

Model Selection in Equations with Many ‘Small’ Effects

JENNIFER L. CASTLE[†], JURGEN A. DOORNIK[‡] and DAVID F. HENDRY[‡]

[†]*Magdalen College and Institute for New Economic Thinking at the Oxford Martin School,
University of Oxford, UK, (jennifer.castle@magd.ox.ac.uk)*

[‡]*Economics Department and Institute for New Economic Thinking at the Oxford Martin School,
University of Oxford, UK, (jorgen.doornik@economics.ox.ac.uk; david.hendry@economics.ox.ac.uk)*

Abstract

High dimensional general unrestricted models (GUMs) may include important individual determinants, many small relevant effects, and irrelevant variables. Automatic model selection procedures can handle 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, possibly representing ‘factor’ structures, as perfect collinearity is also unproblematic. ‘Factors’ can capture small influences that selection may not retain individually. The final model can implicitly include more variables than observations, entering via ‘factors’. We simulate selection in several special cases to illustrate.

JEL classifications: C52, C22.

KEYWORDS: Model selection, high dimensionality, principal components, Monte Carlo.

1 Introduction

Macroeconomic time-series are complicated processes, 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. High dimensional initial models are therefore likely, where the potential set of explanatory variables may include individual determinants with significant explanatory power, irrelevant variables, and relevant variables that may have small effects individually that would not be significant at conventional levels, and hence not retained when selection is undertaken, distorting inference (see Leeb and Pötscher, 2003, for an analysis). Since theory models will always abstract from various aspects of reality, selection is inevitable, so we address how this third group could be captured in part by combining variables with small relevant effects to increase their joint explanatory power, and hence raise the probability of retention. We propose doing so by capturing the otherwise unexplained co-movements of the observable time series using their principal components, which could also embody relevant common forces.

Model selection is then applied from a general unrestricted model (GUM). Absent omniscience, that GUM needs to include all the individual variables as well as their principal components, so will be perfectly collinear. We exploit the ability of automatic selection algorithms to handle such a problem

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

(see e.g., Hendry and Krolzig, 2005, Doornik, 2009a). Alternatively, significant individual variables could be selected first and then principal components computed for the non-retained variables to capture additional small effects. Both procedures are evaluated below.

The structure of the paper is as follows. Section 2 describes the reductions involved when the model is an over-specification of the DGP, with some substantively relevant and some irrelevant variables, as well as many small relevant effects. Section 3 considers representing the last group by their principal components; Section 4 considers the issues of perfect collinearity and more variables than observations introduced by this approach. Sections 5–7 examine Monte Carlo evidence, evaluating the properties of selection: (i) under the null when no variables or factors are relevant; (ii) when principal components are used to parsimoniously approximate many small effects; (iii) when there are both individually relevant variables and small effects, as well as irrelevant variables. Section 8 concludes.

2 Dimension reduction

To formalize 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, both with up to s lags, and $1_{\{i=t\}}, t = 1, \dots, T$ are a saturating set of impulse indicators, then the GUM is:

$$y_t = \sum_{i=1}^n \sum_{j=0}^s \beta_{i,j} x_{i,t-j} + \sum_{i=1}^n \sum_{j=0}^s \kappa_{i,j} z_{i,t-j} + \sum_{j=1}^s \theta_j y_{t-j} + \sum_{i=1}^T \delta_i 1_{\{i=t\}} + e_t \quad (1)$$

resulting in $N > T$ regressors. Of these, L are relevant as defined by non-zero non-centralities of the population t-values in the local data generating process (LDGP: the DGP for the set of variables under consideration, see e.g., Hendry, 2009):

$$y_t = \sum_{i=1}^K \phi_i u_{i,t} + \epsilon_t \quad (2)$$

where $\phi_i \neq 0$ when the $u_{i,t}$ denote the set of relevant variables, with $K \geq L$ as (1) may also omit relevant effects (such as parameter changes), and $\epsilon_t \sim \text{IN}[0, \sigma_\epsilon^2]$.

As we anticipate 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’ represented by their principal components.

Our focus is on the last item only, but we start with a brief discussion of the first four items.

1. We employ *Autometrics* (see Doornik, 2009a, embodied in *PcGive*, Hendry and Doornik, 2009), which is an automatic model selection algorithm that seeks to locate the LDGP. A multi-path general-to-specific search is undertaken, eliminating irrelevant variables while ensuring empirical congruency. The resulting selected model is a valid restriction of the general model, encompassing all other models that are also valid restrictions, also see Doornik (2008).

Castle, Doornik, and Hendry (2011a) discuss the general approach of *Autometrics* and establish its excellent properties when the GUM nests the data-generating process, 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 (2012) propose a procedure for retaining theory-based specifications when there are more variables than observations.

2. Castle, Doornik, and Hendry (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, or use the standard reduction approach of *Autometrics*.

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).

4. Castle and Hendry (2011b) describe an automatic algorithm for non-linearity which includes quadratic, cubic and exponential functions of the principal components, following the test for non-linearity in Castle and Hendry (2010a).

