WORKING PAPERS AND STUDIES

2003 EDITION

ISBN 92-894-5336-2
ISSN 1725-4825
Cat. No. KS-AN-03-008-EN-N

A New Dynamic Factor Estimation Method for Large Datasets

EUROPEAN COMMISSION

Europe Direct is a service to help you find answers to your questions about the European Union

New freephone number:

00 800 6 7 8 9 10 11

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2003

ISBN 92-894-5336-2
ISSN 1725-4825

TABLE OF CONTENTS

COLLOQUIUM ON MODERN TOOLS FOR BUSINESS CYCLE ANALYSIS

28 & 29 NOVEMBER 2002

Luxembourg, Bech building - Room Ampère

A New Dynamic Factor Estimation Method for Large Datasets

George Kapetanios

Queen Mary, University of London

A New Dynamic Factor Estimation Method for Large Datasets

George Kapetanios
Queen Mary, University of London

November 2002

1 Introduction

Recent work in the macroeconometric literature considers the problem of summarising efficiently a large set of variables and using this summary for a variety of purposes including forecasting. Work in this field has been carried out in a series of recent papers by Stock and Watson (1998), Forni and Reichlin (2000), Forni, Hallin, Lippi, and Reichlin (2000) and Forni, Hallin, Lippi, and Reichlin (2001). Factor analysis has been the main tool used in summarising the large datasets.

The main factor model used in the past to extract dynamic factors from economic time series has been a state space model estimated using maximum likelihood. This model was used in conjunction with the Kalman filter in a number of papers carrying out factor analysis (see, among others, Stock and Watson (1989) and Camba-Mendez, Kapetanios, Smith, and Weale (2001)). However, maximum likelihood estimation of a state space model is not practical when the dimension of the model becomes too large due to the computational cost. For the case considered by Stock and Watson (1998) where the number of time series is greater than the number of observations, maximum likelihood estimation is not practically feasible. For this reason, Stock and

Watson (1998) have suggested an approximate dynamic factor model based on principal component analysis. This model can accommodate a very large number of time series and there is no need for the number of observations to exceed the number of variables. Nevertheless, the principal component model is not, strictly speaking, a dynamic model. Stock and Watson (1998) have shown that it can estimate consistently the factor space asymptotically (but the number of time series has to tend to infinity). In small samples and for a finite number of series, the dynamic element of the principal component analysis is not easy to interpret. Forni and Reichlin (2000) suggested an alternative procedure based on dynamic principal components (see Brillinger (1981, ch. 9)). This method incorporates an explicitly dynamic element in the construction of the factors.

This paper discusses an alternative method for estimating factors derived from a factor state space model. This model has a clear dynamic interpretation. Further, the method does not require iterative estimation techniques and due to a modification introduced, can accommodate cases where the number of variables exceeds the number of observations. The computational cost and robustness of the method is comparable to that of principal component analysis because matrix algebraic methods are used. The method forms parts of a large set of algorithms used in the engineering literature for estimating state space models called subspace algorithms. Another advantage of the method is that the asymptotic distribution and therefore the standard errors of the factor estimates are available. Further, as the factor analysis is carried out within a general model, forecasting is easier to carry out than in the currently available procedures where a forecasting model needs to be specified.

The structure of the paper is as follows: Section 2 describes the elements of the suggested factor extraction method. Sections 3-5 discuss aspects of

the new methodology. Section 6 presents an application of the method to the extraction of core inflation and forecasting of UK inflation in the recent past. Section 7 concludes.

2 The method

We consider the following state space model[1].

$$\begin{aligned} x_t &= Cf_t + Du_t, \quad t = 1, \ldots, T \\ f_t &= Af_{t-1} + Bu_{t-1} \end{aligned} \tag{2}$$

x_t is an n-dimensional vector of strictly stationary zero-mean variables observed at time t. f_t is an m-dimensional vector of unobserved states (factors) at time t and u_t is a multivariate standard white noise sequence of dimension n. The aim of the analysis is to obtain estimates of the states f_t, for $t = 1, \ldots, T$.

