Methods explained

A quarterly series of short articles explaining statistical issues and methodologies relevant to ONS and other data. As well as defining the topic areas, the notes explain when, why and how these methodologies are used. Where relevant, we also point the reader to further sources of information.

Data reduction and model selection techniques

Graeme Chamberlin
Office for National Statistics

SUMMARY

Researchers and analysts now have access to increasingly large data sets. This article outlines some of the problems of dealing with a large number of variables and explains some of the techniques that can be used to reduce the number of available indicators to a more manageable size. This can be helpful in analysing the data or in modelling and forecasting work

Due to a proliferation of business and consumer surveys, the development of panels and better access to financial market data, large dimensional data sets have in recent years become increasingly available to statisticians and social scientists. While this undoubtedly offers better opportunities for empirical work, dealing with a large number of variables can present problems for data users. First, there are analytical issues of having to reckon with a large number of competing indicators, all of which measure the underlying variable of interest imperfectly. Second, as the number of variables approaches or exceeds the number of time observations, the problems of degrees of freedom and multicolinearity arise when using the data for modelling and forecasting purposes.

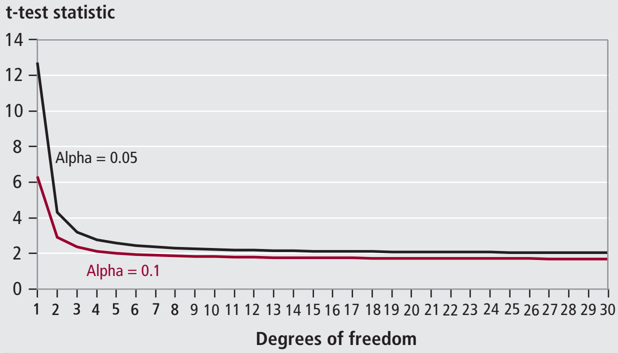
For example, suppose interest was in developing a model to explain the dependent variable y using a total of n available indicators. In principle, the following simple linear model could be estimated:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,.....+\, \beta_n x_n + u \qquad (1)$$

Degrees of freedom are the number of independent bits of information that can be used to estimate each parameter. If the time series has t observations, and there are n coefficients $\beta_1, \beta_2, ...,\beta_n$ then there are t-n degrees of freedom.

When the number of indicators exceeds the number of observations (n>t), there is insufficient information to uniquely determine the coefficients in (1) and the model cannot be estimated. Even if n<t, as n approaches t, the distributions used for hypothesis testing become so wide that it is almost impossible to judge statistical significance. This can be seen in **Figure 1** which plots the required t-test statistic to reject a null hypothesis at the 10 per cent and 5 per cent significance levels. At low degrees of freedom, this test-statistic is unlikely to reject a null hypothesis that any coefficient is significantly different from zero even if it is the case that the variable concerned is a genuine causal factor.

Figure 1
Required t-test statistic to reject a null hypothesis at different numbers of degrees of freedom



Note:
Alpha is the level of significance required

A second potential problem with estimating (1) is multicolinearity. A high degree of correlation between competing indicators makes it difficult to select the relevant variables based on t-tests alone, as standard errors become large. Some of the resulting issues are:

- small changes in the data produce wide swings in parameter estimates
- coefficients may have the ‘wrong’ sign or implausible magnitudes
- coefficients have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high

A lack of degrees of freedom and the presence of multicolinearity mean that multivariate models are usually restricted to lower dimensions. However, how should the best combination of indicator variables from a potentially very large collection be selected? For example, a set of 30 indicators can be arranged into more than 1 billion different models.

This article approaches the problem in two ways:

- **data reduction** – factor analysis is based on the notion that many variables are driven by a reduced number of common factors or shared trends. These can be extracted from the underlying data set using principal components analysis or dynamic factor analysis

- **model selection** – as a relatively large number of different models can be formed from a small number of indicators, identifying the most significant combination of variables is subject to high search costs. However, recent developments in general-to-specific modelling techniques have reduced these, improving the efficiency of model selection

To demonstrate the usefulness of these techniques, examples are based on a number of business survey indicators of the output of UK manufacturing industry, as presented in **Table 1**. Although the methods described are general, the applications discussed in this article relate to time series models.

