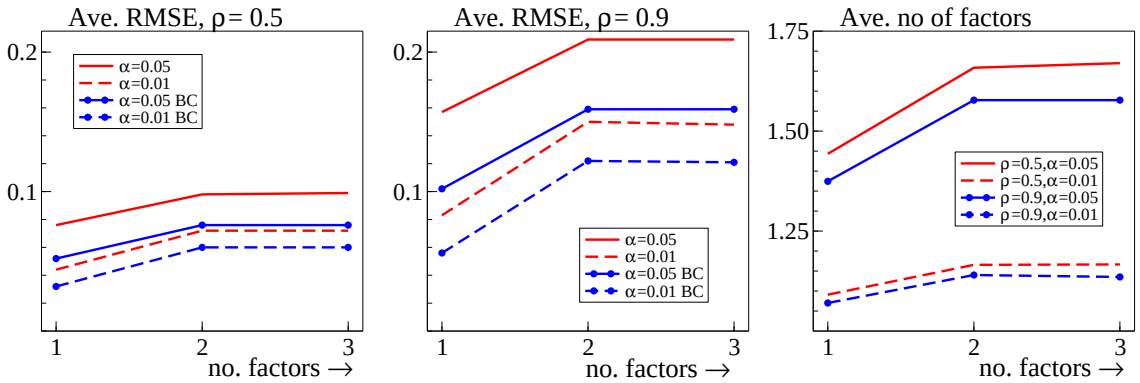

Figure 2: RMSEs and retention rates for varying numbers of theoretically required principal components

Selection with both variables and principal components

Next, we consider selection when there are both individually relevant variables and many small effects which can be captured in part by principal components. The DGP is (4), with the regressors generated by (23), where:

$$\boldsymbol{\beta} = (a, a, b, b, b, b, b, b, b, b)'$$

where a and b are calibrated to give populaton t-statistics equal to 5 and 1 respectively. Setting $\rho = 0.9$ results in $a = 1.568$ and $b = 0.315$.

We compare models [A], [B] and [C] listed above with:

[G] selection commencing from the GUM using the joint procedure, (6), with the resulting parameters denoted $\widetilde{\boldsymbol{\gamma}}_s$;

[H] bias correrection applied to the parameter estimates from the selected model (6), denoted $\widetilde{\boldsymbol{\gamma}}_s^{BC}$. Note that bias correction is applied to both $\widetilde{\boldsymbol{\gamma}}_r$ and $\widetilde{\boldsymbol{\delta}}_s$ from which the $\widetilde{\boldsymbol{\gamma}}_s^{BC}$ are backed out.

[I] selection using the sequential procedure, (9), with the resulting parameters denoted $\widetilde{\boldsymbol{\gamma}}_{o,s}$;

[J] bias correrection applied to the parameter estimates from the selected model (9), denoted $\widetilde{\boldsymbol{\gamma}}_{o,s}^{BC}$.

We compare the models based on the evaluation criteria listed in §3, where $\beta_i^* = \widehat{\beta}_i, \widetilde{\beta}_i, \widetilde{\beta}_i^{BC}, \widetilde{\boldsymbol{\gamma}}_{s,i}$, $\widetilde{\boldsymbol{\gamma}}_{s,i}^{BC}$, $\widetilde{\boldsymbol{\gamma}}_{o,s,i}$ and $\widetilde{\boldsymbol{\gamma}}_{o,s,i}^{BC}$. Ave. $\widetilde{n}$ denotes the average number of variables retained in selection and Ave. $\overline{n}$ denotes the average number of principal components retained.

Table 4, panel a, reports the results for the case with two highly significant variables (large effects) and many marginally significant variables (small effects). The RMSEs for selection of the variables and principal components model are larger than if the DGP was known but needed estimating. There is a small advantage to using the principal components to capture the small effects relative to selecting from the DGP ($\beta_3^* - \beta_{10}^*$), but at a cost of estimating the parameters on the significant variables (β_1^*, β_2^*), where the retention rate falls from almost unity to just over 50%. Hence, the first principal component (which is retained almost 50% of the time) is capturing part of the individual relevant effects, and this impacts on the RMSEs. As a result, the sequential procedure improves the RMSEs for the large effects, at a cost of an increased RMSE for the small effects, which are not as well proxied when the retained variables are forced in the second stage of selection.