The aim is to select all relevant variables and eliminate all irrelevant effects, thereby locating (2) when it is nested in (1). Conversely, the trade-off is between retaining irrelevant and omitting relevant variables, which depends on the chosen significance level, α , and the non-centrality, ψ_i , of the t -distribution for the t -test on the relevant $\phi_i \neq 0$ (see, e.g., Davidson and MacKinnon, 2004, §4.7). 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 $\alpha(N - L) = 1$: on average almost all (999) irrelevant variables will be eliminated. The probability of retaining relevant variables rises the more their non-centralities exceed the corresponding critical value c_α , such that $|t| \geq c_\alpha$. For variables with non-centralities $\psi < 2$ at the available sample size, the probability of retention: $\Pr(|t_{\psi^2 < 4}| \geq c_\alpha)$ would be small at $\alpha = 0.001$, and hence such variables, while relevant, would rarely be retained. When mis-specification tests are undertaken to check congruence, irrelevant variables can be retained to offset chance significant diagnostic tests, but usually only by a small amount.

Here we consider approximating the effects of many small influences, ignoring the additional complications of lags and impulse indicators presented in (1).

3 Model selection with principal components

Principal components

Let us collect the set of candidate variables, which have been transformed to non-integratedness by appropriate differencing, in a $(T \times n)$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)'$, where $\mathbf{x}_t = (x_{1,t}, \dots, x_{n,t})'$. Denoting the standardized¹ data by $\tilde{\mathbf{X}}$, we can take the eigenvalue decomposition of the sample correlation matrix:

$$\hat{\mathbf{C}} = T^{-1} \tilde{\mathbf{X}}' \tilde{\mathbf{X}} = \hat{\mathbf{H}} \hat{\mathbf{\Lambda}} \hat{\mathbf{H}}'$$

where $\hat{\mathbf{\Lambda}}$ is the diagonal matrix of ordered eigenvalues ($\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_n \geq 0$) and $\hat{\mathbf{H}} = (\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_n)$ is the corresponding matrix with eigenvectors in columns, normalized to $\hat{\mathbf{H}}' \hat{\mathbf{H}} = \mathbf{I}_n$. The sample principal components are the n columns of the $(T \times n)$ matrix:

$$\hat{\mathbf{Z}} = \tilde{\mathbf{X}} \hat{\mathbf{H}} = (\tilde{\mathbf{X}} \hat{\mathbf{h}}_1, \dots, \tilde{\mathbf{X}} \hat{\mathbf{h}}_n), \quad (3)$$

with the property that $\hat{\mathbf{Z}}' \hat{\mathbf{Z}} = T \hat{\mathbf{\Lambda}}$. The principal components are orthogonal transformations of the original data used to approximate potentially relevant combinations of variables. So including both principal components and the individual variables in the model will create perfect collinearity.

¹ $\tilde{x}_{i,t} = (x_{i,t} - \bar{x}_i) / \tilde{\sigma}_{x_i}$ $i = 1, \dots, n$ with $\bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{i,t}$ and $\tilde{\sigma}_{x_i}^2 = \frac{1}{T} \sum_{t=1}^T (x_{i,t} - \bar{x}_i)^2$.

Equations with ‘factor’ structures

Factor models are one solution to dimensionality constraints, and 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. Banerjee, Marcellino, and Masten (2008), Stock and Watson (2009) and Castle, Clements, and Hendry (2011) consider the related issue of ‘factor’ forecasts facing structural change. Approximate factor models relax the assumptions of serially uncorrelated and homoskedastic idiosyncratic errors of the static factor model by assuming $N \rightarrow \infty$ (see Chamberlain and Rothschild, 1983, and Stock and Watson, 2002), which still ensures the principal components estimator is consistent and asymptotically normal (see Bai, 2003).

There are two distinct states of nature for a ‘factor structure’ DGP:

- (i) the DGP is a function of ‘common trends’ that are modelled as latent variables; and
- (ii) the DGP contains many small relevant effects that can be approximated by latent factors.

The latter is the case of interest here, so we abstract from many of the issues of factor analysis, namely determining the number of factors, the factor loadings, and the idiosyncratic components.

The simplest DGP over $t = 1, \dots, T$ that focuses on case (i) is given by (abstracting from indicators, lags or other transformations):

$$y_t = \mu + \beta' \mathbf{x}_t + \epsilon_t \quad (4)$$

where $\epsilon_t \sim \text{IN}[0, \sigma_\epsilon^2]$. The $n \ll T$ valid conditioning variables \mathbf{x}_t are not perfectly collinear, and are independent of $\{\epsilon_t\}$. β contains non-zero elements (relevant variables) and zero elements (irrelevant variables). Of the non-zero elements, some β s have small, but non-zero, non-centralities of their absolute t-statistics, namely less than the critical value, c_α , for the chosen significance level.

Joint selection procedure

We consider two alternative selection procedures as follows. In the first procedure, denoted the ‘joint’ procedure, the GUM will include all the individual regressors and their principal components (which have mean zero) computed from (3), leading to $2n + 1$ regressors:

$$y_t = \mu + \gamma' \mathbf{x}_t + \delta' \hat{\mathbf{z}}_t + \nu_t. \quad (5)$$

so the joint set $(1 : \mathbf{x}_t : \hat{\mathbf{z}}_t)$ must be perfectly collinear, with n collinearities. We have re-labelled the coefficients of (4) from β to $\theta = (\gamma : \delta)$ because several principal components may be needed, either alone or in combination with individual regressors.