Table 1
Business survey indicators of UK manufacturing output (1991Q1 to 2007Q1)

Organisation	Survey	Indicator
Chartered Institute of Purchasing and Supply (CIPS)	Report on Manufacturing	Output Deliveries
Confederation of British Industry (CBI)	Quarterly Industrial Trends Survey	Output Home deliveries Export deliveries
British Chambers of Commerce (BCC)	Quarterly Economic Survey	Home deliveries Export deliveries

Data reduction techniques using factor analysis

The basic insight is that strong co-movements between time series offer the opportunity to summarise the information from a large set of data by a smaller number of common factors.

For example, if the set of n indicator variables in (1) can be replaced with a set of m<n factors $f_1, f_2, ...f_m$ which account for the underlying common trends, then model (2) represents a feasible alternative:

$$y = \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + ... + \theta_m f_m + v \qquad (2)$$

There are two main approaches to extracting factors from a set of data. These are principal components and dynamic factor analysis.

Principal components

The basic methodology was developed by Hotelling (1933) and later applied by Stone (1947) to show that most of the variation in a large number of national accounts series could be interpreted by just three components: trend, cycle and rate of change of cycle.

A principal component (PC) is simply a linear combination of the variables in the data set, where each is designed in turn to account for the maximal variance of that data. So, for a set of n indicators, there will be n corresponding PCs, where the first PC is constructed to account for maximal variance, the second to account for maximal variance of that not accounted for by the first PC, and so on. If the underlying data are driven by a small number of factors, then most of the variance in that data will be accounted for by a relatively small number of PCs. Furthermore, PCs are designed to be orthogonal to each other, so the problem of multicollinearity that might otherwise beset estimation of (1) is reduced.

The methodology is based on the eigenvalues and eigenvectors for the variance-covariance matrix of the set of indicators. Eigenvalues and eigenvectors essentially describe the transformation properties of a matrix, where the eigenvector describes the direction of the transformation and the corresponding eigenvalue the strength. Hence, the first PC reflects a combination of indicators based on the eigenvector associated with the largest eigenvalue of the variance-covariance matrix. The second PC is based on the eigenvector associated with the second largest eigenvalue, and so on. If the data exhibit strong co-movements between indicators (that is, sets of indicators are strongly correlated with each other), then it will be the case that the transformation properties of the matrix are dominated by relatively few eigenvectors. This will be apparent if the first few eigenvalues are relatively large.

Table 2 shows the PC analysis of the set of seven indicators listed in Table 1. Here, the first PC accounts for over 65 per cent of the total variance in the set of indicators, whereas the first two PCs together account for almost 80 per cent of the total.

Table 2
Principal component analysis of the set of seven manufacturing indicators from Table 1

Principal component	Eigenvalue	Variance proportion explained	Cumulative variance proportion
1	4.562	0.652	0.652
2	0.982	0.140	0.792
3	0.782	0.112	0.904
4	0.444	0.063	0.967
5	0.147	0.021	0.988
6	0.054	0.008	0.996
7	0.029	0.004	1.000

In selecting the number of relevant PCs, a conventional rule of thumb is to look for a step change in the eigenvalues, which in this case occurs between the first and second PCs. Alternatively, when the data have been standardised as in this case, another rule of thumb is to select the PCs corresponding to eigenvalues greater than one. This suggests that the first PC on its own is an adequate representation of the set of seven indicators.

The composition of the first PC can be observed in **Table 3** by looking at the eigenvector associated with the largest eigenvalue. If all the data are driven by a common factor, it is normally the case that the factor loadings in the first PC are fairly equal. The evidence here suggests that the CIPS data, particularly that relating to deliveries, is less correlated with the rest of the sample. Because it has more independent variation from the rest, the second principal

Table 3
The normalised eigenvectors associated with the two largest eigenvalues, forming the basis for the first two PCs

Variable	Eigenvector 1	Eigenvector 2
CIPS output	0.1198	0.3657
CIPS deliveries	0.0551	1.3570
CBI output	0.1714	–0.1507
CBI home deliveries	0.1669	0.0029
CBI export deliveries	0.1588	–0.1788
BCC home deliveries	0.1670	–0.1759
BCC export deliveries	0.1610	–0.2203

component would be expected to be quite correlated with this variable, which is confirmed by looking at the factor loadings of the second eigenvector in Table 3.

In **Figure 2, Figure 3** and **Figure 4,** the first two principal components are plotted against the CIPS, CBI and BCC survey data, respectively. Here, it can be seen that the first principal component is strongly correlated with the CBI, BCC and CIPS output data, whereas the second principal component mirrors the movements in the CIPS deliveries data. It can be concluded that the original set of seven indicators can be summarised by one or two principal components.