We next consider the case where there are some highly significant individual regressors, many 'small effects', and also irrelevant variables. The DGP is as above but we set $n = 20$, $\rho = 0.9$, and define:

$$\boldsymbol{\beta} = (a, a, b, b, b, b, b, b, b, b, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$$

where a and b are calibrated to give populaton t-statistics equal to 5 and 1 respectively, to give two variables that are highly significant, eight that are 'small effects' and ten irrelevant regressors.

Table 4, panel b, reports the results with bias and RMSE reported as an average across (i) highly significant variables, (ii) small effects, and (iii) irrelevant variables. For variables with large non-centralities, the sequential procedure delivers a smaller RMSE than the joint procedure. Selection picks up the large effects and forces them in the second stage, thus avoiding any contamination by the retention of principal components. The RMSE of the sequential procedure is also smaller for irrelevant variables, as the principal components are only calculated over variables omitted from the first stage. so a lower weight is placed on them in selection. This leads to a smaller bias as well as RMSE. The joint procedure has a lower RMSE than the sequential procedure for the small effect variables. For the given correlation structure, the principal components can proxy the small effects well, and indeed, the RMSE of the joint

procedure after bias correction is close to that of just estimating the GUM for small effects, so the benefits of the joint procedure dominate when the DGP contains many small effects and few large effects or irrelevant variables.

Finally, we consider the case where *Autometrics* undertakes expanding as well as contracting searches by ensuring $N \geq T$. We set $n = 50$ with the first 10 parameters calibrated to give t-statistics of 5, the next 20 parameters calibrated to give t-statistics of 1, and the final 20 parameters calibrated to give t-statistics of 0. Thus, there are $2n = 100$ regressors in the GUM and $T = 100$:

$$\boldsymbol{\beta}' = \left(a \times \mathbf{1}'_{10}, b \times \mathbf{1}'_{20}, 0 \times \mathbf{1}'_{10}\right),$$

where $a = 2.18, b = 0.44$ and $\rho = 0.9$.

Table 4, panel c, reports the results, averaging across variables with ts of $5, 1$ and 0. Ensuring expanding as well as contracting searches increases the probability of retaining the highly significant variables (above 90%) with the first principal component also being retained 50% of the time for the joint procedure. This results in a smaller bias for the highly significant variables, although selecting from the model with both variables and factors is worse than selecting from the model with just variables. The advantage to including factors is observed in the 'small effects' regressors, which have a smaller bias and RMSE relative to selecting from the variables, and can even dominate estimation of the GUM after bias correction. The main benefit to selection comes from eliminating irrelevant variables, but the sequential procedure delivers RMSEs close to that of selection from just variables, so additional searching for relevant factors is not very costly and yields smaller RMSEs for the small effects. Both procedures retain fewer relevant variables than the DGP, with very few additional factors being retained under the sequential procedure.

Pure 'factor structure' DGPs

The above results suggest that the sequential procedure works well as a 'diagnostic test', picking up additional common effects once the main determinants have been selected and forced. However, if the DGP has a pure factor structure, the joint procedure should be able to retain the factors and exclude individual regressors, whereas any erroneously retained regressors in the sequential procedure will be omitted from the principal components, resulting in poor proxies for the DGP factors. We also investigated this scenario by Monte Carlo analysis.

The DGP is given by:

$$y_t = \beta_1 \widehat{z}_{1,t} + \beta_2 \widehat{z}_{2,t} + \nu_t, \quad t = 1, \ldots, T \tag{29}$$

where $\widehat{\mathbf{z}}_t$ is given by (3) and the data are generated by (23). We set $n = 10$, $T = 100$, $M = 10000$, $\beta_1 = \beta_2 = 1$ and $\rho = 0.9$. We consider both the joint and sequential procedure.

We evaluate the results based on the retention of variables, including factors, given by (11). In this scenario we can also evaluate on potency and gauge, although they will not be comparable across procedures because N differs in the sequential procedure. We also compute the bias and RMSE of the coefficients on the individual variables by comparing to the theory DGP coefficient values from the eigenvectors of the correlation matrix. From (29) we obtain:

$$y_t = \beta_1 \widehat{\mathbf{H}}_{(1)} \widetilde{\mathbf{x}}_t + \beta_2 \widehat{\mathbf{H}}_{(2)} \widetilde{\mathbf{x}}_t + \nu_t \tag{30}$$

where $\widehat{\mathbf{H}}_{(k)}$ corresponds to the kth eigenvector. The DGP coefficient values for $\mathbf{x}_t$ given the above parameters are $\boldsymbol{\kappa} = (0.63, -0.42, -0.42, -0.42, -0.42, -0.42, -0.42, -0.42, -0.42, -0.42)'$.

