

The cyclical component factor model

Christian M. Dahl* Henrik Hansen
Department of Economics Department of Economics
Purdue University University of Copenhagen

John Smidt
The Danish Economic Council

August 28, 2005

Abstract

Recently, factor models based on large data sets have received ample attention due to their ability to increase forecast accuracy with respect to a wide range of key macroeconomic variables, see, e.g., Stock and Watson (1998, 2002a,b), Marcellino et al. (2001,2003) and Artis et al. (2005). In this paper we propose to estimate the factors based on the pure cyclical components of the series entering the "large" data set. We provide an empirical illustration showing that this procedure indeed improves on pseudo real time forecast accuracy.

1 Introduction

In this paper we propose a relatively simple method to potentially improve the forecast accuracy based on diffusion index models and provide an empirical illustration showing its usefulness. Our approach is inspired by the work of Camacho and Sancho (2004) and Kaiser and Maravall (1999) and 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 before estimating the factors. Consequently, according to this view, it is the pure cyclical component that allows a factor representation. In situations where the irregular component is relatively large, we conjecture that our approach will provide a more accurate estimate of the factors. As argued by Dahl, Hansen and Smidt (2005) the irregular component in Danish data seems to be much more dominating than in, for example, US data. This might explain why diffusion index models estimated based on Danish data perform relatively disappointing. We show that by estimating the factors

*Corresponding author. Address: 403 West State Street, Purdue University, West Lafayette, IN 47907-2056. E-mail: dahlc@mgmt.purdue.edu. Phone: 765-494-4503. Fax: 765-496-1778.

based on an estimate of the "pure" cyclical components the predictive accuracy of the diffusion index model is improved substantially.

2 The modelling framework

Consider the "large" collection of variables $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_T)$, where $\mathbf{X}_t = (x_{1t}, x_{2t}, \dots, x_{Nt})'$. Let g_{tj} denote a trend component, c_{tj} the business cycle component, s_{tj} a seasonal component and e_{tj} the irregular component and assume that x_{it} can be represented as

$$x_{it} = g_{it} + c_{it} + s_{it} + e_{it}, \quad (1)$$

for $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$. In most of the existing work on diffusion models it is common to make prior adjustments of 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 6 times the standard error). Consequently, using the 'traditional' Stock and Watson (1998,2002a,b) approach, the estimator of the common factors is based on the relation

$$\hat{x}_{it} = \boldsymbol{\lambda}_i \mathbf{F}_t + \eta_{it}, \quad (2)$$

where \hat{x}_{it} is the trend and seasonally adjusted series (assuming that there are no outliers), $\mathbf{F}_t = (f_{1t}, f_{2t}, \dots, f_{rt})'$ are the common factors and $\boldsymbol{\lambda}_i = (\lambda_{1i}, \lambda_{12}, \dots, \lambda_{1r})$ are the factor loadings. In addition, it is typically assumed that η_{it} is an idiosyncratic error term.

The main new contribution of this paper is to propose a modification of (2) by explicitly assuming that it is the business cycle component in (1) that admits a linear factor representation with r common factors, i.e.,

$$c_{it} = \boldsymbol{\lambda}_i \mathbf{F}_t + v_{it}, \quad (3)$$

where v_{it} has the same properties as η_{it} . According to (1), define the "true" trend and seasonally adjusted series $\tilde{x}_{it} \stackrel{def.}{=} c_{it} + e_{it}$, and assume that the estimator \hat{x}_{it} satisfies the condition

$$\hat{x}_{it} = \tilde{x}_{it} + \hat{\epsilon}_{it}, \quad (4)$$

where $\hat{\epsilon}_{it}$ is the estimation error associated with the trend and seasonal adjustment procedure. Note, that given (1), (3) and (4) we can write

$$\hat{x}_{it} = \boldsymbol{\lambda}_i \mathbf{F}_t + v_{it} + e_{it} + \hat{\epsilon}_{it}, \quad (5)$$