This model is quite general. Its aim is to use the states as a summary of the information available from the past on the future evolution of the system. A large literature exists on the identification issues related with the state space representation given in (2). An extensive discussion may be found in Hannan and Deistler (1988). As we have mentioned in the introduction, maximum likelihood techniques either using the Kalman filter or otherwise may be used to estimate the parameters of the model under some identification scheme. For large datasets this is likely to be computationally intensive. Subspace algorithms avoid expensive iterative techniques and instead rely on matrix algebraic methods to provide estimates for the factors as well as the

[1]Note that the model we present is equivalent to the more common form given by

$$\begin{aligned} x_t &= Cf_t + u_t, \quad t = 1, \ldots, T \\ f_t &= Af_{t-1} + v_t \end{aligned} \tag{1}$$

as proven in Hannan and Deistler (1988, pp. 17-18).

parameters of the state space representation.

There are many subspace algorithms and vary in many respects but a unifying characteristic is their view of the state as the interface between the past and the future in the sense that the best linear prediction of the future of the observed series is a linear function of the state. A very good review of existing subspace algorithms is given by Bauer (1998) in an econometric context. Another review with an engineering perspective may be found in Van Overschee and De Moor (1996).

The starting point of most subspace algorithms is the following representation of the system which follows from the state space representation and the assumed nonsingularity of D.

$$X_t^f = \mathcal{O}\mathcal{K}X_t^p + \mathcal{E}E_t^f \tag{3}$$

where $X_t^f = (x_t', x_{t+1}', x_{t+2}', \ldots)'$, $X_t^p = (x_{t-1}', x_{t-2}', \ldots)'$, $E_t^f = (u_t', u_{t+1}', \ldots)'$, $\mathcal{O} = [C', A'C', (A^2)'C', \ldots]'$, $\mathcal{K} = [\bar{B}, (A - \bar{B}C)\bar{B}, (A - \bar{B}C)^2\bar{B}, \ldots]$, $\bar{B} = BD^{-1}$ and

$$\mathcal{E} = \begin{pmatrix} D & 0 & \ldots & 0 \\ CB & D & \ddots & \vdots \\ CAB & \ddots & \ddots & 0 \\ \vdots & & CB & D \end{pmatrix}$$

The derivation of this representation is easy to see once we note that (i) $X_t^f = \mathcal{O}f_t + \mathcal{E}E_t^f$ and (ii) $f_t = \mathcal{K}X_t^p$. The best linear predictor of the future of the series at time t is given by $\mathcal{O}\mathcal{K}X_t^p$. The state is given in this context by $\mathcal{K}X_t^p$ at time t. The task is therefore to provide an estimate for $\mathcal{K}$. Obviously, the above representation involves infinite dimensional vectors.

In practice, truncation is used to end up with finite sample approximations given by $X_{s,t}^f = (x_t', x_{t+1}', x_{t+2}', \ldots, x_{t+s-1}')'$ and $X_{q,t}^p = (x_{t-1}', x_{t-2}', \ldots, x_{t-q}')'$. Then an estimate of $\mathcal{F} = \mathcal{O}\mathcal{K}$ may be obtained by regressing $X_{s,t}^f$ on $X_{q,t}^p$.

Following that, the most popular subspace algorithms use a singular value decomposition of an appropriately weighted version of the least squares estimate of $\mathcal{F}$, denoted by $\hat{\mathcal{F}}$. In particular the algorithm we will use, due to Larimore (1983), applies a singular value decomposition to $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$, where $\hat{\Gamma}^f$, and $\hat{\Gamma}^p$ are the sample covariances of $X^f_{s,t}$ and $X^p_{q,t}$ respectively. These weights are used to determine the importance of certain directions in $\hat{\mathcal{F}}$. Then, the estimate of $\mathcal{K}$ is given by