We solve out the retained principal components to retrieve the original specification in terms of \mathbf{x}_t , in order to evaluate the impact of selection. After selection, denote retained \mathbf{x}_t and $\hat{\mathbf{z}}_t$ by \mathbf{x}_t^r and $\hat{\mathbf{z}}_t^r$ respectively, with estimated coefficients $\tilde{\gamma}_r$ and $\tilde{\delta}_r$:

$$y_t = \tilde{\mu} + \tilde{\gamma}_r' \mathbf{x}_t^r + \tilde{\delta}_r' \hat{\mathbf{z}}_t^r + \tilde{\nu}_t$$

where from (3) the retained principal components are the relevant columns of $\hat{\mathbf{Z}}$ with $\hat{\mathbf{H}}_r$ the corresponding eigenvectors:

$$\hat{\mathbf{z}}_t^r = \hat{\mathbf{H}}_r' \tilde{\mathbf{x}}_t$$

Solving out from the principal components in terms of the original set of variables results in:

$$y_t = \tilde{\mu} + \tilde{\gamma}_r' \mathbf{x}_t^r + \tilde{\delta}_r' \hat{\mathbf{H}}_r' \tilde{\mathbf{x}}_t + \tilde{\nu}_t = \tilde{\mu}_s + \tilde{\beta}_s' \mathbf{x}_t + \tilde{\nu}_t. \quad (6)$$

Sequential selection procedure

An alternative procedure, denoted the ‘sequential’ procedure, first selects the variables, then adds the principal components of the omitted variables for a second round selection.

The GUM in the first stage consists of the variables only. Let \mathbf{x}_t^r denote the variables retained in the first-stage selection using significance level α_1 , and define the set of excluded variables as $\mathbf{x}_t^o = \{S : S \in \mathbf{x}_t, S \notin \mathbf{x}_t^r\}$. Principal components are then computed from the omitted set:

$$\tilde{\mathbf{z}}_t^o = \hat{\mathbf{H}}_o' \tilde{\mathbf{x}}_t^o. \quad (7)$$

In the second stage, selection is undertaken over the \mathbf{x}_t^{rr} and $\tilde{\mathbf{z}}_t^o$ at significance α_2 . Denote the selected set of variables as \mathbf{x}_t^{rr} , and factors by $\tilde{\mathbf{z}}_t^{or}$, so:

$$y_t = \bar{\mu} + \bar{\gamma}_o' \mathbf{x}_t^{rr} + \bar{\delta}_o' \tilde{\mathbf{z}}_t^{or} + \bar{\eta}_t.$$

Solving out for the principal components as in (6) yields:

$$y_t = \bar{\mu} + \bar{\gamma}_o' \mathbf{x}_t^{rr} + \bar{\delta}_o' \hat{\mathbf{H}}_o' \tilde{\mathbf{x}}_t^o + \bar{\eta}_t = \bar{\mu}_s + \bar{\beta}_s' \mathbf{x}_t + \bar{\eta}_t. \quad (8)$$

This method allows retained ‘large effects’ from the first round to be altered by additional ‘small effects’ captured by the principal components. Choosing α_1 smaller than α_2 will potentially offer more influence to the factors.

4 Perfect collinearity

The previous section noted that a linear model containing both factors and variables is perfectly collinear: all factors (or all variables) can be removed without affecting the likelihood. When there are more observations than regressors, our OLS procedure will identify the singular subset and assign them a coefficient of zero.²

Adding the factors can also create perfect collinearity when it results in more regressors than the sample size. In IIS, Section 2, T impulse dummies are added, and the theoretical analysis can proceed because they are added in blocks. Selection is done from each block, after which the surviving impulses are added jointly for a further selection. Both cases of perfect collinearity are handled by *Autometrics*.

More variables than observations and sparsity

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

²The order in which this happens is not under our direct control, but depends on the data through the size of the pivot in the QR method, see Golub and Van Loan (1996, §5.5).

Selection from a GUM with more variables than observations appears to hinge on the sparsity of the LDGP parameter vector ϕ in (2). 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 entails more non-zero elements than T , as in (6) where $\dim(\gamma_r) + \dim(\delta_r) < T$ yet nothing precludes $\dim(\beta_s) > T$. Thus, when selecting over principal components in conjunction with variables, sparsity in the final model is no longer required.

Selection can be applied when there are more regressors than observations. However, increasing the dimension will require selection to be undertaken at a tighter significance level to avoid over-fitting, and this will reduce the probability of detecting any individually small yet relevant effects.

Multi-path selection with perfect collinearity

To illustrate the analysis, consider the simple DGP:

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

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

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

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 keep $(x_{1,t} - x_{2,t})$ but also drop $x_{1,t}$ as now insignificant; similarly for the third path commencing without $x_{1,t}$.

The cost of doubling the number of variables n by also including n perfectly collinear combinations is that the procedure will now retain approximately $2\alpha n$ irrelevant regressors: in a t-test based approach, combinations can be significant under the null even when their components would not be (see Castle and Hendry, 2011a). This is pertinent when the joint procedure leads to $N \geq T$ and the blocks search is used.

5 Evaluation and design of the Monte Carlo experiments

Evaluating the selected model

Selection can be evaluated by its success at retaining ‘relevant’ and excluding ‘irrelevant’ variables and factors, provided both variables and latent factors enter the DGP directly as part of the explanatory regressors. Let the first L regressors in (1) be relevant, with the remaining $N - L$ irrelevant. Let $\tilde{\beta}_{i,m}^*$ denote the OLS coefficient estimate after selection for the coefficient on the i th regressor in replication m , with M replications. A variable that is not selected has corresponding entrance in $\tilde{\beta}_{i,m}^*$ set to zero. When $1(\cdot)$ is the indicator variable, *potency* and *gauge* respectively calculate the retention frequencies of relevant and irrelevant variables (ignoring the intercept which is always kept in the model) as:

$$\begin{aligned} \text{retention rate: } \tilde{p}_i &= \frac{1}{M} \sum_{m=1}^M 1(\tilde{\beta}_{i,m}^* \neq 0), \quad i = 1, \dots, N, \\ \text{potency: } p &= \frac{1}{L} \sum_{i=1}^L \tilde{p}_i, \\ \text{gauge: } g &= \frac{1}{N-L} \sum_{i=L+1}^N \tilde{p}_i. \end{aligned}$$