which is observational equivalent to (2) with

$$\eta_{it} = v_{it} + e_{it} + \hat{\epsilon}_{it},$$

where also $\widehat{\epsilon}_{it}$ is assumed to be an idiosyncratic innovation term. Given the representation in (1) and the factor model (3), and assuming that one could actually observe c_{it} , it would obviously be more informative to estimate \mathbf{F}_t based on (3) instead of (2). In reality, however, one does not observe c_{it} . Yet, an alternative to using the model in (2) is to use an estimate of the cyclical component, based on the individual series. Consequently, if we let \widehat{c}_{it} be an estimator of c_{it} and let $\widehat{\epsilon}_{it}^c$ denote the associated estimation error, equation (3) can be represented as

$$\widehat{c}_{it} = \boldsymbol{\lambda}_i \mathbf{F}_t + v_{it} + \widehat{\epsilon}_{it}^c, \quad (6)$$

The estimator of \mathbf{F}_t based on (6) is no longer guaranteed to be more informative relative to the estimator based on (2) and it becomes 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 favor the estimation approach based on (6).

An estimate of c_{it} can be obtained under some additional assumption on 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). In particular, following Koopman et al. (1999), we specify a local linear trend model for the trend component, a trigonometric representation of the seasonal component, while the cyclical component is assumed to have a stochastic and time varying sine wave representation. The irregular component is a Gaussian white noise process.¹ The estimation procedures, based on using the linear Gaussian State Space representation and the Kalman filter, was carried out by applying the SSF-package by Koopman et al. (1999).²

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

$$(y_{t+h} - y_t) = \sum_{j=1}^r \beta_j^c \widehat{f}_{jt}^c + \sum_{j=1}^p \alpha_j^c \Delta y_{t-j} + v_t^c, \quad (7)$$

for $t = 1, 2, \dots, T$, where the estimated factors $\widehat{\mathbf{F}}_t^c = (\widehat{f}_{1t}^c, \dots, \widehat{f}_{rt}^c)'$ are based on (6). We wish to compare (7) to the "standard" factor forecasting model given as

$$(y_{t+h} - y_t) = \sum_{j=1}^r \beta_j^s \widehat{f}_{jt}^s + \sum_{j=1}^p \alpha_j^s \Delta y_{t-j} + v_t^s, \quad (8)$$

¹The exact mathematical expressions for each of the unobserved components (which are standard) can be found in Koopman et al. (1999) and Durbin and Koopman (2001).

²Alternatively the business cycle component estimate could be obtained using the bandpass filter, see, e.g., Baxter and King (1999).

where the estimator $\widehat{\mathbf{F}}_t^s = (\widehat{f}_{1t}^s, \dots, \widehat{f}_{rt}^s)'$ is obtained based on (2) and to the pure autoregressive linear model

$$(y_{t+h} - y_t) = \sum_{j=1}^p \alpha_j^l \Delta y_{it-j} + v_t^l. \quad (9)$$

3 Empirical Illustration

The empirical illustration will be 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 as shown in the previous section this may explain why the traditional 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.

3.1 The data and the estimated factors

The data set for Denmark, our \mathbf{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 follows closely Stock and Watson (2002a,b) suggestions, 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).³

It is important to note that the estimator of c_{it} obtained by applying the SSF-package by Koopman et al. (1999) is two-sided, i.e., the estimator is based on the entire sample and not observations up to time t only. By ignoring this feature the forecasts of $(y_{t+h} - y_t)$ based on $\widehat{\mathbf{F}}_t^c$ would be heavily favored, by construction, as it would be based on a larger information set relative to the forecast based on (8) and (9). We account for this by computing the estimator for \mathbf{C} (given by

³The data and documentation can be obtain from the corresponding author.

$\widehat{\mathbf{C}} = (\widehat{\mathbf{C}}'_1, \widehat{\mathbf{C}}'_2, \dots, \widehat{\mathbf{C}}'_S)$, where $\widehat{\mathbf{C}}_s = (\widehat{c}_{1s}, \widehat{c}_{2s}, \dots, \widehat{c}_{N_s s})'$ recursively, i.e., each time the sample is expanded in the pseudo out-of-sample interval. Subsequently, $\widehat{\mathbf{F}}_S^c$ is computed using $\widehat{\mathbf{C}}$ for each S in the out-of-sample interval and is used to predict $(y_{S+h} - y_S)$ only. By this recursive procedure the information set on which the forecasts are conditioned will be identical across all three forecasting models given by (7), (8) and (9). In addition, this procedure will to a larger extent reflect the situation an actual real time forecaster is facing.

