CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2008-44

The cyclical component factor model

Christian M. Dahl, Henrik Hansen and John Smidt

School of Economics and Management
University of Aarhus
Building 1322, DK-8000 Aarhus C
Denmark

Aarhus School of Business
University of Aarhus
Handelshøjskolen
Aarhus Universitet

UNIVERSITY OF
COPENHAGEN

The cyclical component factor model*

Christian M. Dahl†
CREATES, School of Economics and Management,
University of Aarhus,

Henrik Hansen
Institute of Food and Resource Economics,
University of Copenhagen,

John Smidt
Danish Economic Councils

August 28, 2008

Abstract

Forecasting using factor models based on large data sets have received ample attention due to the models' ability to increase forecast accuracy with respect to a range of key macroeconomic variables in the US and the UK. However, forecasts based on such factor models do not uniformly outperform the simple autoregressive model when using data from other countries. In this paper we propose to estimate the factors based on the pure cyclical components of the series entering the large data set. Monte Carlo evidence and an empirical illustration using Danish data shows that this procedure can indeed improve on pseudo real time forecast accuracy.

Key words: Factor model; Cyclical components; Estimation; Real time forecasting.

JEL Codes: C13, C22, C52, G53.

1 Introduction

The diffusion index model developed by Stock and Watson (1998, 2002a, b) has been shown to have very good forecasting properties when predicting macroeconomic variables, mainly using US, UK and Euro-wide data. See Stock and

*We wish to thank an anonymous referee and the associated editor for very useful comments and suggestions. The first author gratefully acknowledges the research support of CREATES (funded by the Danish National Research Foundation).

†Corresponding author. Room 225, Building 1326, DK-8000 Aarhus C. Phone: +45 8942 1559. E-mail: cdahl@econ.au.dk.

Watson (1998, 2002a, b), Marcellino, Stock and Watson (2003), Artis, Banerjee and Marcellino (2005) and Banerjee and Marcellino (2006) among others. However, the diffusion index model does not outperform simpler models in all cases, and several recent studies propose extensions, and improvements, of the model. Bai and Ng (forthcoming) focus on the estimation of the latent factors by looking at a polynomial extension of the factor series (that is, including squared terms of the series) and by looking at a selection procedure for the factor series. In Bai and Ng (2008) the focus is on improvements of the forecasting equation, taking the estimated factors as given. Moving in a slightly different direction, Armah and Swanson (2007) look at construction of "factor proxies" which is, essentially, a modification of the classical leading indicator model in which the leading indicators are selected based on their similiarity with the latent factors, estimated using the diffusion index methodology.

In this paper we propose a relatively simple method to potentially improve the forecast accuracy of the diffusion index model. We illustrate the usefulness of the modification in a small Monte Carlo study and by an empirical illustration based on Danish data. Our approach is inspired by the work of Camacho and Sancho (2004) and Kaiser and Maravall (1999). The basic idea is to remove not only the trend, the seasonal components and outliers but also the irregular component in all series entering the large data set which is used for estimation of the factors. In some cases this might just be a minor modification of the pre-filtering of the data, but in situations where the irregular component is relatively large, we conjecture that this modification of the pre-filtering in the diffusion index model will provide more accurate estimates of the factors.

As argued by Dahl et al. (2005) the irregular component in Danish data seems to be much more dominating than in, say, US data. This might explain why the forecast performance of diffusion index models estimated on Danish data is relatively disappointing in the sense that the forecast accuracy of the diffusion index model is not significantly better than the accuracy of standard autoregressive models. Here, we show that when the factors are based on estimates of the "pure" cyclical components, the predictive accuracy of the diffusion index model is improved substantially.

The paper is structured as follows. After explaining the basic idea in Section 2, we present results of a simple Monte Carlo study in Section 3. The Monte Carlo study shows that pre-filtering the data has the potential of improving the forecast accuracy of the traditional diffusion index model. In Section 4 we provide an empirical illustration using Danish data to forecast four key macroeconomic variables out of sample. The illustration provides additional evidence that pre-filtering the data can improve forecasting accuracy when the time series contain large irregular components.

2 The modelling framework

Consider the large collection of time series $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \ldots, \boldsymbol{X}_T')$, where $\boldsymbol{X}_t = (x_{1t}, x_{2t},...,x_{Nt})'$. Assume that each series, x_{it}, can be represented as

$$x_{it} = g_{it} + c_{it} + s_{it} + e_{it} \tag{1}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$, where g_{it} denotes a trend component, c_{it} the business cycle component, s_{it} a seasonal component and e_{it} the irregular component.