If the latent factors do not directly enter the DGP, but principal components approximate combinations of relevant variables in the DGP, then the potency and gauge of these are not useful measures:

retained principal components would be counted in the gauge yet may also contribute to potency. Consequently, we evaluate selection based on the conditional ‘solved out’ estimated coefficients from (6) and (8). Let $\tilde{\beta}_{i,m}^*$ now denote the solved out OLS coefficient when principal components are involved. Biases and root mean squared errors (RMSE) can then be computed in the standard way:

$$\text{bias}_i = \frac{1}{M} \sum_{m=1}^M \left(\tilde{\beta}_{i,m}^* - \beta_i \right), \quad (9)$$

$$\text{RMSE}_i = \left[\frac{1}{M} \sum_{m=1}^M \left(\tilde{\beta}_{i,m}^* - \beta_i \right)^2 \right]^{1/2}. \quad (10)$$

Data generation process

The DGP is always given by (4), with $\mu = 0$, and regressors generated by:

$$\mathbf{x}_t \sim \text{IN}_n \left[\mathbf{0}, \sigma_x^2 \mathbf{C}_x \right], \quad t = 1, \dots, T,$$

where $c_{i,i} = 1$ for $i = 1, \dots, n$ and $c_{i,j} = \rho$, $\forall i \neq j$, and $\sigma_x^2 = 1$. Variations of this DGP are created through the specification of the coefficients β , the dimensionality n , the value of ρ , and the sparsity (i.e. zeros in β). We use sample size $T = 100$ and $M = 10,000$ replications in all reported simulations, with regressors drawn independently for each replication (i.e., not fixed in repeated samples). The experiments are run in Ox 6, Doornik (2007), using PcGive 13 for selection and estimation.

Estimation approaches

Model selection procedures are applied starting from the following GUMs:

- (1) n variables only: (5) with $\delta = \mathbf{0}$;
- (2) n factors only: (5) with $\gamma = \mathbf{0}$;
- (3) joint procedure: $2n$ variables and factors, (5);
- (4) sequential procedure: n variables first at α_1 , then extending by principal components from the omitted variables, reselecting at α_2 .

The model selection algorithm used is *Autometrics*, with no diagnostic checking, where selection from assuming (4) is known provides an upper bound on performance.³ The significance levels for selection are set to $\alpha = 5\%$ and $\alpha = 1\%$ (if no selection takes place we write $\alpha = 100\%$). All estimated models include an intercept that is ignored in the evaluation.

Null retention frequency (gauge)

The baseline experiment sets $\beta = \mathbf{0}$ in (4) to establish the gauge when principal components $\hat{z}_{i,t}$ are entered. We set $\rho = 0.5$ and 0.9 . Table 1 records the gauge, showing that the inclusion of principal components instead of variables has no effect. In both cases, the gauge is close to the nominal significance level for small n , but slightly under for $n = 50$.

Inclusion of both regressors and principal components in the joint procedure results in perfect collinearity. We compute the gauge as the retention rate out of $2n$ regressors.⁴ For small n , only non-singular reduction paths are searched, which reduces the gauge relative to the nominal significance level. When $n = 50$, there are as many regressors as observations, so expanding as well as contracting

³*Autometrics* is also used in the factors only case. Because the factors are orthogonal, selection according to t -values would be a simple alternative.

⁴The gauge doubles if we take the perspective that there are only n unique variables.

TABLE 1:
Gauge for selection with (i) variables, (ii) principal components,
and (iii) jointly; all coefficients in the DGP are zero, $T = 100$.

		variables only		factors only		joint procedure	
		$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$
$n = 2$	$\rho = 0.5$	0.0528	0.0115	0.0507	0.0103	0.0201	0.0039
$n = 10$	$\rho = 0.5$	0.0481	0.0088	0.0387	0.0064	0.0009	0.0001
$n = 50$	$\rho = 0.5$	0.0313	0.0047	0.0351	0.0032	0.0467	0.0110
$n = 2$	$\rho = 0.9$	0.0525	0.0113	0.0516	0.0106	0.0207	0.0037
$n = 10$	$\rho = 0.9$	0.0567	0.0128	0.0390	0.0064	0.0010	0.0001
$n = 50$	$\rho = 0.9$	0.0317	0.0057	0.0351	0.0032	0.0480	0.0130

block searches are used, which results in a gauge close to α . Thus, forcing a more complete search by doubling the number of irrelevant variables also doubles the number of adventitious retentions. Nevertheless, at tight significance levels, this will be a relatively small cost.

6 Monte Carlo evidence on small effects

We next investigate whether principal components can capture many small effects more effectively than variables. Models that combine factors and variables are considered in the next section.

The DGP in this section only has small effects. In the first case the DGP corresponds to a factor structure, while the second case uses alternating sign to remove the factor structure. In this section we use $n = 10$ for $T = 100$ and $\rho = 0.5, 0.9$.