When combining monthly as well as quarterly data in \mathbf{X} and \mathbf{C} they become unbalanced data matrices. When estimating the factors we therefore employ the EM algorithm described in Stock and Watson (1998,2002a,b), where the optimal number of factors are determined by the selection criterion suggested by Bai and Ng (2002).

3.2 The forecasting framework

As already mentioned we are interested in comparing the out-of-sample forecast accuracy of the model based on (7) relative to (8) and (9). Each of these forecasting equations are estimated recursively using ordinary least squares, applying a standard general-to-specific procedure. To make a fair comparison model selection is obvious very important. As pointed out by Dahl et al. (2005) it is typically possible to find a model configuration based on (8) that can outperform the linear autoregressive model in terms of predictive accuracy simply by doing a search wide enough. Obviously, this is a result 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. We will try to avoid this pitfall by reporting the predictive outcome of all the models configurations selected by an automated general-to-specific selection mechanism. In that respect it is important to note that all the variables entering (7), (8) and (9) are estimated. It is unclear what effect this may have on the general-to-specific model selection procedure, i.e., what is the appropriate critical value when faced with generated right and left hand side variables. In particular, it turns out that the outcome of the general-to-specific procedure actually is very sensitive to the initial choice of r and p , see also the discussion by Phillips (2005). As a consequence, the search is expanded to include a general-to-specific model selection procedure for all possible combinations of initial settings for $r = 1, 2$, and $p = 1, 2, \dots, 8$. This implies that 16 alternative measures of MSFE is computed over the out-of-sample period for each of the diffusion models (8 for the autoregressive model). We choose to report results based on a relative low number of choices of r primarily for simplicity but also to obtain a certain degree of robustness of the results as pointed out by Artis et al. (2005).⁴ The automated model selection procedure we employ for each r

⁴We also did the analysis for $r = 1, 2, \dots, 4$ and $p = 1, 2, \dots, 8$. The main results did not change.

Table 1: Recursive out-of-sample forecast comparisons using the estimated cyclical components. Initial sample: 1986q1-1994q12. Final sample: 1986q1-2003q12. Only the results based on the "best" performing models are reported. For a description of the specification search, see discussion in main text. F, AR and CF denote the diffusion index model, the autoregressive model and the cyclical diffusion index model respectively. CF/F, CF/AR and F/AR are the forecast accuracy ratios and finally h denotes the forecast horizon.

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

and p is the traditional general-to-specific approach where the most insignificant terms are removed sequentially based on individual t-statistics. The first period used for estimation is 1986q1. The first period in the pseudo out-of-sample is 1995q1 and the last period is 2003q4.

In Table 1 the pseudo out-of-sample MSFE and relative MSFE for the four variables of interest are reported based on (7), (8) and (9). Only the results based on the most accurate specifications are reported. By inspection of the last column in Table 1, we see that overall the gains in forecast accuracy by applying the "traditional" diffusion index model over the autoregressive model are relative 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. Compared to the "traditional" diffusion index model the improvement is substantial for most horizons and variables – with a maximum reduction of MSFE of 35% (GDP at a one-year horizon).

In Figures 1 and 2 we have depicted not only the MSFE associated with the best performing models but all the MSFE's that was calculated based on our search over alternative initial settings for r and p in the forecasting equation.