In the existing work on diffusion index models it is common to make prior adjustments of each time series, x_{it}, (i) by removing the seasonal component, s_{it}, (applying the popular X11 filter); (ii) by removing the trend component, g_{it}, (applying first (log) differences), and (iii) by screening for outliers (say, by removing observations in excess of some predetermined threshold value). Consequently, using the "traditional" Stock and Watson (1998, 2002a,b) approach, the estimator of the common factors is based on the relation

$$\hat{x}_{it} = \boldsymbol{\alpha}_i \boldsymbol{F}_t + \eta_{it}, \tag{2}$$

where $\hat{x}_{it}$ is the trend and seasonally adjusted series (assuming there are no outliers), $\boldsymbol{F}_t = (f_{1t}, f_{2t}, ..., f_{kt})'$ are the common factors and $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ik})$ are the factor loadings. In addition, it is typically assumed that η_{it} is an idiosyncratic error term.

The main contribution of this paper is to propose a modification of (2) by explicitly assuming that it is the business cycle component of the series, c_{it}, that admits a linear factor representation with k common factors, i.e.,

$$c_{it} = \boldsymbol{\alpha}_i \boldsymbol{F}_t + \upsilon_{it}, \tag{3}$$

where υ_{it} has the same properties as η_{it}.

Clearly, this assumption is admissible within the traditional diffusion index model. We can define the true trend and seasonally adjusted series $\tilde{x}_{it} \stackrel{\text{def}}{=} c_{it} + e_{it}$, and assume that the estimator $\hat{x}_{it}$ satisfies the condition

$$\hat{x}_{it} = \tilde{x}_{it} + \hat{\epsilon}_{it,} \tag{4}$$

where $\hat{\epsilon}_{it}$ is the estimation error associated with the trend and seasonal adjustment procedure. Given (1), (3) and (4) we have the factor model

$$\hat{x}_{it} = \boldsymbol{\alpha}_i \boldsymbol{F}_t + \upsilon_{it} + e_{it} + \hat{\epsilon}_{it}, \tag{5}$$

which is observationally equivalent to (2) with

$$\eta_{it} = \upsilon_{it} + e_{it} + \hat{\epsilon}_{it}.$$

Given the representation in (1) and the factor model (3), and assuming that we could actually observe c_{it}, it would obviously be more informative to estimate $\boldsymbol{F}_t$ based on (3) instead of (2). In reality, however, we do not observe c_{it}. Still, an alternative to model (2) is to use estimates of the cyclical components based on the individual series. Specifically, if we let $\hat{c}_{it}$ be an estimator of c_{it} and let $\hat{\epsilon}^c_{it}$ denote the associated estimation error, equation (3) can be represented as[1]

$$\hat{c}_{it} = \boldsymbol{\alpha}_i \boldsymbol{F}_t + \upsilon_{it} - \hat{\epsilon}^c_{it}, \tag{6}$$

Now, the estimator of $\boldsymbol{F}_t$ based on (6) is not guaranteed to be more informative relative to the estimator based on (2). In (3) the error term is $\upsilon_{it} + e_{it} + \hat{\epsilon}_{it}$ while model (6) has the error term $\upsilon_{it} - \hat{\epsilon}^c_{it}$. Thus, assuming orthogonality of the error components the relative efficiency of the estimators depends on the variance of $e_{it} + \hat{\epsilon}_{it}$ and the variance of $\hat{\epsilon}^c_{it}$. The relative size of these variances cannot be determined analytically as it depends on the variances of the idiosyncratic components, the estimator of the cyclical components–and, hence, the time series dimension, T–and the number of series in the factor model, N. Thus, it is primarily an empirical question as to which approach is most informative/efficient. However, if data is very noisy due to large variance in the irregular component of the series, this will tend to favour the estimation approach based on (6) for given dimensions of the data matrix, $\boldsymbol{X}$.

Our main interest is out-of-sample forecasting of, say, y_t which typically is an element of $\widehat{\boldsymbol{X}_t} = (\hat{x}_{1t}, \hat{x}_{2t}, \ldots, \hat{x}_{Nt})'$. Following the approach by Stock and Watson (2002a,b), the approximating cyclical diffusion index h-periods ahead forecasting model can be represented as

$$(y_{t+h} - y_t) = \sum_{j=1}^{k} \beta^c_j \hat{f}^c_{jt} + \sum_{j=1}^{p} \gamma^c_j \Delta y_{t-j} + \upsilon^c_t, \tag{7}$$

for $t = 1, 2, \ldots, T$, where the estimated factors $\widehat{\boldsymbol{F}}^c_t = (\hat{f}^c_{1t}, ..., \hat{f}^c_{kt})'$ are based on (6) using principal components. We wish to compare (7) to the "regular" diffusion index forecasting model

$$(y_{t+h} - y_t) = \sum_{j=1}^{k} \beta^r_j \hat{f}^r_{jt} + \sum_{j=1}^{p} \gamma^r_j \Delta y_{t-j} + \upsilon^r_t, \tag{8}$$

where the estimator $\widehat{\boldsymbol{F}}^r_t = (\hat{f}^r_{1t}, \ldots, \hat{f}^r_{kt})'$ is obtained based on (2) and to the pure autoregressive linear model

$$(y_{t+h} - y_t) = \sum_{j=1}^{p} \gamma^l_j \Delta y_{t-j} + \upsilon^l_t. \tag{9}$$