Data generation processes

(a) 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:

$$\mathbf{x}_t \sim \text{IN}_n [\mathbf{0}, \sigma_x^2 \mathbf{C}] \quad (11)$$

where $\mathbf{C} = (1 - \rho) \mathbf{I}_n + \rho \iota \iota'$, so all the variables have a common correlation ρ . E.g., when $n = 3$:

$$\mathbf{C} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \quad (12)$$

which has an unnormalized eigenvector⁵ of $\mathbf{h}_1 = \iota = (1 \dots 1)'$ so there is a ‘factor’ $f_{1,t}$ such that:

$$\phi f_{1,t} = \phi \iota' \mathbf{x}_t = \beta' \mathbf{x}_t. \quad (13)$$

⁵In general, there is one eigenvalue of $(n - 1)\rho + 1$ and $n - 1$ of $1 - \rho$. The unnormalized eigenmatrix is:

$$\begin{pmatrix} 1 & n-1 & 0 & \cdots & 0 \\ 1 & -1 & n-2 & & 0 \\ 1 & -1 & -1 & & 0 \\ \vdots & \vdots & \vdots & & 1 \\ 1 & -1 & -1 & & -1 \end{pmatrix}.$$

The first column is a unique eigenvector up to scale and sign. Columns 2... n correspond to eigenvalue $1 - \rho$, so any non-trivial linear combination of these is also an eigenvector.

Therefore this one-factor DGP corresponds to the case where all coefficients have a common value β . This must be one of the most favourable cases, since (4) becomes:

$$y_t = \mu + \phi f_{1,t} + \epsilon_t. \quad (14)$$

The many small effects are represented by population t-statistics with non-centralities of unity: $\psi_i = 1, i = 1, \dots, 10$. This amounts to $\beta_i = 0.141$ and $\beta_i = 0.315$ for $n = 10$, $\sigma_x^2 = 1$ when $\rho = 0.5$ and $\rho = 0.9$ respectively. The Monte Carlo outcomes can be analyzed analytically in this case.⁶

(b) Next, consider the case when the link of y_t to \mathbf{x}_t in (4) does not match the correlation structure in (11), say the β_i alternate in sign with the same magnitude, so are $\pm\phi$. The other two unnormalized eigenvectors of (12) are (linear combinations of):

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

Now $\beta' \mathbf{x}_t = \phi(1 : -1 : 1) \mathbf{x}_t = (\phi/3)(\mathbf{h}_1 - 4\mathbf{h}_2 + 2\mathbf{h}_3)' \mathbf{x}_t$, which requires all three factors.

For $n = 10$ we set $\beta_i = a(-1)^{i+1}$, for $i = 1, \dots, n$ where a is calibrated to deliver $|\psi_i| = 1$ ($a = 0.141, 0.315$ for $\rho = 0.5, 0.9$).

Simulation results

Simulation evidence is presented in Table 2 and Figure 1. Reported are n_r :vars, the average number of variables retained, n_r :factors, the average number of principal components retained, Bias, the averaged bias (9), and RMSE, the averaged root mean squared error (10). In all cases where selection was used, $\hat{\sigma}_\epsilon$ is close to unity, so is not substantively downwards biased despite selection.

(a) When the correlation structure between regressors matches that of the correlation structure with y , the first principal component picks up most of the variation. The first panel of Table 2 confirms the advantage of representing the individually insignificant effects from \mathbf{x}_t by principal components in the ‘factor DGP’, with a vast reduction in RMSEs relative to just estimating the correct model.⁷

Figure 1a shows the retention rates for variables, which are all at 35%. Panel b, on the right, shows that the first principal component is always selected, and factors 2–10 5% of the time ($\alpha = 5\%$).

⁶By inverting $\Sigma = \sigma_x^2 \mathbf{C}$, the non-centrality of the t-test in (4) of the null hypothesis that $\beta_i = 0$ is:

$$\psi_i = E[t_{\beta_i=0} | \beta_i \neq 0] = E\left[\frac{\hat{\beta}_i}{SE[\hat{\beta}_i]}\right] \simeq \phi \frac{\sigma_x \sqrt{1 + (n-2)\rho} - (n-1)\rho^2}{\sigma_\epsilon \sqrt{1 + (n-2)\rho}}.$$

The non-centrality τ of the t-test of $\phi = 0$ in (14) is based on:

$$\tau = \frac{\hat{\phi}}{SE[\hat{\phi}]} \simeq \phi \frac{\sigma_x \sqrt{n(1 + (n-1)\rho)}}{\sigma_\epsilon}$$

as:

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

where:

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

Thus, at $\sigma_x^2 = \sigma_\epsilon^2 = 1$, $\rho = 0.5$ and $n = 3$, $\psi_i \simeq \sqrt{0.67}\phi$, whereas $\tau \simeq \sqrt{6}\phi$; and at $\rho = 0.9$, $\psi_i \simeq \sqrt{0.15}\phi$, whereas $\tau \simeq \sqrt{8.4}\phi$. The ratios τ/ψ are $\sqrt{9} = 3$ for $\rho = 0.5$ and $\sqrt{57.0} \simeq 7.5$ for $\rho = 0.9$ when $n = 3$. The ratios are 10 and 29 at $n = 10$ for $\rho = 0.5, 0.9$ (versus MC outcomes of 10.5 and 30 at $T = 100$); and 50 and 149 at $n = 50$. So with many small effects, principal components will be retained with a high probability relative to individual regressors.

⁷Coefficient estimates after selection will be biased. Hendry and Krolzig (2005) present an approximate correction, but we ignore this issue here.

TABLE 2:
Approximating small effects by principal components.
The DGPs are (a),(b) with $n = 10$, $T = 100$; $1(\cdot)$ is the indicator function. The GUM contains either all variables or all factors. Selection by *Autometrics* at $\alpha = 5\%$, 1% .