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

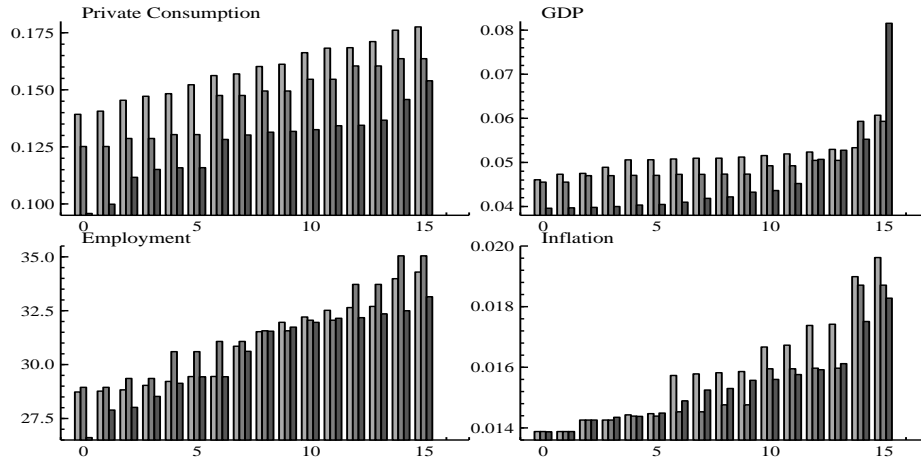
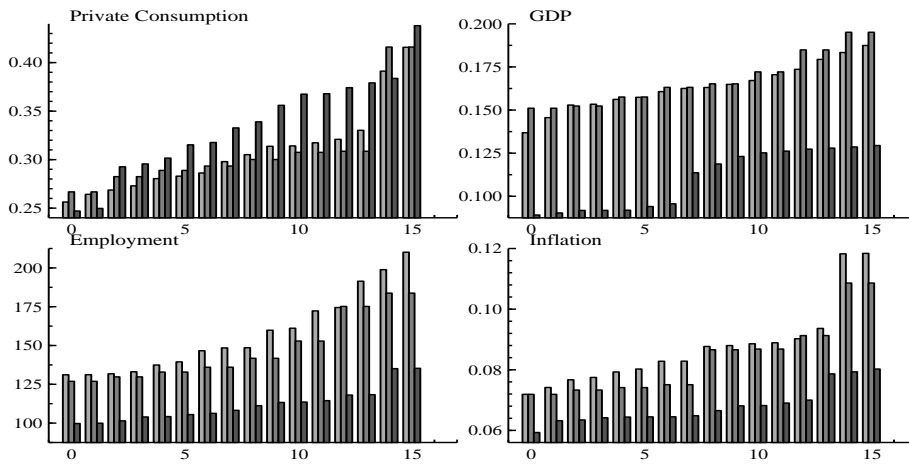


Figure 2: Distribution of MSFE ($h=4$) for the traditional diffusion index model (1'st bar), the linear AR model (2'nd bar) and the cyclical diffusion index model.



The MSFE's have been sorted based on the size, with the best performing model (identical to the results in table 1) to the left, and the worst performing model (or initial setting) to the right. The figures show, that the forecasting accuracy is indeed sensitive to the initial settings.

By inspection of Figure 1, which is for a forecast horizon of one quarter, we notice that the MSFE's are generally lowest for the cyclical diffusion index model. However, the improvement in forecasting accuracy compared to the "traditional" diffusion index model and the autoregressive model is not universal. For some of the initial settings the cyclical diffusion index model performs no better or 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 2). The MSFE's based on the cyclical diffusion index model are again generally lower than the MSFE's of the other models. For the case of GDP, employment and inflation the cyclical diffusion index model outperforms the "traditional" diffusion index model and the autoregressive model irrespectively of the initial settings: 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, the improvement of forecasting accuracy is contingent on choosing the "best" initial settings. Overall, however, the evidence based on Figures 1 and 2 is very encouraging as it indicates a reasonable degree of robustness in our findings on the increased forecast accuracy of the cyclical component factor model.

4 Conclusion

We have suggested a new and simple approach to improve the out-of-sample forecast of factor models based on large data sets. 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. Our empirical illustration demonstrates that our approach can improve the out-of-sample forecast accuracy substantially relative to the traditional diffusion index model suggested by Stock and Watson (1998, 2002a,b).

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

- 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.
- Baxter, M and R.G. King (1999). Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series. *Review of Economics and Statistics*, 81, 575-593.
- 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*, (forthcoming).
- 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 Cycles, Bank of Spain, Mimeo.
- 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, 113-66.
- Marcellino M., Stock, J.H. and M.W. Watson (2001), A dynamic factor analysis of the Euro area. Mimeo.
- 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.