[1] An estimate of c_{it} can be obtained given some additional assumptions about the underlying stochastic processes driving each of the unobserved components in (1) as shown by Harvey (1989), and more recently discussed by Durbin and Koopman (2001).

3 A Monte Carlo simulation study

In this section we provide a simple Monte Carlo simulation study illustrating the potential efficiency of the cyclical component factor model relative to the regular factor model. The comparison will be based on relative MSE measured in-sample. It should be emphasized that the only purpose of the Monte Carlo study is to illustrate that there *can* exist situations in which the cyclical component factor model has smaller MSE than the regular factor model. Whether this is actually the case based on real data and out-of-sample is an entirely different and mainly empirically question which we will address in the subsequent section.

3.1 A simple sampling scheme for generating observables

A convenient and simple method of generating so-called similar/common cycles has been suggested by Harvey and Koopman (1997) and Carvalho, Harvey and Trimbur (2007). Following their approach the cyclical component $\boldsymbol{C}_t = (c_{1t}, c_{2t}, ...c_{Nt})'$ takes the representation

$$\begin{pmatrix} \boldsymbol{C}_t \\ \boldsymbol{C}_t^* \end{pmatrix} = \left[\rho \begin{pmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{pmatrix} \otimes \boldsymbol{I}_N \right] \begin{pmatrix} \boldsymbol{C}_{t-1} \\ \boldsymbol{C}_{t-1}^* \end{pmatrix} + \begin{pmatrix} \boldsymbol{\kappa}_t \\ \boldsymbol{\kappa}_t^* \end{pmatrix} \tag{10}$$

for $t = 1, ..., T$, where $\boldsymbol{\kappa}_t$ and $\boldsymbol{\kappa}_t^*$ are Gaussian disturbances such that $\mathrm{E}(\boldsymbol{\kappa}_t \boldsymbol{\kappa}_t') = \mathrm{E}(\boldsymbol{\kappa}_t^* \boldsymbol{\kappa}_t^{*\prime}) = \boldsymbol{\Sigma}_{\boldsymbol{\kappa}}$ and $\mathrm{E}(\boldsymbol{\kappa}_t \boldsymbol{\kappa}_t^{*\prime}) = \boldsymbol{0}$. In the representation $\rho \in [0; 1)$ is denoted the dampening factor, while the cycle parameter λ_c satisfies $0 \leq \lambda_c \leq \pi$.

Note, that if $\boldsymbol{\Sigma}_{\boldsymbol{\kappa}}$ has reduced rank then there exists common cycles. Consequently, if the rank of $\boldsymbol{\Sigma}_{\boldsymbol{\kappa}}$ is two then there exists two common cycles according to the representation.

Based on the generated matrix $\boldsymbol{C} = (\boldsymbol{C}_1', ..., \boldsymbol{C}_T')'$ we compute the true factors $\boldsymbol{F}_t = (f_{1t}, ..., f_{kt})$ by standard principal components routines. We have hereby explicitly assumed that $\boldsymbol{C}$ has a factor representation. Finally, the observables $(y_t, \boldsymbol{X}_t)$ are generated recursively as

$$\begin{aligned} y_t &= \gamma y_{t-1} + \sum_{j=1}^{k} \beta_j f_{jt} + v_t, \\ \boldsymbol{X}_t &= \boldsymbol{C}_t + \boldsymbol{e}_t, \end{aligned}$$

for all $t = 1, 2, \ldots, T$ where $v_t \sim \mathrm{N}(0, \sigma_v^2)$ and $\boldsymbol{e}_t \sim \mathrm{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N)$.