Figure 3 records the potency and gauge in panels a and b corresponding to selection from the factor model (24), the joint and sequential procedures. Panel c records the retention rate of the relevant factors, $\widehat{z}_{1,t}$ and $\widehat{z}_{2,t}$ in the first two columns, and the average retention of the irrelevant individual variables,

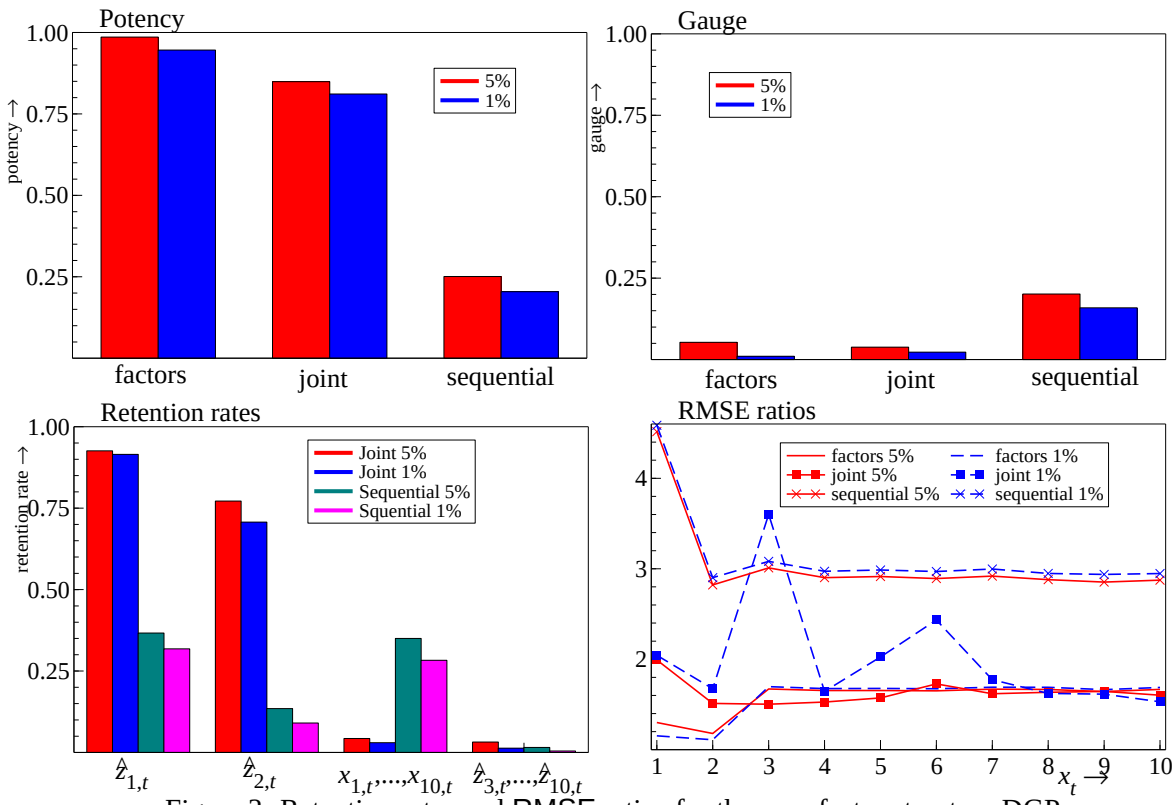

Figure 3: Retention rates and RMSE ratios for the pure factor structure DGP

$x_{1,t}, \ldots, x_{10,t}$, and irrelevant factors $\widehat{z}_{3,t}, \ldots, \widehat{z}_{10,t}$ in the final two columns. Panel d records the RMSE ratio against estimating the DGP and calculating the corresponding coefficients from (30). A ratio greater than one indicates a larger RMSE than the estimated DGP.