	α	variables only			factors only		
		100%	5%	1%	100%	5%	1%
(a) factor DGP: $\psi_i = 1, \rho = 0.5$							
n_r :vars	10		3.48	2.74		0	0
n_r :factors	0		0	0		10	1.48
Bias		-0.0002	-0.009	-0.016		-0.0002	-0.0002
RMSE		0.144	0.189	0.211		0.144	0.078
(a) factor DGP: $\psi_i = 1, \rho = 0.9$							
n_r :vars	10		3.67	2.99		0	0
n_r :factors	0		0	0		10	1.48
Bias		-0.0002	-0.003	-0.004		-0.0002	-0.0000
RMSE		0.320	0.431	0.491		0.320	0.171
(b) alternating sign DGP: $\psi_i = (-1)^{i+1}, \rho = 0.5$							
n_r :vars	10		1.89	0.76		0	0
n_r :factors	0		0	0		10	1.58
Bias(+)		-0.0002	-0.081	-0.113		-0.0002	-0.064
RMSE		0.144	0.156	0.151		0.144	0.145
(b) alternating sign DGP: $\psi_i = (-1)^{i+1}, \rho = 0.9$							
n_r :vars	10		2.05	0.99		0	0
n_r :factors	0		0	0		10	1.59
Bias(+)		0.0002	-0.172	-0.237		0.0002	-0.143
RMSE		0.320	0.349	0.341		0.320	0.322

(b) Bias(+) is the bias averaged over all positive coefficients, Bias(-) is essentially the same with opposite sign. Despite the factor structure not capturing the correlations with y , the RMSEs of the principal components model still yield a small improvement over using variables, with no cost in terms of RMSE relative to estimating the DGP. Few principal components are retained on average, despite requiring all of them to capture the factor structure. Figure 1b shows that the first principal component is now the least selected.

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 effects’ and is highly significant. This is the optimal case of 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 so are harder to detect.

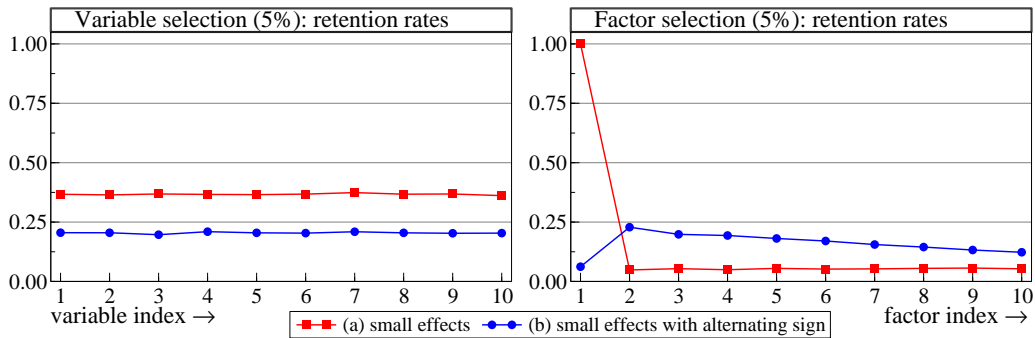


Figure 1: Retention rates of variables and factors, experiments (a),(b), $\rho = 0.9$ and $\alpha = 5\%$.

7 Monte Carlo evidence on selection with both variables and factors

Next, we consider models which combine variables and principal components in more general settings. The two procedures, joint and sequential selection, are described in Section 3. In joint selection, the GUM consists of all variables and principal components ($2n$ free regressors), whereas in the sequential procedure the first stage GUM holds all variables (n regressors) and the second stage GUM the selected variables and the principal components of the omitted variables, so n regressors again. As before, we start with a factor DGP, see (a) in Section 6. Throughout we have $T = 100$ and $\rho = 0.9$.

Data generation processes

(c) This DGP is as (a), but replacing two small coefficients with highly significant ones, as well as using alternating signs: $\beta = (b, -b, a, -a, a, -a, a, -a, a, -a)'$, with b and a calibrated to give population ψ equal to 5 and 1 respectively: $b = 1.568$ and $a = 0.315$.

(d) This is DGP (c) without the alternating sign, but extended by 10 irrelevant variables, so $n = 20$. We set $\beta = (b, b, a, a, a, a, a, a, a, a, 0, \dots)'$, again with $\psi_1 = \psi_2 = 5$ and $\psi_3 = \dots = \psi_{10} = 1$.

(e) Now $n = 50$ with the first 10 parameters calibrated to give $\psi = 5$, the next 20 with $\psi = 1$, and the final 20 having $\psi = 0$: $\beta' = (b \times \mathbf{1}_{10}', a \times \mathbf{1}_{20}', 0 \times \mathbf{1}_{20}')$, where $b = 2.18, a = 0.44$. Thus, the GUM in the joint procedure has $2n = 100$ regressors with an intercept for 100 observations, requiring *Autometrics* to undertake expanding as well as contracting searches.

Simulation results

Simulation evidence is presented in Table 3 and Figure 2. The bias is averaged over coefficients with the same non-centrality, e.g. Bias(1) is averaged over those with $\psi = 1$ in the DGP. The argument in the RMSE also indicates over which it has been averaged. The sequential procedure is listed with both significance levels, e.g. 0.1, 5% corresponds to $\alpha_1 = 0.1\%, \alpha_2 = 5\%$.