In order to make the data generating process empirically relevant to some degree we have chosen population parameter configurations similar to estimated magnitudes reported in Table 8 in Carvalho, Harvey and Trimbur (2007) based on US/Canadian data. That is, in terms of generating $\boldsymbol{C}$ according to (10) we have chosen $\rho = 0.9$, $\lambda_c = 0.2$ and $k = 2$. Furthermore, we have used that we can write $\boldsymbol{\kappa} = \boldsymbol{\Lambda}\boldsymbol{\omega}_1$ and $\boldsymbol{\kappa}^* = \boldsymbol{\Lambda}\boldsymbol{\omega}_2$, where $\boldsymbol{\Lambda}$ is an $N \times k$ matrix and $\boldsymbol{\omega}_s \sim \mathrm{N}(\boldsymbol{0}_k, \boldsymbol{I}_k)$ for

$s = 1, 2$. Hence $\boldsymbol{\Sigma}_{\boldsymbol{\kappa}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ has reduced rank, equal to k (the number of common cyclical factors), as desired. In the simulations $\boldsymbol{\Lambda}$ is drawn from the (independent) uniform distribution on the unit interval. In order to generate $\boldsymbol{X}$ we have chosen to consider a relative dense sequence of values for $\sigma_e^2 = (0.25, 0.5, \ldots, 4.75, 5)$ since this parameter will be pivotal for the relative efficiency of the cyclical component factor model approach as we have argued above. Finally, in order to generate y_t we have chosen $\gamma = 0.5$, $\beta_1 = \beta_2 = 1$ and $\sigma_v^2 = 1$.

3.2 The estimation procedure and the results

Given the matrix of observables, $(y_t, \boldsymbol{X}_t)$ the estimation procedure for the cyclical components factor model and the regular factor model can be summarized as follows:

Cyclical components factor model (c)

1. For each of the N time series, x_{it}, the cyclical component is estimated using using the linear Gaussian State Space representation and the Kalman filter.[2]

2. Based on the estimated cyclical components, $\hat{c}_{it}$, the factors, $\hat{f}_{jt}^c$, for $j = 1, 2$ and $t = 1, 2, ..., T$ are estimated using a principal components decomposition.

3. The variable y_t, is regressed on y_{t-1} and $\hat{f}_{jt}^c$, for $j = 1, 2$ and the mean squared error, $MSE(c) = \frac{1}{T}\sum_t (\widehat{v}_t^c)^2$, is computed.

Regular factor model (r)

1. Based on the time series x_{it}, the factors, $\hat{f}_{jt}^r$, for $j = 1, 2$ and $t = 1, 2, ..., T$ are estimated using a principal components decomposition.

2. The variable y_t, is regressed on y_{t-1} and $\hat{f}_{jt}^r$ for $j = 1, 2$ and the mean squared error, $MSE(r) = \frac{1}{T}\sum_t (\widehat{v}_t^r)^2$, is computed.

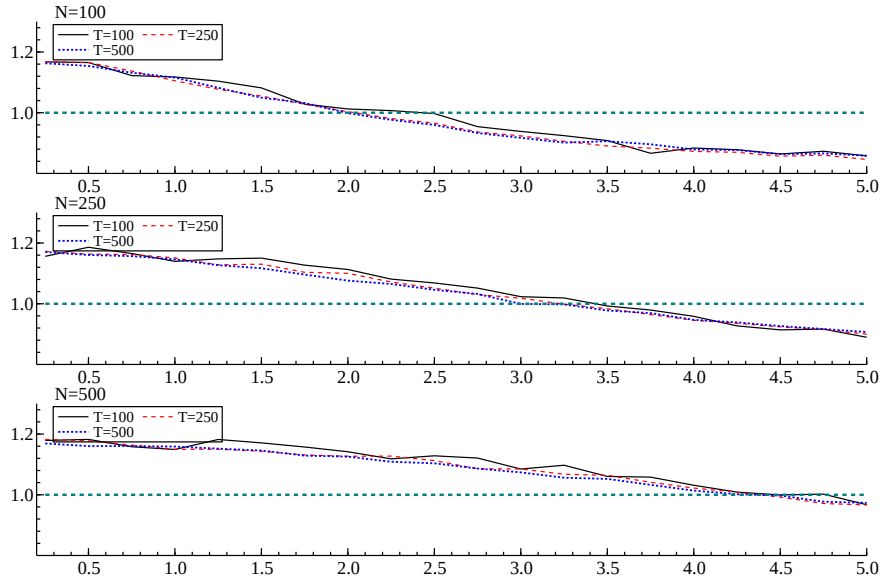
In Figure 1 we report the results on the relative efficiency of the cyclical component factor model defined as $MSE(c)/MSE(r)$ for alternative values of σ_e^2 and for different values of N and T.[3]

All three panels in Figure 1 clearly indicate that a larger variance in the irregular component increases the efficiency of the cyclical component factor model relative to the regular factor model. For example, in the case where $N = 100$ and $T = 500$ (top panel) the cyclical component factor model becomes more efficient whenever $\sigma_e^2 > 2$. Furthermore, and perhaps somewhat surprisingly, the relative efficiency is almost unaffected by changes in the sample size. This implies that

[2] We used the SSF-package for Ox by Koopman et al. (1999)

[3] The results are based on 1000 Monte Carlo replications for each value of σ_e^2, N, T.