The power of the PC approach is greatest when the indicator set is very large. A recent article by this author (Chamberlin 2007) showed that a set of over 400 business survey and financial markets indicators could adequately be described by eight PCs. The approach is also very good at isolating sources of idiosyncratic movements and potential outliers, as these are often identified as individual PCs and can therefore be discarded.

Figure 2
Principal components and the CIPS survey
Standardised units

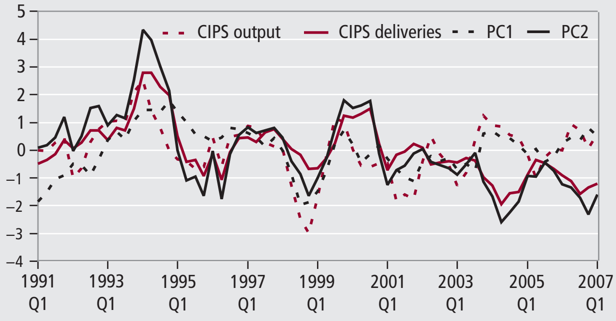


Figure 3
Principal components and the CBI survey
Standardised units

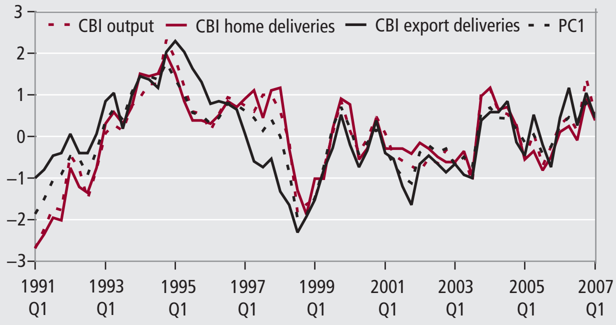
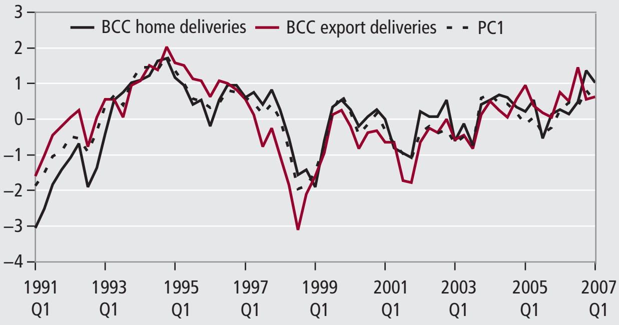


Figure 4
Principal components and the BCC survey
Standardised units



The main problem is that a PC which explains a very small proportion of the variation in the set of indicators might explain a large part of the variation of the dependent variable *y* in the model of interest. For example, if interest were in constructing a set of variables to model and forecast the official Index of Manufacturing published by the Office for National Statistics (ONS), then it cannot be discounted that the CIPS data, or specifically the CIPS delivery data, might outperform the first PC. The variable-specific parts to the CIPS data set that reduces its correlation with other surveys might just be an important ingredient in explaining movements in *y*.

Recent work by Forni *et al* (2003) has extended this basic approach. Traditional factor analysis looks to partition variables into common and variable-specific parts, but it is assumed that there is no cross-correlation at any lead or lag between the variable-specific components. This could be a problem. Suppose two industries are represented by an input-output relationship, possibly with a lag so that an idiosyncratic shock to B may eventually propagate to A. Their generalised technique, often referred to as dynamic principal components, allows a limited degree of cross-correlation between the idiosyncratic components, allowing more information to be extracted from large panels.

Dynamic factor models

This is essentially a generalisation of the PC approach and is designed to take account of the dynamic interrelationships between variables. Stock and Watson (1989) pioneered the method which has subsequently been widely applied and updated: for example, see Garratt and Hall (1996) for a UK application. The aim is to extract from a set of variables a latent variable which can be interpreted as the underlying common trend in the data. Therefore, each standardised data series can be expressed as a combination of this common variable, known as the state (S_t) and a variable-specific component $e_{i,t}$:

$$\text{CIPS output} = S_t + e_{1,t} \qquad [\text{Var}(e_1) = C_1] \qquad (3)$$
$$\text{CIPS deliveries} = S_t + e_{2,t} \qquad [\text{Var}(e_2) = C_2] \qquad (4)$$
$$\text{CBI output} = S_t + e_{3,t} \qquad [\text{Var}(e_3) = C_3] \qquad (5)$$
$$\text{CBI home deliveries} = S_t + e_{4,t} \qquad [\text{Var}(e_4) = C_4] \qquad (6)$$
$$\text{CBI export deliveries} = S_t + e_{5,t} \qquad [\text{Var}(e_5) = C_5] \qquad (7)$$
$$\text{BCC home deliveries} = S_t + e_{6,t} \qquad [\text{Var}(e_6) = C_6] \qquad (8)$$
$$\text{BCC export deliveries} = S_t + e_{7,t} \qquad [\text{Var}(e_7) = C_7] \qquad (9)$$
$$S_t = S_{t-1} + w_t \qquad [\text{Var}(w) = 1] \qquad (10)$$