$$\hat{\mathcal{K}} = \hat{S}_m^{1/2} \hat{V}_m \hat{\Gamma}^{p^{-1}}$$

where $\hat{U}\hat{S}\hat{V}$ represents the singular value decomposition of $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$, $\hat{S}_m$ denotes the matrix containing the first m columns of $\hat{S}$ and $\hat{V}_m$ denotes the heading $m \times m$ submatrix of $\hat{V}$. $\hat{S}$ contains the singular values of $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$ in decreasing order. Then, the factor estimates are given by $\hat{\mathcal{K}} X^p_t$. For what follows it is important to note that the choice of the weighting matrices are important but not crucial for the asymptotic properties of the estimation method. They are only required to be nonsingular. A second thing to note is that consistent estimation of the factor space requires that q tends to infinity at a certain rate as T tends to infinity as pointed out by Bauer (1998, pp. 54). Once estimates of the factors have been obtained and if estimates of the parameters (including the factor loadings) are subsequently required, it is easy to see that least squares methods may be used to obtain such estimates. These estimates have been proved to be $\sqrt{T}$-consistent and asymptotically normal in Bauer (1998, ch.4). We note that the identification scheme used above is implicit and depends on the normalisation used in the computation of the singular value decomposition. Finally, we must note that the method is also applicable in the case of unbalanced panels. In analogy to the work of Stock and Watson (1998) use of the EM algorithm, described there, can be made to provide estimates both of the factors and of the missing elements in the dataset.

3 Dealing with large datasets

Up to now we have outlined an existing method for estimating factors which requires that the number of observations be larger than the number of elements in X_t^p. Given the work of Stock and Watson (1998) this is rather restrictive. We therefore suggest a modification of the existing methodology to allow the number of series in X_t^p be larger than the number of observations. The problem arises in this method because the least squares estimate of $\mathcal{F}$ does not exists due to rank deficiency of $X^{p'}X^p$ where $X^p = (X_1^p, \ldots, X_T^p)'$. As we mentioned in the previous section we do not necessarily want an estimate of $\mathcal{F}$ but an estimate of the states $X^p\mathcal{K}'$. That could be obtained if we had an estimate of $X^p\mathcal{F}'$ and used a singular value decomposition of that. But it is well known (see e.g. Magnus and Neudecker (1988)) that although $\hat{\mathcal{F}}$ may not be estimable $X^p\mathcal{F}'$ always is using least squares methods. In particular, the least squares estimate of $X^p\mathcal{F}'$ is given by

$$\widehat{X^p\mathcal{F}'} = X^p(X^{p'}X^p)^+X^{p'}X^f$$

where $X^f = (X_1^f, \ldots, X_T^f)'$ and A^+ denotes the unique Moore-Penrose inverse of matrix A. Once this step is modified then the estimate of the factors may be straightforwardly obtained by applying a singular value decomposition to $\widehat{X^p\mathcal{F}'}$. We choose to set both weighting matrices to the identity matrix in this case.

4 Number of factors

A very important question relates to the determination of the number of factors, i.e. the dimension of the state vector. This issue has only recently received attention in the econometric literature. Stock and Watson (1998) suggest using information criteria for determining this dimension. Bai and Ng (2002) provide modified information criteria and justification for their

use in the case where the number of variables goes to infinity as well as the number of observations. We suggest a simple information theoretic method for determining the number of factors in our model. Its simplicity comes from the fact that both the number of series and factors are assumed to be finite.

The search simply involves (i) fixing a maximum number of factors f^{max} to search over, (ii) estimating the factors for each assumed number of factors $m = 1, \ldots, m^{max}$ and (iii) minimising the negative penalised loglikelhood of the regression

$$x_t = C\hat{f}_t + u_t,$$

i.e. minimising $ln|\hat{\Sigma}_u^m| + c_T(m)$ where $\hat{\Sigma}_u^m$ is the estimated covariance matrix of u_t and $c_T(m)$ is a penalty term depending on the choice of the information criterion used. The theoretical properties of the new methodology are discussed in detail in Kapetanios (2002).