Figure 1: Efficiency of the cyclical component factor model relative to the regular factor model



In each plot the x-axis shows the variance of the irregular component σ_e^2 while the y-axis shows the MSE of the cyclical component factor model relative to the MSE of the regular factormodel: $MSE(c)/MSE(r)$.

the algorithms employed (the SSF-package) for estimating the cyclical components appear quite efficient even in small to moderate samples. Finally, as the number of time series increases from $N = 100$ over $N = 250$ to $N = 500$ the relative efficiency of the cyclical component factor model falls uniformly over T and σ_e^2 and in the case where $N = 500$ (bottom panel) the cyclical component factor model only becomes efficient when $\sigma_e^2 > 4.5$.

Summing up, the limited simulation evidence provided in this section clearly illustrates that when the variance of the irregular component is relatively high and the dimension of $\boldsymbol{X}$ moderate (when N is of small to moderate size) the cyclical component factor model may be a potent alternative to the regular factor model approach.

4 Empirical Illustration

The empirical illustration is based on Danish data, which in general is characterized by being much more volatile relative to US data. For example, as pointed out by Dahl et al. (2005), the volatility in the Danish GDP growth rate is about twice as high as the volatility in US GDP growth, whereas the volatility in the industrial production in Denmark is about seven times higher than the volatility

in US industrial production. Dahl et al. (2005) argues, that this could be due to the presence of more noise in the Danish data and this may explain why the regular diffusion index model based on Danish data does not perform well in terms of forecast accuracy as shown by Dahl et al. (2005). This provides a strong motivation for improving the factor model by computing the factors based on preliminary estimates of the cycle component, which should be a less noisy signal of the underlying business cycle component. In this study our main interest is on forecasting private consumption, GDP, employment and the deflator for private consumption (inflation), which are all important policy variables and are all measured on a quarterly basis.

4.1 The data and the estimated factors

The data set for Denmark, our $\boldsymbol{X}$, contains 172 monthly and 74 quarterly series over the period 1986m1 - 2003m12. To obtain a good representation of the Danish economy we include a wide range of output variables, labour market variables, prices, monetary aggregates, interest rates, stock prices, exchange rates, imports, exports, net trade, and other miscellaneous series. This selection procedure closely follows the suggestions in Stock and Watson (2002a,b) and is aimed at getting as balanced and complete a list of important variables as possible. A description of the entire list of the variables is reported in Dahl et al. (2005).[4]

When combining monthly as well as quarterly data in $\boldsymbol{X}$ and $\boldsymbol{C}$ they become unbalanced data matrices. We therefore employ the EM algorithm described in Stock and Watson (1998,2002a,b) to fill out the missing observations. The number of factors in the factor model is determined by the information criteria (IC_{p1}), suggested by Bai and Ng (2002).

4.2 The forecasting framework

We wish to compare the out-of-sample forecast accuracy of the cyclical diffusion index model, (7), relative to the forecasts of the traditional diffusion index model, (8), and a pure autoregressive model, (9). Specifically we want to forecast the growth rates of private consumption, GDP, employment and inflation one and four quarters ahead, respectively. The first period in the pseudo out-of-sample is 1995q1 and the last period is 2003q4. The forecasting framework is best illustrated by an example. Consider the one period ahead forecast of the growth rate of say GDP: First, we estimate the factors using data from 1986m1 to 1994m12. These factors are estimated at a monthly frequency and subsequently collapsed into quarters. Then, we estimate the forecasting equations by ordinary least squares, applying an automated general-to-specific procedure, using data up to and including 1994q4. Given that the right-hand-side variables, including the

[4]The data and documentation can be obtained from the corresponding author.

factors, are lagged, only observations up to 1994q3 are used in the estimation. Finally, the forecasts (and the forecast errors) are calculated using the specific version of the estimated equations. In the one period ahead forecast for the growth rate in 1995q1 observations of the factors in 1994q4 are used (along with lagged values of the variable to forecast).