Equations (3) to (9) are measurement equations, describing the relationship between the observed manufacturing indicators and the unobserved state variable. Equation (10) describes the dynamic process that represents movements in the state variable. In this case it is a simple random walk. If the dynamic term in (10) were removed, so that $S_t = w_t$, this model would become static and produce a similar outcome to the first PC in the above analysis. For this reason, the PC methodology is often referred to as static factor analysis.

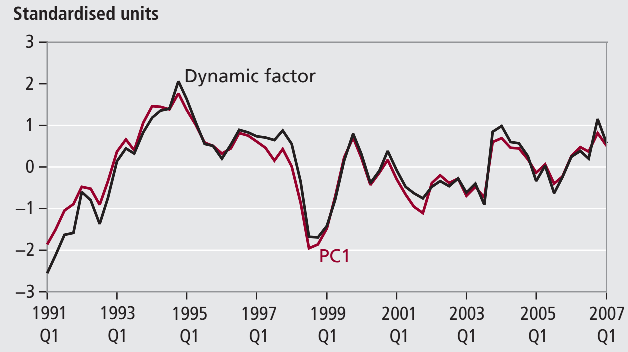
The system of equations (3) to (10) can be estimated using the Kalman filter. This is a recursive algorithm which updates its estimates of the unobserved state variable as each new data point arrives. A good description of the Kalman filter is given in Harvey (1991). Because almost any linear model can be expressed in the required state-space form, this constitutes a very flexible modelling approach. The above system is just one example. In practice, the modeller has almost free range to determine the number of unobserved variables, the dynamic structure of the state equations and the form of the measurement equations.

The key elements in this system are the noise-to-signal ratios. As the variance of the error term in the state equation is normalised to 1, these are given by the coefficients C_i, for i = 1 to 7, which determines for each indicator how much of the variable is driven by the common trend and how much by the variable-specific part. The lower the noise-to-signal ratio, the more the series is represented by the underlying common trend, and less by its own idiosyncratic features. These hyperparameters can be imposed or, as in this case, estimated using maximum likelihood methods.

Figure 5 plots the estimated state variable compared with the first PC from above. There is a fairly close association between the two which is unsurprising given the limited dynamics in the model.

Figure 5

Comparing static and dynamic factor analysis of business survey indicators of manufacturing output



Analysing the noise-to-signal ratios (**Table 4**) implies that the common trend is strongly related to the CBI data on output, home deliveries and the BCC data on home deliveries. The CIPS data, especially those on deliveries, are again given a lower weight, with more of the variance in these series explained by the indicator-specific component. As a factor extraction technique, the same criticisms made of the PC approach apply here.

Table 4

Noise-to-signal ratios for the system (3) to (10), estimated by maximum likelihood

Variable	Coefficient	Standard error	Z-statistic	Probability
C(1)	0.916	0.113	8.106	0.000
C(2)	1.384	0.254	5.445	0.000
C(3)	0.035	0.012	2.996	0.003
C(4)	0.070	0.021	3.254	0.001
C(5)	0.469	0.105	4.475	0.000
C(6)	0.122	0.027	4.506	0.000
C(7)	0.484	0.130	3.728	0.000

Factor reduction techniques offer a convenient way of summarising the main features of a data set. This can be useful when a lack of degrees of freedom or multicolinearity make estimation of a model such as (1) infeasible, but in doing so potentially useful independent sources of information are often discarded, so the alternative model (2) may not be best-fitting. The next section on model selection suggests how this problem might be addressed.

Model selection

When dealing with a large number of indicators, a common approach is to attempt to select a subset that best explains the variable of interest. The general-to-specific (GETS) modelling approach consists of starting from a very general statistical model, which captures the essential characteristics of the underlying data set, and then using standard testing procedures to reduce its complexity by eliminating statistically insignificant variables. At each stage of deletion, the validity of the reductions made should be checked to ensure the selected model continues to pass diagnostic tests (that is, it is congruent).