The potency and gauge of selection of the factors model correspond to the properties of selection for individual regressors. Potency is close to one reflecting the large non-centralities of the two factors, and the gauge is close to 5% and 1% at a selection significance level of $\alpha = 0.05$ and 0.01 respectively. Hence, if a principal component GUM is postulated for a factor DGP, selection can proceed as normal, and there is no discernable difference between factors and variables despite estimating the principal components in advance of selection. The joint procedure has a potency that is lower than selecting the principal components alone, due to a slightly lower retention rate for the second principal component, which has a smaller non-centrality. The gauge is slightly undersized at $\alpha = 0.05$ and oversized at $\alpha = 0.01$, but there is not a significant difference between retention of variables and principal components. The sequential procedure is not designed for a case where the factors enter the DGP, as can be seen by the potency and gauge. Variables are retained as they enter the relevant principal components, resulting in fewer principal components being retained in the second stage. Judging the procedures on RMSE indicates that this can be costly if, indeed, the DGP has a pure factor structure.

7 Conclusions

Factor structure models are frequently used to condense a large amount of information into a parsimonious form, and there is a substantial literature discussing the merits of such methods for modelling and forecasting. An alternative methodology for obtaining a more parsimonious model is to use model selection. This paper attempts to reconcile the two approaches to dimension reduction by undertaking

selection of both variables and principal components. The paper does not take a stand on the appropriate DGP structure. Rather, we propose a general method that allows for either individual variables, common factors, or a combination of both. We motivate the use of principal components by demonstrating that they can capture 'small effects', resulting in combinations of variables with larger non-centralities than the individual regressors alone. However, this is dependent on the correlation structure between the regressors and the dependent variable being similar to the correlation structure within the regressors.

Two strategies are considered. A joint procedure selects from a GUM that contains both variables and principal components, resulting in perfect collinearity. We demonstrate that this is not a problem for a model selection algorithm such as *Autometrics* that undertakes a comprehensive multi-path search. Simulation evidence suggests that the procedure works well when the DGP has either a factor structure or contains many 'small effect' variables that are well proxied by principal components. The second strategy proposed is a sequential procedure, which acts like a diagnostic test, first selecting relevant regressors and then using principal components of the remaining variables to 'mop up' any remaining systematic variation. This procedure enables the relevant variables to be pinned down initially, so outperforms the joint procedure when there are highly relevant variables and irrelevant variables. However, if the DGP has a pure factor structure, the sequential procedure will have little power.

To conclude, traditional regression models with individual variables and principal component models need not be treated separately, but can be combined using a selection procedure that can handle perfect collinearity. This allows the data to determine whether the underlying DGP has a factor structure, or the regressand can instead be explained by a few variables. Extensions to non-linear models, dynamics and multiple breaks are feasible by combining the present approach with the earlier results mentioned above.

References

Abadir, K. M., Distaso, W., and Žikeš, F. (2010). Model-free estimation of large variance matrices. Working paper series 17-10, Rimini Centre for Economic Analysis, Rimini, Italy.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171.

Banerjee, A., Marcellino, M., and Masten, I. (2008). Forecasting macroeconomic variables using diffusion indexes in short samples with structural change. In Rapach, D. E., and Wohar, M. E.(eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty: Frontiers of Economics and Globalization Volume 3*, pp. 149–194. Bingley, UK: Emerald Group.

Belsley, D. A., and Klema, V. (1974). Detecting and assessing the problems caused by multi-collinearity: A use of the singular-value decomposition. NBER working papers 0066, National Bureau of Economic Research, Cambridge, MA.

Bontemps, C., and Mizon, G. E. (2008). Encompassing: Concepts and implementation. *Oxford Bulletin of Economics and Statistics*, **70**, 721–750.

Campos, J., and Ericsson, N. R. (1999). Constructive data mining: Modeling consumers' expenditure in Venezuela. *Econometrics Journal*, **2**, 226–240.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2011a). Evaluating automatic model selection. *Journal of Time Series Econometrics*, **3(1)**, DOI: 10.2202/1941-1928.1097.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2011b). Model selection when there are multiple breaks. *Journal of Econometrics*, forthcoming.

Castle, J. L., and Hendry, D. F. (2010a). A low-dimension, portmanteau test for non-linearity. *Journal of Econometrics*, **158**, 231–245.

Castle, J. L., and Hendry, D. F. (2010b). Model selection in under-specified equations with breaks. Discussion paper 509, Economics Department, Oxford University.