It turns out that the results are very sensitive to the initial choice of the maximum number of factors, k, and the maximum number of autoregressive lags, p, used in the general specifications of the forecasting equations. As pointed out by Dahl et al. (2005) it is possible to choose combinations of k and p which make the forecasts based on (7) outperform the other models, while other combinations do not. Obviously, this indicates that there is a risk of data snooping as described by White (2000) and one should be careful interpreting such findings as an indication in favor of the diffusion index model.[5]

We try to avoid the data snooping pitfall by reporting the predictive outcome of the models selected by an automated general-to-specific selection mechanism starting from a range of different general models. In particular, we make forecasts for all possible combinations of initial settings of $k = 1, 2$ and $p = 1, \ldots, 8$. This implies that for each variable and forecast horizon 16 measures of forecast accuracy (we use the mean squared forecast error, MSFE) are computed over the out-of-sample period for the two diffusion index models. For the autoregressive model we compute 8 MSFEs for each variable and forecast horizon. The automated model selection procedure we employ for each k and p is the traditional general-to-specific approach in which regressors are omitted sequentially based on SIC. The search for improvements in SIC is done in the direction of sequentially removing the variable with the smallest t-value first. It should be noted that by using this fully automated specification procedure we may end up with identical forecasting equations for some periods as the estimated factors may be excluded from the forecast equations whereby they become simple autoregressions.

Pseudo out-of-sample MSFE for forecast horizons of 1 and 4 quarters, and the relative MSFEs, for the four variables of interest are reported in Table 1. The results shown in the Table are from the initial choice of k and p that resulted in the most accurate out-of-sample forecast. Thus, these specific results are only fully valid if the forecaster is believed to know the "optimal" parameters to be used in the general-to-specific procedure.

The last column in Table 1 shows that, overall, the gains in forecast accuracy by applying the regular diffusion index model over the autoregressive model are modest. These results confirm the findings based on monthly Danish data reported in Dahl et al. (2005). Most noticeable, however, is the amount by which the MSFE is reduced by employing the cyclical factor model given by (7). The improvement in forecast accuracy is present for all forecast horizons and variables.

[5] See also Phillips (2005).

Table 1: Recursive out-of-sample forecast comparisons using the estimated cyclical components.

	MSFE			Relative MSFE		
	F	AR	CF	CF/F	CF/AR	F/AR
Private Consumption						
$h = 1$	0.139	0.125	0.095	0.688	0.765	1.112
$h = 4$	0.256	0.266	0.247	0.963	0.926	0.960
GDP						
$h = 1$	0.046	0.045	0.039	0.859	0.869	1.011
$h = 4$	0.136	0.151	0.089	0.651	0.590	0.906
Employment						
$h = 1$	28.729	28.945	26.612	0.926	0.919	0.992
$h = 4$	131.140	126.950	99.752	0.760	0.785	1.033
Inflation						
$h = 1$	0.013	0.013	0.013	0.999	0.999	1.000
$h = 4$	0.071	0.071	0.059	0.824	0.824	1.000

Initial sample: 1986q1-1994q4. Final sample: 1986q1-2003q4. Only the results based on the "best" performing models are reported. For a description of the specification search, see discussion in main text. CF, F and AR denote the cyclical diffusion index model (7), the regular diffusion index model (8) and the autoregressive model (9), respectively. CF/F, CF/AR and F/AR are the forecast accuracy ratios and, finally, h denotes the forecast horizon.

Compared to the regular diffusion index model the improvement is substantial for most horizons and variables – with a maximum reduction of MSFE of 35% (GDP at a 4-quarter (one-year) horizon).

As previously mentioned the results reported in Table 1 are based on the "best" initial choice of k and p, i.e., the values of k and p that yield the lowest MSFE in the pseudo out-of-sample forecasting). As these parameters are not known in real time, the results provided in Table 1 should be interpreted with caution. However, in Figures 2 and 3 we have depicted not only the MSFE associated with the best performing models but all the 16 MSFEs that were calculated for each model based on our search over alternative initial settings for k and p in the forecasting equation.[6] For each model, the MSFEs have been sorted based on the size, with the best performing initial specification (identical to the results in Table 1) to the left, and the worst performing initial specification to the right. The figures show that the forecasting accuracy is indeed sensitive to the initial settings.

When the forecast horizon is one quarter (Figure 2), we notice that the MSFEs

[6]For the linear autoregressive model there are only 8 MSFEs for each variable and forecast horizon. To increase the readability of the plots we report each of these results twice, resulting in the 16 bars shown in the figures.

Figure 2: Distribution of MSFE ($h = 1$) for the regular diffusion index model (1'st bar), the linear AR model (2'nd bar) and the cyclical diffusion index model (3'rd bar).