(a) The DGP has a factor structure, and the joint procedure is able to retain the factors and exclude individual regressors almost as well as using just factors (Table 2), reflected in a similarly small RMSE. Any erroneously retained regressors in the sequential procedure will be omitted from the principal components, resulting in poor proxies for the DGP factors. However, doing the first stage selection very strictly alleviates this (last two columns of Table 2). The top-left panel of Figure 2 shows that retention of the first factor drops from 99% in the joint procedure to about 90% in the sequential procedure ($\alpha = 5\%$ and $\alpha_1 = 0.1\%, \alpha_2 = 5\%$ respectively).

(c) This case has no factor structure, and many factors are needed to describe the small effects: these, however, have small non-centralities. Joint selection does worse than selecting from just variables through the increased bias of the large effects. The sequential procedure has a small advantage here from using the principal components to capture the small effects, after a strict first-stage selection. Figure 2 shows that, as in (b) before, the first factor is the one that is selected the least.

Using factors only is not shown in the table, but we note that it fares worse: 5% selection retains 4.4 factors on average, for Bias(5) = -0.16 and RMSE of 0.39.

(d) We next consider the addition of irrelevant variables. Comparing joint selection to variables only shows an improvement in the RMSE of the small effects, against a deterioration for the large and irrelevant effects. Sequential selection strikes a better balance: an improvement for small and irrelevant effects, at (almost) no cost in the RMSE of large effects.

(e) The large DGP has 50 variables for 100 observations, and estimating the DGP results in an RMSE of 0.45. So all selection methods improve in overall RMSE. Joint selection has more variables

TABLE 3:
Combining variables and factors. The DGPs are (a),(c)–(e) with two large and many small effects, as well as sparsity in (d),(e), $\rho = 0.9$, $T = 100$.

α	variables only		joint selection		sequential selection			
	5%	1%	5%	1%	5,5%	1,1%	0.1,5%	0.1,1%
(a) factor DGP with small effects: $\psi_i = 1, n = 10$								
n_r :vars	3.67	2.99	0.16	0.03	2.45	1.46	1.65	0.96
n_r :factors	0	0	1.30	1.06	0.69	0.73	1.04	0.90
Bias(1)	−0.003	−0.004	−0.000	−0.000	−0.001	−0.002	−0.001	−0.001
RMSE	0.431	0.491	0.180	0.104	0.335	0.328	0.270	0.274
(c) mixed effects: $\psi_i = (-1)^{i+1}[5 \times 1(i \leq 2) + 1(3 \leq i \leq 10)], n = 10$								
n_r :vars	3.48	2.51	1.77	1.48	3.40	2.46	2.09	2.08
n_r :factors	0	0	1.75	1.19	0.37	0.25	1.21	0.44
Bias(5)	−0.001	−0.006	−0.071	−0.136	−0.001	−0.006	−0.007	−0.014
Bias(1)	−0.183	−0.261	−0.100	−0.161	−0.137	−0.217	−0.139	−0.221
RMSE	0.322	0.333	0.361	0.391	0.338	0.331	0.321	0.324
(d) mixed effects, sparsity: $\psi_i = 5 \times 1(i \leq 2) + 1(3 \leq i \leq 10), n = 20$								
n_r :vars	5.04	4.04	2.02	2.11	4.40	3.40	3.04	2.69
n_r :factors	0	0	2.86	1.26	0.67	0.51	1.66	0.84
Bias(5)	0.136	0.258	−0.183	−0.137	0.087	0.136	0.046	0.102
Bias(1)	−0.087	−0.116	−0.048	−0.098	−0.084	−0.119	−0.098	−0.137
Bias(0)	0.041	0.039	0.075	0.105	0.049	0.067	0.069	0.088
RMSE(5)	0.375	0.451	0.498	0.664	0.366	0.408	0.383	0.431
RMSE(1)	0.396	0.425	0.321	0.315	0.371	0.371	0.329	0.332
RMSE(0)	0.224	0.202	0.316	0.260	0.241	0.198	0.231	0.186
(e) large: $\psi_i = 5 \times 1(i \leq 10) + 1(11 \leq i \leq 30), n = 50$								
n_r :vars	15.5		10.9		12.9		11.6	
n_r :factors	0		3.08		1.33		1.72	
Bias(5)	0.239		−0.050		0.066		0.021	
Bias(1)	−0.162		−0.117		−0.161		−0.168	
Bias(0)	0.041		0.143		0.128		0.156	
RMSE(5)	0.518		0.503		0.431		0.430	
RMSE(1)	0.543		0.331		0.413		0.353	
RMSE(0)	0.246		0.322		0.262		0.252	
RMSE	0.419		0.362		0.356		0.328	

than observations here, as well as perfect collinearity, but this does not have much impact: compared to (d) there is a smaller bias for the highly significant variables and a larger one for the irrelevant ones. The main benefit of the sequential procedure is that it gives a clearer partitioning into variables that matter, and factors to represent the small effects.