We briefly discuss an alternative class of testing procedures for determining the number of factors prevalent in the state space model literature. The testing procedures are based on the well known fact that the rank of certain block matrices referred to as Hankel matrices is equal to the dimension of the state vector. The most familiar Hankel matrix is the covariance Hankel matrix. The autocovariance Hankel matrix is a block matrix made up of the autocovariances of the observed process x_t. It is given by

$$\begin{pmatrix} \Gamma_1 & \Gamma_2 & \Gamma_3 \ldots \\ \Gamma_2 & \Gamma_3 & \ldots \\ \Gamma_3 & \ldots & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

where Γ_i denotes the i-th autocovariance of x_t. Its finite truncation may be estimated by $1/TX^{f'}X^p$. Tests of rank may be used to estimate the rank of the covariance Hankel matrix from its estimate. A thorough investigation of the properties of the information criteria and the testing procedures in

determining the rank of the Hankel matrix may be found in Camba-Mendez and Kapetanios (2001b). Further issues are discussed in Camba-Mendez and Kapetanios (2001a). A related discussion of the tests of rank used may also be found in Camba-Mendez, Kapetanios, Smith, and Weale (2000).

5 Extensions

The analysis of large datasets based on a state space model and estimated using subspace methods can be extended in a number of ways. Up to now we have not entertained the possibility of idiosyncratic serially correlated errors for particular variables. This extension is straightforward in the state space model context, as these errors may simply be modelled as extra factors, that enter one or a few variables. In that sense the analysis does not change. However, one may wish to draw a more clear distinction between common factors and idiosyncratic errors. Such a distinction can be accommodated by assuming that the number of variables tends to infinity following the ideas of Stock and Watson (1998). Crucially, the computational aspects of the analysis do not change.

Another important extension can be envisaged in terms of developing structural models for large datasets in the spirit of structural VAR (SVAR) models popularised in the 90's. Considering the state space model of the form

$$\begin{aligned} x_t &= Cf_t + u_t, \quad t = 1, \ldots, T \\ f_t &= Af_{t-1} + v_t \end{aligned} \tag{4}$$

we may distinguish between the shocks u_t and v_t and attribute structural meaning to linear combinations of v_t following the SVAR literature. Many possible identification schemes are possible and research in them is carried out in Kapetanios and Marcellino (2002).

6 An Application: Extracting Core Inflation

In this section we provide an application of the dynamic factor methodology to the modelling of UK core inflation. We take as our measure of inflation the RPIX (RPI minus mortgage interest payments) inflation used by the Bank of England at the target measure for monetary policy.

Core inflation is a fuzzy concept which has been defined in various ways in the literature. We will not attempt to provide even a partial review of a huge literature. In general, when people use the term core inflation they seem to refer to the long-run or persistent component of the measured price index. A clear definition of core inflation requires a model of how prices and money are determined in the economy. We choose to follow an atheoretical approach to the definition of core inflation by specifying it to be the major dynamic factor underlying the components used to construct the retail price index.

More specifically let the set of individual price component growth rates be denoted by x_t. These growth rates are obtained by differencing the logarithm of the respective component price index. Then, x_t is specified to follow a model of the form (4). Core inflation at time t is defined to be the first factor in the vector f_t as defined by the ordered singular values of the singular value decomposition of $\mathcal{F} = \mathcal{O}\mathcal{K}$ in (3). This definition although in no way related to a theoretical economic model is consistent with the prior idea that core inflation is the main persistent component of inflation.

We fit a state space model to the components of the RPIX price index for the period of January 1987 to August 2002. Monthly data are used. Information on the components used are given in the data appendix. We set the truncation indices to $s = 1$ and $q = 3$ respectively. We note that q has to tend to infinity as the sample size grows in order to get a consistent estimate

of the factors. We have chosen to set this to 3 because the resulting estimate of core inflation does not change perceptibly as q is increased from this value. Component series were normalised to have mean equal to zero and variance equal to one prior to estimation of the factor. We present RPIX inflation and our measure of the core inflation in Figure 1. Note that the core inflation has been normalised to have the same mean and variance as observed inflation over the sample period.