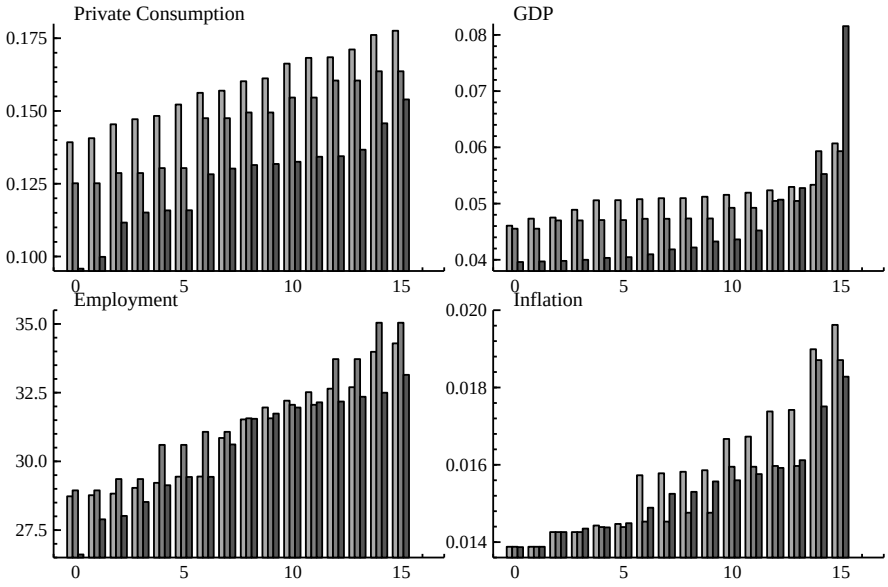
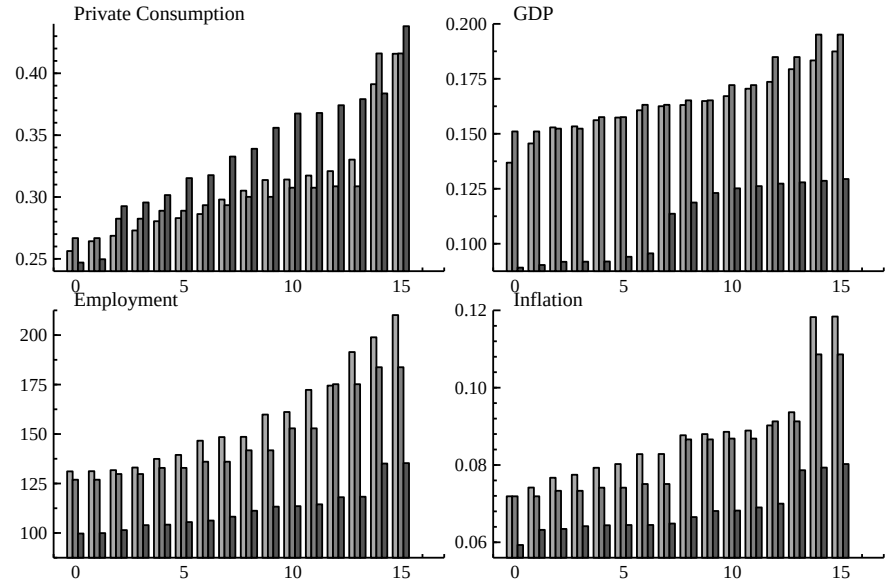


Figure 3: Distribution of MSFE ($h = 4$) for the regular diffusion index model (1'st bar), the linear AR model (2'nd bar) and the cyclical diffusion index model (3'rd bar).



are generally lowest for the cyclical diffusion index model. Yet, the improvement in forecasting accuracy compared to the regular diffusion index model and the autoregressive model is not universal. For some of the initial settings the cyclical diffusion index model performs no better, and in some cases even worse, than the best performing variants of the other models (compare the rightmost MSFE's of the cyclical diffusion model with the leftmost MSFE's of the other models).

For the one-year-ahead forecast horizon the improvement in forecasting accuracy is very pronounced (Figure 3). The MSFEs based on the cyclical diffusion index model are again generally lower than the MSFEs of the other models. For the case of GDP, employment and inflation the cyclical diffusion index model outperforms the regular diffusion index model and the autoregressive model irrespectively of the initial settings. That is, even the worst performing variant of the cyclical model is better than the best performing rival model for these three variables. Only for private consumption is the improvement of forecasting accuracy contingent on choosing the "best" initial settings. Overall, we find the evidence based on Figures 2 and 3 to be very encouraging as it indicates a reasonable degree of robustness in our findings on the increased forecast accuracy of the cyclical component factor model.

5 Conclusion

We have suggested a new and simple approach to improve the out-of-sample forecast of factor models based on large data sets which was introduced by Stock and Watson (1998, 2002a,b). The basic idea is to assume that it is the pure cyclical component of the series that allows a factor representation. We suggest using an estimator of the pure cyclical components based on the SSF-package of Koopman et al. (1999) which numerically is easy to obtain. Monte Carlo simulations suggest that the modification may actually improve the forecast performance of the factor model when the variances of the irregular components are large while the number of time series in the factor model is relatively small. Our empirical illustration demonstrates that our approach improves the out-of-sample forecast accuracy substantially relative to the regular diffusion index model for four Danish macroeconomic variables.