The main criticism of GETS is that it suffers from high 'search costs' and path dependence, meaning that it is very difficult to retrieve the best model from among all the possible combinations of variables. A study by Lovell (1983) of trying to select a small relation (0 to 5 regressors) hidden in a large database (40 variables) found a low success rate.

High search costs can easily be understood from the theory of repeated testing. Conducting 40 independent tests at the 5 per cent significance level means that there is only a $(1–0.05)^{40}$ = 0.13 chance that no tests reject by chance. A type one error is the probability of rejecting a hypothesis that is true; in this case there is a 1–0.13 = 0.87 chance that in 40 tests one or more irrelevant variables will be maintained in the model. This is quite large and shows how repeated testing can generate spurious results. Failing to reject irrelevant variables means that they may stay in the regression and act as proxies for variables that do matter, and which are subsequently omitted. Therefore, under repeated testing, the probability of retaining variables that should not enter a relationship would be high because a multitude of tests on irrelevant variables must deliver some significant outcomes by chance.

A possible solution is to raise the size of the test by using larger critical values. For example, at a 0.5 per cent significance level, there is a $(1–0.005)^{40}$ = 0.89 chance that no tests reject simply by chance. Raising the size of the test lowers the probability of type 1 errors from 0.87 to 0.11. Unfortunately, more stringent criteria for avoiding rejections when the null is true lower the power of rejection when it is false. That is, in attempting to lower the probability of maintaining irrelevant variables by raising critical values, the chance probability of rejecting the relevant ones is increased (a type 2 error). The size versus power trade-off is a well-known phenomenon in econometric modelling.

Path dependence refers to the fact that the order in which the variables are deleted generally matters, so the final model is dependent on the path taken to get there. Hence, a multitude of terminal models can result from the same starting point, making it difficult to identify the best underlying model.

Recent advancements in automating GETS procedures have reduced the search costs associated with exploring multiple deletion paths from a general model and choosing between alternative terminal models. Krolzig and Hendry (2001) have significantly aided this with the development of their PcGets software, which has built upon the earlier innovations by Hoover and Perez (1999).

Starting with a general unrestricted model (GUM), the PcGets algorithm works like a series of sieves, searching multiple deletion paths, checking that congruence is maintained at each stage and then using encompassing tests to select between terminal models. The rationale for these steps is as follows:

- **search many reduction paths** – this is designed to mitigate the problem of path dependence, leading to misspecification, as important variables are deleted and irrelevant variables are retained as proxies. Exploring several paths gives the opportunity for error correction in the light of wrong decisions. Searching all feasible paths increases the probability that some models will retain the variables that matter while eliminating those that do not

- **maintaining congruence** – the algorithm only undertakes reductions which leaves diagnostic tests as insignificant. Diagnostics act as a constraint on reduction and the choice of diagnostics and their significance levels adds to the size of the selection process

- **selection of the terminal model by encompassing** – each search path is terminated when there are no further possible reductions or when deletion induces a diagnostic test failure. Encompassing is the notion of being able to account for the results obtained by rival models given one's own findings. Therefore, if model A encompasses model B, then model A accounts for all of the variance in the dependent variable explained by model B. In this sense, encompassing implies variance-dominance, that is, a badly-fitting model cannot account for the variance of a well-fitting model

In the encompassing stage of the PcGets algorithm, all distinct non-nested models are collected and encompassing is used to eliminate those which are dominated. If a unique choice does not result, it implies that the remaining models are incomplete, that is, each explains some variance in the dependent variable not accounted for by other models, but no model is dominant. The PcGets algorithm then forms the union of resulting models which becomes the new starting point for path searches. The algorithm repeats until the union is unchanged between successive rounds.

Simply choosing the best-fitting model offers no protection against picking a spurious relationship. When a given path eliminates a variable that matters, other variables proxy such an effect, leading to spuriously large and misspecified models. However, some other paths will retain that variable and in the encompassing tests the proxies will be frequently revealed as conditionally redundant, inducing a smaller final model focused on the genuine causal factors.

Although PcGets is an automatic procedure, there is still a role to be played by the practitioner. This predominately involves choosing the form of the GUM and the significance levels of the variable deletion and diagnostic tests, which act as constraints on the paths the algorithm explores and therefore have an important bearing on the terminal models produced. If required, the practitioner can also initiate forced searches that maintain certain variables of interest in the model.

