Improved Construction of Diffusion Indexes for Macroeconomic Forecasting

Christiaan Heij[*], Dick J. van Dijk, Patrick J.F. Groenen
Erasmus University Rotterdam

Econometric Institute Report EI 2006 -03

Revised version, August 30, 2006

Abstract

This article proposes an improved method for the construction of diffusion indexes in macroeconomic forecasting using principal component regression. The method aims to maximize the amount of variance of the original predictor variables retained by the diffusion indexes, by matching the data windows used for constructing the principal components and for estimating the diffusion index models. The method is analyzed by means of extensive Monte Carlo simulations, as well as an empirical application to forecast eight monthly US macroeconomic time series, using the data set of Stock and Watson (2002a). The results show that the proposed modification leads, on average, to better forecasts than previously used principal component methods.

Keywords forecasting, principal components, factor construction

JEL classification C32, C53, E17

[*] Corresponding author, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, email: heij@few.eur.nl, fax: +31-10-4089162, tel: +31-10-4081269.

1 INTRODUCTION

One of the basic questions in empirical forecasting is which information should be included in the forecast model. For instance, in many macroeconomic and financial applications, a large number of predictor variables is available. The forecaster then faces the challenge to employ the available information in the best possible way. Various methods for forecasting with many predictors have been proposed in the literature, including forecast combination, model averaging, variable selection, and predictor combination. We refer to Stock and Watson (in press) for a survey. Several empirical studies in macroeconomic forecasting indicate that, on average, the best forecast results are obtained by principal component regression (PCR), see Stock and Watson (1999, 2005) and Lin and Tsay (2005), among others. In PCR, the predictors are summarized by means of a limited number of factor components.

In this article, we show that further gains in the forecast accuracy of PCR can be achieved by constructing the principal components somewhat differently as compared to the method usually employed in the literature. We call our method 'matched PCR' (MPCR), as it matches two data windows that are used in PCR. More precisely, the distinction between PCR and MPCR lies in the construction of the factors. In PCR, the h-step ahead forecast made at time T is based on the principal components computed from the (standardized) predictor variables using observations up to and including time T. The diffusion index forecast model, however, is estimated using the observations only up to time $T - h$. In MPCR, the data window used for constructing the principal components is matched with the data window used for estimating the diffusion index models, by extracting the principal components from the (standardized) predictor variables up to time $T - h$. This modification better achieves the goal of principal components, namely, to retain the maximal amount of variance of the original predictor variables.

The article is structured as follows. In Section 2, we outline the current method of forecasting with diffusion indexes and we present our method of matched PCR. The relative forecast performance of the original and matched PCR methods is evaluated in Section 3 by means of a simulation experiment, based on Stock and Watson (2002b). Section 4 contains an empirical application involving forecasts of four real economic variables and four price variables from a set of 146 macroeconomic predictor variables, using the data set from Stock and Watson (2002a). Section 5 concludes.

2 FORECASTING WITH DIFFUSION INDEXES

2.1 Diffusion index models

In this section, we briefly summarize the method of principal component regression (PCR) proposed by Stock and Watson (1999, 2002a,b), to which we refer for further details. The corresponding forecast models are also called 'diffusion index' models, as the principal components can be interpreted as indexes that summarize the common movements in the underlying macroeconomic predictor variables.

Let y denote the economic variable of interest and let X denote a set of N predictor variables. In PCR, the information in the N predictor variables is summarized by means of k factors f, with k (much) smaller than N. These factors are used to forecast y by means of a linear regression model. Let h be the forecast horizon and let t denote the current time moment, then the h-step-ahead forecast model is written as

$$y_{t+h}^{h} = \alpha + \sum_{j=1}^{m} \beta_j' f_{t-j+1} + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + \varepsilon_{t+h}^{h}. \tag{1}$$

Here y_{t+h}^{h} denotes the h-step-ahead variable to be forecasted. Following Stock and Watson (1999, 2002a,b), we will forecast the h-period average of y_t, so that $y_{t+h}^{h} = \frac{1}{h}\sum_{j=1}^{h} y_{t+j}$. The model (1) is denoted by DI-AR-Lag. DI-AR is the model without lagged factors ($m = 1$), and DI is the model with f_t as the only regressor variable ($m = 1$ and $p = 0$). Let data on y and X be available over a period of length T, then y_{t+h}^{h} can be computed for $t \leq T - h$. Regression in (1) requires that the effective sample size is at least as large as the number of unknown parameters, so that $T - h \geq 1 + km + p$. In particular, this requires that

$$T - h > km. \tag{2}$$

The PCR forecast at time T is computed in two steps. First, the factors f are estimated by means of the leading k principal components of X over the time interval $[1, T]$, where the predictor variables are standardized to have zero mean and unit variance on this interval. Second, the parameters in (1) are estimated by a regression on the time interval $[l + 1, T - h]$, where $l = \max(m - 1, p - 1)$, replacing the terms f_{t-j+1} by their corresponding principal component values. In practice, appropriate values for k, m, and p have to be selected, for instance, by means of the Bayes Information Criterion (BIC) over a set of models with $k \leq K$, $m \leq M$ and $p \leq P$. The results in Ng and Perron (2005) motivate the use of equal effective sample sizes for all candidate models, so that all models are estimated on the time interval $[L + 1, T - h]$ where $L = \max(M - 1, P - 1)$ is the maximum considered lag.