Clearly, the factor model estimate of the core inflation is smoother than actual observed inflation. However, at business cycle frequencies it exhibits pronounced cyclicality. The departure from observed inflation in the spike of the late eighties and early nineties can be traced back to tax changes (including the repeal of the poll tax) in that period. Our measure of core inflation can explain on average 44% of a given component series whereas addition of an extra factor raises this to 53%.

Having obtained a means of estimating core inflation we now examine the forecasting abilities of this measure. In particular we consider three models. One is a simple benchmark AR model where the lag order is chosen automatically using the Akaike information criterion. The second is the benchmark model augmented by the growth rate of money and in particular M0. Lag selection is again carried out by the Akaike informatin criterion for both inflation and the money growth rate. Finally, the third model is the benchmark model augmented with the currently available estimate of the core inflation.

We evaluate the three models over the period June 1998-August 2002. We have allowed for a year following the introduction of independence for the Bank of England to carry out monetary policy though an inflation targeting regime. We examine both relative RMSEs compared to the model which includes the factor and the Diebold and Mariano (1995) test statistic for

equality in predictive ability between two different forecasts. All models are estimated recursively (including lag order selection). The forecasts are examined for horizons of 1 to 4 months ahead. All results are presented in Table 1.

Table 1: Results on forecasting performance

Horizon	DM[a]	DM[b]	RMSE[c]	RMSE[d]
1	1.42	0.66	0.95	0.68
2	0.13	0.26	0.99	0.98
3	0.48	0.67	0.97	0.92
4	0.61	1.04	0.97	0.89

[a]Diebold-Mariano test statistic against benchmark AR model

[b]Diebold-Mariano test statistic against money growth rate model

[c]Relative RMSE compared to benchmark AR model. Values less than 1 indicate superiority of factor model

[d]Relative RMSE compared to money growth rate model. Values less than 1 indicate superiority of factor model

The results show that the factor model can indeed help in forecasting. The factor model performs 32% better than the money growth model for forecasts one month ahead. The factor model always has a lower RMSE compared to the other models. Although the factor model may appear to have a similar performance compared to the AR model the Diebold-Mariano statistic, although not rejecting in favour of the factor model, indicates that with a probability value of 0.078 is close to rejection.

7 Conclusion

In this paper we have discussed a new factor based method for forecasting time series introduced by Kapetanios (2002). This work follows closely in spirit the work of Stock and Watson (1998), Stock and Watson (1999) and subsequent, as yet unpublished papers by these authors and their co-authors on the one hand and the work by Forni and Reichlin (2000), Forni, Hallin, Lippi, and Reichlin (2000) and Forni, Hallin, Lippi, and Reichlin (2001) on

the other hand. The innovation lies in providing an alternative method for obtaining factor estimates.

One strand of the literature on factor extraction relies on explicitly dynamic state space models to estimate factors via computationally expensive and, in small samples, non-robust maximum likelihood estimation. The other strand of the literature based on the work of Stock and Watson (1998) uses principal components to extract the factors. This methodology is robust, computationally feasible with very large datasets and asymptotically valid for dynamic settings. Unfortunately, these methods are approximately dynamic in that the dynamic structure of the factors is not explicitly modelled in finite samples but captured only asymptotically where both the number of observations and the number of series used, grows to infinity. We propose a new methodology which while sharing all the advantages of the principal component extraction method is explicitly dynamic. This method is based on linear algebraic techniques for estimating the state and, if need be, the parameters of a general linear state space model.

We evaluate the new methodology by investigating a model of core inflation for the UK. The measure of core inflation obtained is shown have predictive ability for inflation in the UK over a relatively long evaluation period.

References

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1).

Bauer, D. (1998): "Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms," Ph.d. Thesis.

Brillinger, D. (1981): *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco.

Camba-Mendez, G., and G. Kapetanios (2001a): "Statistical Testing Procedures and Information Criteria in Rank Determination for System Identification," Mimeo.

——— (2001b): "Testing the Rank of the Hankel Covariance matrix: A Statistical Approach," *Institute of Electrical and Electronic Engineers Transactions on Automatic Control*, 46(2), 331–336.