Castle, J. L., and Hendry, D. F. (2011a). Automatic selection of non-linear models. In Wang, L., Garnier, H., and Jackman, T.(eds.), *System Identification, Environmental Modelling and Control*, forthcoming. New York: Springer.

Castle, J. L., and Hendry, D. F. (2011b). A tale of 3 cities: Model selection in over-, exact, and under-specified equations. Discussion paper 523, Economics Department, Oxford University.

Castle, J. L., and Shephard, N.(eds.)(2009). *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Oxford: Oxford University Press.

Chamberlain, G., and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, **51**, 1305–1324.

Doornik, J. A. (2007). *Object-Oriented Matrix Programming using Ox* 6th ed. London: Timberlake Consultants Press.

Doornik, J. A. (2009a). Autometrics. in Castle, and Shephard (2009), pp. 88–121.

Doornik, J. A. (2009b). Econometric model selection with more variables than observations. Working paper, Economics Department, University of Oxford.

Favero, C., Marcellino, M., and Neglia, F. (2005). Principal components at work: the empirical analysis of monetary policy with large datasets. *Journal of Applied Econometrics*, **20**, 603–620.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized factor model: Identification and estimation. *The Review of Economics and Statistics*, **82**, 540–554.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In Mills, T. C., and Patterson, K. D.(eds.), *Palgrave Handbook of Econometrics*, pp. 3–67. Basingstoke: Palgrave MacMillan.

Hendry, D. F., and Doornik, J. A. (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.

Hendry, D. F., and Johansen, S. (2010). Model selection when forcing retention of theory variables. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, **33**, 317–335. Erratum, 337–339.

Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

Hendry, D. F., and Mizon, G. E. (1978). Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. *Economic Journal*, **88**, 549–563.

Hendry, D. F., and Mizon, G. E. (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, **3(1)**, DOI: 10.2202/1941-1928.1100.

Hendry, D. F., and Richard, J.-F. (1989). Recent developments in the theory of encompassing. In Cornet, B., and Tulkens, H.(eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, MA: MIT Press.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. in Castle, and Shephard (2009), pp. 1–36.

Jones, C. S. (2006). A nonlinear factor analysis of S&P 500 index option returns. *The Journal of Finance*, **61(5)**, 2325–2363.

Mayo, I., and Espasa, A. (2009). Forecasting aggregates and disaggregates with common features. Working paper, Universidad Carlos III, Madrid.

Mizon, G. E., and Richard, J.-F. (1986). The encompassing principle and its application to non-nested

hypothesis tests. *Econometrica*, **54**, 657–678.

OECD (2008). Handbook on constructing composite indicators: Methodology and user guide. ISBN: 978-92-64-04345-9. http://www.oecd.org/dataoecd/37/42/42495745.pdf.

Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics*, **4**, 393–397.

Sargan, J. D. (1980). Some tests of dynamic specification for a single equation. *Econometrica*, **48**, 879–897. Reprinted as pp. 191–212 in Sargan J. D. (1988), *Contributions to Econometrics*, Vol. 1, Cambridge: Cambridge University Press.

Stock, J. H., and Watson, M. W. (1998). Diffusion indexes. Working paper No. 6702, NBER.

Stock, J. H., and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business and Economic Statistics*, **20**, 147–162.

Stock, J. H., and Watson, M. W. (2009). Forecasting in dynamic factor models subject to structural instability. in Castle, and Shephard (2009), Ch. 7.

Yalcin, I., and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical Science*, **16**, 275–294.

Table 4: Model selection with variables and principal components.