Acknowledgements: The authors would like to thank S. Koopman, N. Shephard and J. Doornik, who wrote the SSF-package for Ox (Doornik, 2001) which was used for estimation of the cyclical components in this paper. The authors are also grateful for financial support received from the Danish Social Science Research Council.

References

Armah, N.A. and N. Swanson (2007). Seeing inside the black box: using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments. Mimeo. Department of Economics, Rutgers University. http://ssrn.com/abstract=966438.

Artis, M., A. Banerjee and M. Marcellino (2005). Factor forecasts for the UK. *Journal of Forecasting*, 24, 279-298.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70, 191-221.

Bai, J. and S. Ng (2008). Boosting Diffusion Indexes. Working Paper. University of Michigan, Ann-Arbor.

Bai, J. and S. Ng (forthcoming). Forecasting economic time series using targeted predictors. *Journal of Econometrics*.

Banerjee, A. and M. Marcellino (2006). Are there any reliable leading indicators for US inflation and GDP growth? *International Journal of Forecasting*, 22, 137-151.

Camacho, M. and I. Sancho (2003). Spanish diffusion indexes. *Spanish Economic Review*, 5, 173-203

Dahl, C., H. Hansen and J. Smidt (2005). Makroøkonomiske forudsigelser baseret på diffusionsindeks. *The Danish Economic Journal*, 143, 125-152.

Doornik, J. (2001). *Ox An Object-oriented Matrix Programming Language*. London: Timberlake Consultants Ltd.

Durbin, J., and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Kaiser, R. and A. Maravall, (1999). Short-Term and Long-Term Trends, Seasonal Adjustment, and the Business Cycle. Working Paper 99-10(2). Statistics and Econometrics Series, Universidad Carlos III de Madrid

Koopman, S. J., N. Shephard og J. A. Doornik (1999). Statistical algorithms for models in state space form using SsfPack 2.2. *Econometrics Journal*, 2, 113-166.

Marcellino M., Stock, J.H. and M.W. Watson (2003). Macroeconomic forecasting in the euro area: country specific versus euro wide information. *European Economic Review* 47, 1-18.

Phillips P.C.B (2005). Automated Discovery in Econometrics, *Econometric Theory*, 21, 3-21.

Stock, J.H. and M.W. Watson (1998). Diffusion indexes. Working Paper 6702. National Bureau of Economic Research. Cambridge, MA.

Stock, J.H. and M.W. Watson (2002a). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economics Statistics*, 20, 147-162.

Stock, J.H. and M.W. Watson (2002b). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97, 1167-1179.

White H (2000). A Reality Check For Data Snooping, *Econometrica*, 68, 1097-1127.

Research Papers 2008

2008-29: Per Frederiksen, Frank S. Nielsen and Morten Ørregaard Nielsen: Local polynomial Whittle estimation of perturbed fractional processes

2008-30: Mika Meitz and Pentti Saikkonen: Parameter estimation in nonlinear AR-GARCH models

2008-31: Ingmar Nolte and Valeri Voev: Estimating High-Frequency Based (Co-) Variances: A Unified Approach

2008-32: Martin Møller Andreasen: How to Maximize the Likelihood Function for a DSGE Model

2008-33: Martin Møller Andreasen: Non-linear DSGE Models, The Central Difference Kalman Filter, and The Mean Shifted Particle Filter

2008-34: Mark Podolskij and Daniel Ziggel: New tests for jumps: a threshold-based approach

2008-35: Per Frederiksen and Morten Ørregaard Nielsen: Bias-reduced estimation of long memory stochastic volatility

2008-36: Morten Ørregaard Nielsen: A Powerful Test of the Autoregressive Unit Root Hypothesis Based on a Tuning Parameter Free Statistic

2008-37: Dennis Kristensen: Uniform Convergence Rates of Kernel Estimators with Heterogenous, Dependent Data

2008-38: Christian M. Dahl and Emma M. Iglesias: The limiting properties of the QMLE in a general class of asymmetric volatility models

2008-39: Roxana Chiriac and Valeri Voev: Modelling and Forecasting Multivariate Realized Volatility

2008-40: Stig Vinther Møller: Consumption growth and time-varying expected stock returns

2008-41: Lars Stentoft: American Option Pricing using GARCH models and the Normal Inverse Gaussian distribution

2008-42: Ole E. Barndorff-Nielsen, Silja Kinnebrock and Neil Shephard: Measuring downside risk — realised semivariance

2008-43: Martin Møller Andreasen: Explaining Macroeconomic and Term Structure Dynamics Jointly in a Non-linear DSGE Model

2008-44: Christian M. Dahl, Henrik Hansen and John Smidt: The cyclical component factor model