Camba-Mendez, G., G. Kapetanios, R. J. Smith, and M. R. Weale (2000): "Tests of Rank in Reduced Rank Regression Models," Forthcoming in Journal of Business and Economic Statistics.

——— (2001): "An Automatic Leading Indicator of Economic Activity: Forecasting GDP growth for European Countries," *Econometrics Journal*, 4, 56–90.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82(4), 540–554.

——— (2001): "The Generalised Dynamic Factor Model: One Sided-Estimation and Forecasting," Mimeo, Université Libre de Bruxelles.

Forni, M., and L. Reichlin (2000): "Let's Get Real: A Factor Analytic Approach to Business Cycle Dynamics," *Review of Economic Studies*, 65(3), 453–473.

Hannan, E. J., and M. Deistler (1988): *The Statistical Theory of Linear Systems*. John Wiley.

Kapetanios, G. (2002): "Factor analysis using subspace factor models: Some theoretical results and an application to UK inflation forecasting," *Working paper no. 466*.

Kapetanios, G., and M. Marcellino (2002): "Structural Impulse Response Analysis of Dynamic Factor Models for Large Datasets," *Mimeo*.

Larimore, W. E. (1983): "System Identification, Reduced Order Filters and Modelling via Canonical Variate Analysis," *Proc. 1983 Amer. Control Conference*, 2, 445–451.

Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications to Statistics and Econometrics*. John Wiley.

Stock, J. H., and M. Watson (1989): "New Indexes (Sic) of Coincident and Leading Indicators," *NBER Macroeconomics annual*, 4, 351–394.

Stock, J. H., and M. W. Watson (1998): "Diffusion Indices," Forthcoming in Journal of Business and Economic Statistics.

——— (1999): "Forecasting Inflation," *Journal of Monetary Economics*, 44(2), 293–335.

Van Overschee, P., and B. De Moor (1996): *Subspace Identification for Linear Systems*. Kluwer Academic Publishers.

Data Appendix

RPIX components and their ONS (Office of National Statistics) codes.

bread DOAA.M
cereals DOAB.M
biscuits DOAC.M
beef DOAD.M
lamb DOAE.M

pork DOAG.M

bacon DOAH.M

poultry DOAI.M

other meat DOAJ.M

fish DOAK.M

butter DOAM.M

oil and fat DOAN.M

cheese DOAO.M

eggs DOAP.M

milk DOAQ.M

milk products DOAR.M

tea DOAS.M

coffee DOAT.M

soft drink DOAU.M

sugar DOAV.M

sweets chocolates DOAW.M

potatoes DOAX.M

vegetables DOAZ.M

other foods DOBD.M

restaurant meals DOBE.M

canteen meals DOBF.M

take aways DOBG.M

beer DOBH.M

wine DOBK.M

cigarettes DOBN.M

other tobacco DOBO.M

rent DOBP.M

council tax DOBR.M

water DOBS.M

repairs and maintenance DOBT.M

DIY DOBU.M

insurance and ground rent DOBV.M

coal DOBW.M

electricity DOBX.M

gas DOBY.M

oil and other fuel DOBZ.M

furniture DOCA.M

furnishings DOCB.M

appliances DOCC.M

other eqpt DOCD.M

consumables DOCE.M

pet care DOCF.M

postage DOCG.M

telephones DOCH.M

dom services DOCI.M

fees and subs DOCJ.M

clothing men DOCK.M

clothing women DOCL.M

clothing children DOCM.M

clothing other DOCN.M

footwear DOCO.M

personal articles DOCP.M

chemist goods DOCQ.M

personal services DOCR.M

purchase cars DOCS.M

maintenance cars DOCT.M

petrol and oil DOCU.M

tax and insurance DOCV.M

rail fares DOCW.M

bus and coach fares DOCX.M

other travel DOCY.M

audio visual DOCZ.M

CDs tapes DODA.M

toys and sports goods DODB.M

books and newspapers DODC.M

garden products DODD.M

tv licences DODE.M

entertainment and other recreation DODF.M

Figure 1: Observed and Core Inflation