In **Table 5** and **Table 6,** the PcGets software is used to find a relationship between the set of business survey indicators and ONS's Index of Manufacturing. Estimation of the GUM is shown in Table 5, and the final dominant model in Table 6. In **Table 7,** a number of different measures of 'goodness of fit' are presented for each model.

Table 5
GUM: dependent variable – ONS Index of Manufacturing, three-month on three-month growth rate (1991Q1 to 2007Q1)

Variable	Coefficient	Standard error	t-value	t-probability
Constant	–5.264	2.164	–2.433	0.018^{**}
CIPS output	0.072	0.036	2.000	0.051^{*}
CIPS deliveries	0.032	0.023	1.401	0.167
CBI output	0.016	0.031	0.506	0.615
CBI home deliveries	0.010	0.028	0.359	0.721
CBI export deliveries	–0.009	0.017	–0.517	0.607
BCC home deliveries	–0.005	0.020	–0.222	0.825
BCC export deliveries	0.005	0.018	0.287	0.776
Seasonal Q1	–0.025	0.295	–0.084	0.933
Seasonal Q2	–0.146	0.280	–0.521	0.604
Seasonal Q3	0.241	0.284	0.849	0.400

Table 6
Final model estimated by PcGets from the GUM in Table 5

Variable	Coefficient	Standard error	t-value	t-probability
Constant	–4.231	1.554	–2.722	0.008^{**}
CIPS output	0.083	0.029	2.856	0.006^{**}
CBI output	0.019	0.007	2.737	0.008^{**}

Table 7
Information criteria for goodness of fit and parsimonious specification of the model

	GUM (Table 5)	Final model (Table 6)
Residual sum of squares	32.35	35.58
R^2	0.37	0.31
Adjusted R^2	0.26	0.28
Akaike Information Criterion	–0.36	–0.51
Schwartz Criterion	0.01	–0.41

In terms of the residual sum of squares and R^2 statistics, the original GUM is a better-fitting model. However, these statistics can never deteriorate when more variables are added to the model, so a judgement based on these criteria could lead to over-fitting. This might lead one away from the best forecasting model because adding variables can increase the variance of the forecast error.

Alternative measures such as adjusted R^2, the Akaike Information Criterion and the Schwartz Criterion are measures of goodness of fit that increasingly penalise the loss of degrees of freedom that results from adding more variables to the model. These statistics suggest that the reduced-form final model can be accepted as a more parsimonious representation of the GUM.

Although PcGets is a powerful tool aiding model selection, there are some obvious limitations in its use. Because it starts with estimating a GUM, then all the problems identified in estimating (1) still apply. The GUM will be indeterminate if there are insufficient degrees of freedom, and the presence of multicolinearity can reduce the efficiency of the algorithm leading to a proliferation of final models.

Recent work by Castle and Hendry (2006) has started to explore how PcGets might deal with these problems. They find that the procedure is still quite successful if the set of indicators is divided into smaller subgroups, where in each the variables are selected to reduce the incidence of multicolinearity. PcGets is then run on these models and a union of the final models formed as a new GUM.

CONTACT

✉ elmr@ons.gsi.gov.uk

REFERENCES

Castle J L and Hendry D F (2006): *Extending the boundaries of PcGets: non-linear models*. Oxford University Department of Economics mimeo.

Chamberlin G (2007) 'Forecasting GDP using external data sources', *Economic & Labour Market Review* 1(8), pp 18–23.

Garratt A and Hall S G (1996): 'Measuring underlying economic activity', *Journal of Applied Econometrics*, 11, pp 135–51.

Harvey A C (1991): *Forecasts, structural time series and the Kalman filter*. Cambridge University Press.

Hendry D F (2000): *Econometrics: Alchemy or Science?* Oxford University Press.

Hendry D F and Krolzig H-M (2001): *Automatic econometric model selection*. London, Timberlake Consultants' Press.

Hoover K D and Perez S J (1999): 'Data mining reconsidered: Encompassing and the general-to-specific approach to specification search', *Econometrics Journal*, 2, pp 1–25.

Hotelling H (1933): 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology,* 24, pp 417–41, 489–520.

Lovell M C (1983): 'Data Mining', *Review of Economics and Statistics*, 65, pp 1–12.

Stock J H and Watson M W (1989): 'New indexes of coincident and leading economic indicators', *NBER Macroeconomics Annual*, 4, pp 351–94.

Stone R (1947): 'On the interdependence of blocks of transactions', *Supplement to the Journal of the Royal Statistical Society*, 11, pp 1–32.