	[A]	[B]	[C]	[B]	[C]	[G]	[H]	[G]	[H]	[I]	[J]	[I]	[J]
α		5%	5%	1%	1%	5%	5%	1%	1%	5%	5%	1%	1%
Panel a: $n = 10$; $\psi_i = 5$ for $i = 1, 2$, $\psi_i = 1$ for $i = 3, \ldots, 10$													
$\widehat{\sigma}$	0.998	0.996	0.996	1.012	1.012	0.996	0.996	1.011	1.011	0.992	0.992	1.005	1.005
$\widetilde{n}$		4.594	4.594	3.891	3.891	2.377	2.377	2.180	2.180	4.594	4.594	3.891	3.891
$\overline{n}$						1.907	1.907	1.555	1.555	0.166	0.166	0.156	0.156
Bias $\psi_i = 5$	-0.008	17.89	17.44	30.09	29.40	-2.149	-4.395	-3.236	-5.696	14.35	13.69	23.96	22.66
Bias $\psi_i = 1$	-0.020	-4.749	-9.958	-7.966	-11.78	0.433	-0.112	0.633	0.046	-3.798	-10.66	-6.332	-12.01
RMSE $\psi_i = 5$	0.317	0.383	0.389	0.467	0.473	0.445	0.462	0.536	0.550	0.373	0.383	0.441	0.456
RMSE $\psi_i = 1$	0.319	0.406	0.366	0.441	0.400	0.344	0.335	0.360	0.349	0.386	0.350	0.413	0.374
Panel b: $n = 20$; $\psi_i = 5$ for $i = 1, 2$, $\psi_i = 1$ for $i = 3, \ldots, 10$, $\psi_i = 0$ for $i = 11, \ldots, 20$													
$\widehat{\sigma}$	0.996	0.979	0.979	1.001	1.001	0.994	0.994	1.032	1.032	0.976	0.000	0.998	0.000
$\widetilde{n}$		5.044	5.044	4.047	4.047	2.942	2.942	2.480	2.480	5.044	5.044	4.047	4.047
$\overline{n}$						2.548	2.548	1.577	1.577	0.129	0.129	0.055	0.055
Bias $\psi_i = 5$	-0.236	13.56	12.99	25.69	24.89	-5.726	-8.819	-7.568	-10.59	13.13	12.52	24.95	24.07
Bias $\psi_i = 1$	-0.044	-8.666	-12.95	-11.57	-14.69	-6.686	-8.010	-9.377	-10.52	-8.206	-12.78	-11.30	-14.62
Bias $\psi_i = 0$	0.095	4.082	3.207	3.838	3.091	6.414	6.716	8.836	8.917	3.809	2.987	3.782	3.029
RMSE $\psi_i = 5$	0.346	0.375	0.383	0.448	0.455	0.504	0.527	0.668	0.688	0.374	0.383	0.444	0.453
RMSE $\psi_i = 1$	0.347	0.395	0.361	0.424	0.389	0.360	0.345	0.364	0.346	0.395	0.360	0.422	0.387
RMSE $\psi_i = 0$	0.346	0.223	0.176	0.200	0.166	0.319	0.299	0.302	0.283	0.233	0.181	0.207	0.169
Panel c: $n = 50$; $\psi_i = 5$ for $i = 1, \ldots, 10$, $\psi_i = 1$ for $i = 11, \ldots, 30$, $\psi_i = 0$ for $i = 31, \ldots, 50$													
α		1%	1%	0.5%	0.5%	1%	1%	0.5%	0.5%	1%	1%	0.5%	0.5%
$\widehat{\sigma}$	0.992	1.031	1.031	1.068	1.068	0.996	0.996	1.029	1.029	1.024	1.024	1.056	1.056
$\widetilde{n}$		15.64	15.64	14.70	14.70	13.32	13.32	12.71	12.71	15.57	15.57	14.69	14.69
$\overline{n}$						2.899	2.899	2.506	2.506	0.161	0.161	0.162	0.162
Bias $\psi_i = 5$	-0.603	22.87	22.32	28.45	27.72	10.67	8.759	12.41	9.607	22.01	21.49	26.07	25.29
Bias $\psi_i = 1$	-0.176	-15.95	-20.55	-18.34	-22.78	-12.60	-16.58	-14.15	-18.27	-15.64	-20.64	-17.79	-22.92
Bias $\psi_i = 0$	0.469	4.372	3.470	3.946	3.093	7.220	7.024	7.882	7.633	4.517	3.587	4.615	3.632
RMSE $\psi_i = 5$	0.454	0.517	0.525	0.566	0.574	0.528	0.559	0.562	0.604	0.506	0.514	0.537	0.547
RMSE $\psi_i = 1$	0.445	0.548	0.507	0.559	0.516	0.453	0.433	0.455	0.433	0.538	0.497	0.543	0.501
RMSE $\psi_i = 0$	0.448	0.250	0.204	0.238	0.195	0.337	0.303	0.319	0.287	0.250	0.202	0.236	0.189

Notes: $\psi_i = 5$ averages across variables with population t-statistics of 5, $\psi_i = 1$ averages across variables with population t-statistics of 1 and $\psi_i = 0$ averages across irrelevant variables.