8 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. We do 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

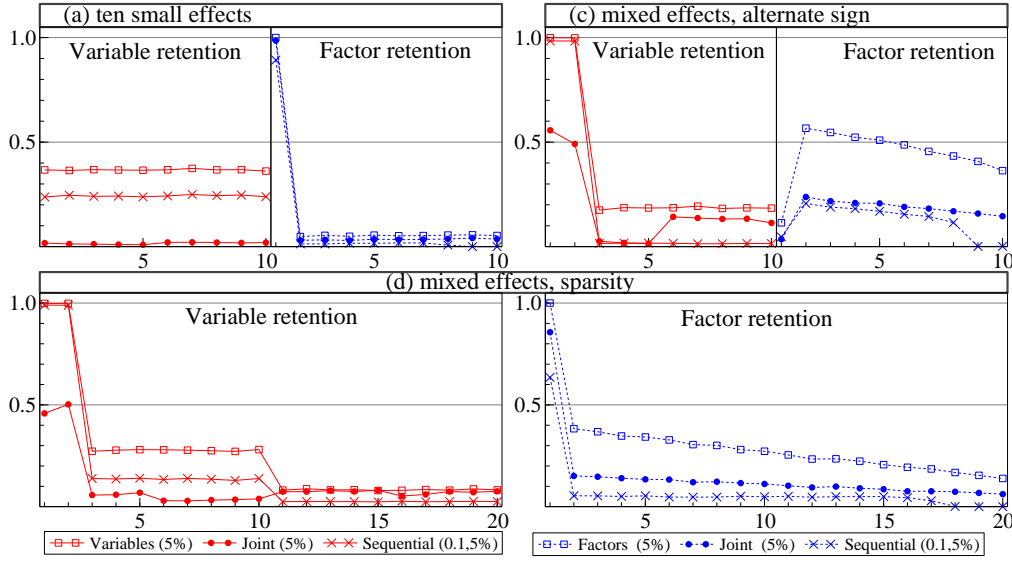


Figure 2: Retention rates of variables and factors for experiments (a),(c),(d).

than the individual regressors alone. The extent to which this is successful depends on the similarity of the correlation structure within the regressors to that between regressors and the dependent variable.

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 a factor structure, as shown for the case where many ‘small effect’ variables 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 determined initially, so outperforms the joint procedure unless the DGP has a pure factor structure.

To conclude, traditional regression models with variables and principal component models need not be treated separately, but can be combined to capture large effects through variables and small effects through factors. The sequential procedure with conservative first-stage variable selection seems to strike a good balance between factors and variables. Extensions of the present approach to non-linear models, dynamics and multiple breaks are feasible.

References

- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71, 135–171.
- Banerjee, A., M. Marcellino, and I. Masten (2008). Forecasting macroeconomic variables using diffusion indexes in short samples with structural change. In D. E. Rapach and M. E. Wohar (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.
- Castle, J. L., M. P. Clements, and D. F. Hendry (2011). Forecasting by factors, by variables, by both or neither? Working paper, Economics Department, University of Oxford.
- Castle, J. L., J. A. Doornik, and D. F. Hendry (2011a). Evaluating automatic model selection. *Journal of Time Series Econometrics* 3 (1), DOI: 10.2202/1941–1928.1097.
- Castle, J. L., J. A. Doornik, and D. F. Hendry (2011b). Model selection when there are multiple breaks. *Journal of Econometrics*, forthcoming.

- Castle, J. L. and D. F. Hendry (2010a). A low-dimension, portmanteau test for non-linearity. *Journal of Econometrics* 158, 231–245.
- Castle, J. L. and D. F. Hendry (2010b). Model selection in under-specified equations with breaks. Discussion paper 509, Economics Department, Oxford University.
- Castle, J. L. and D. F. Hendry (2011a). A Tale of 3 Cities: Model Selection in Over-, Exact, and Under-specified Equations. In M. Kaldor and P. Vizard (Eds.), *Arguing About the World*, pp. 31–55. London: Bloomsbury Academic.
- Castle, J. L. and D. F. Hendry (2011b). Automatic selection of non-linear models. In L. Wang, H. Garnier, and T. Jackman (Eds.), *System Identification, Environmental Modelling and Control*, pp. 229–250. New York: Springer.
- Castle, J. L. and N. Shephard (Eds.) (2009). *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Oxford: Oxford University Press.
- Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–1324.
- Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*. New York: Oxford University Press.
- Doornik, J. A. (2007). *Object-Oriented Matrix Programming using Ox* (6th ed.). London: Timberlake Consultants Press.
- Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics* 70, 915–925.
- Doornik, J. A. (2009a). Autometrics. See 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., M. Marcellino, and F. Neglia (2005). Principal components at work: the empirical analysis of monetary policy with large datasets. *Journal of Applied Econometrics* 20, 603–620.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized factor model: Identification and estimation. *The Review of Economics and Statistics* 82, 540–554.
- Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations* (3rd ed.). Baltimore: The Johns Hopkins University Press.
- Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In T. C. Mills and K. D. Patterson (Eds.), *Palgrave Handbook of Econometrics*, pp. 3–67. Basingstoke: Palgrave MacMillan.
- Hendry, D. F. and J. A. Doornik (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.
- Hendry, D. F. and S. Johansen (2012). Model Discovery and Trygve Haavelmo’s Legacy. *Econometric Theory* forthcoming.
- Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 33, 317–335. Erratum, 337–339.
- Hendry, D. F. and H.-M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal* 115, C32–C61.
- Hendry, D. F. and G. E. Mizon (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics* 3 (1), DOI: 10.2202/1941–1928.1100.
- Johansen, S. and B. Nielsen (2009). An analysis of the indicator saturation estimator as a robust regression estimator. See Castle and Shephard (2009), pp. 1–36.
- Leeb, H. and B. M. Pötscher (2003). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory* 19, 100–142.
- Mayo, I. and A. Espasa (2009). Forecasting aggregates and disaggregates with common features. Working paper, Universidad Carlos III, Madrid.
- Stock, J. H. and M. W. Watson (1998). Diffusion indexes. Working paper No. 6702, NBER.
- Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business and Economic Statistics* 20, 147–162.
- Stock, J. H. and M. W. Watson (2009). Forecasting in dynamic factor models subject to structural instability. See Castle and Shephard (2009), Chapter 7.