2.2 Modified factor construction

The quality of the PCR forecasts depends on the quality of (i) the forecast model (1), (ii) the estimates of the factors f_t, and (iii) the estimates of the parameters $(\alpha, \beta_j, \gamma_j)$. The use of principal components in (ii) can be motivated by the fact that these components account for the largest possible variance of the predictor variables, see Anderson (1984) and Jolliffe (2002). This property is of importance, as a larger predictor variance reduces the standard errors of the parameter estimates in (1) and, therefore, the forecast variance. However, PCR maximizes the accounted predictor variance over the time interval $[1, T]$, whereas the parameter estimates in (1) are obtained from observations on the smaller interval $[L + 1, T - h]$, using the factor components for $t = L - M + 2, \ldots, T - h$. This imbalance motivates our modification, that is, to construct the principal components on the relevant estimation interval.

Figure 1 summarizes the data windows for the original PCR and our matched PCR methods. Note that our modification includes normalizing the original predictor variables X to have mean zero and unit variance on the interval $[L-M+2, T-h]$ as well. More precisely, the k PCR factors f_t consist of the leading k principal components of the N predictors x_t which are normalized on the interval $[1, T]$. These factors consist of linear combinations $f_t = Ax_t$ for $1 \leq t \leq T$, where A is a $(k \times N)$ matrix. On the other hand, the MPCR factors $f_{m,t}$ are constructed by extracting the leading k principal components from x_t normalized on the interval $[L - M + 2, T - h]$, with $f_{m,t} = A_m x_t$ for $L - M + 2 \leq t \leq T - h$, where A_m is a $(k \times N)$ matrix. This matrix is then used to construct the MPCR factors $f_t = A_m x_t$ also for $1 \leq t < L - M + 2$ and $T - h < t \leq T$.

<< **FIGURE 1 to be inserted somewhere over here.** >>

3 SIMULATION EXPERIMENT

3.1 Monte Carlo design

In this section, we analyze the relative merits of the original and our matched PCR methods by means of Monte Carlo simulations. As was discussed in Section 2.2, matched PCR is based on the idea to match the window for extracting the factors with the estimation window of the forecast equation (1). The simulations are meant to clarify which aspects of the data are important for the relative forecast quality of the two methods. Relevant design parameters in this respect are the number of observations (T), the forecast horizon (h), the number of predictors (N), the number of latent factors (k) and the number of factor lags (q), as well as the correlations within the set of predictors and the correlations between the predictors and the latent factors.

We use a similar Monte Carlo design as in Stock and Watson (2002b). The predictors x_{it}, the variable of interest y_t and the latent factors f_t are generated by the equations

$$
\begin{aligned}
x_{it} &= \sum_{j=0}^{q} \lambda_{ijt} f_{t-j} + e_{it}, && (3)\\
\lambda_{ijt} &= \lambda_{ij,t-1} + (c/T) w_{ijt}, && (4)\\
f_t &= \alpha f_{t-1} + u_t, && (5)\\
e_{it} &= \rho e_{i,t-1} + \gamma e_{i-1,t} - \rho\gamma e_{i-1,t-1} + v_{it}, && (6)\\
y_t &= \sum_{j=0}^{q} (1, \ldots, 1) f_{t-j} + \varepsilon_t, && (7)
\end{aligned}
$$

where $i = 1, \ldots, N, t = 1, \ldots, T, j = 0, \ldots, q$, and $w_{ijt}, v_{it}, \varepsilon_t$, and all k components of u_t are mutually independent NID(0,1) random variables. This design corresponds to a dynamic factor model for the N predictor variables x_{it} in terms of k latent factors f_t, with time varying loadings (if $c \neq 0$) and first-order correlations across time (if $\rho \neq 0$) and across variables (if $\gamma \neq 0$). Initial values of f_0 and e_{i0} are drawn from the (stationary) marginal distributions corresponding to (5) and (6), with variance $v_f = 1/(1-\alpha^2)$ and $v_e = 1/((1-\rho^2)(1-\gamma^2))$ respectively. Further, initial values of λ_{ij0} are drawn in line with Stock and Watson (2002b, p. 1171), allowing for varying importance of the predictors and even for irrelevant predictors. That is, for the i-th predictor variable, λ_{ij0} is generated as follows: (i) draw R_i^2, with probability π for the value 0 and with probability $(1-\pi)$ from the uniform distribution on $[0.1, 0.8]$; (ii) draw the $1 \times k$ vector $\overline{\lambda}_{ij0}$ from $N(0, I_r)$, independent of all other error terms; (iii) define $\lambda_{ij0} = d_i \overline{\lambda}_{ij0}$, where d_i is chosen such that the fraction of the variance of x_{i0} explained by the factors f_0 is equal to R_i^2. The last step is solved by taking $q = 0$ in (3), in which case x_{i0} has variance $d_i^2 v_f + v_e$ and explained variance $d_i^2 v_f$, so that $R_i^2 = d_i^2 v_f/(d_i^2 v_f + v_e)$ from which d_i is easily solved in terms of $(R_i^2, \alpha, \rho, \gamma)$.

The variable to be forecasted is the h-period average $y_{t+h}^h = \frac{1}{h}\sum_{j=1}^h y_{t+j}$. The horizon h is one of the main parameters of interest in our comparison of PCR and MPCR. To achieve comparable predictability for different forecast horizons, the value of α is chosen as a function of h, as follows. Let $\hat{y}_{t+h}^h$ denote the optimal forecast of y_{t+h}^h that is based on exact knowledge of all design parameters in (3)-(7) and let $R_f^2 = \text{var}(\hat{y}_{t+h})/\text{var}(y_{t+h})$ be the corresponding forecast R-squared. A rough approximation is given by $R_f^2 \approx \alpha^2/(h(1-\alpha^2))$ (see the appendix), so that a given forecast R-squared is approximately obtained by taking $\alpha = \sqrt{hR_f^2/(1+hR_f^2)}$.

3.2 Design parameters

The simulation design (3)-(7) has the following ten design parameters (the considered values are in parentheses and are mostly in line with the specifications in Stock and Watson (2002b, Table 1)): the forecast horizon h (5, 10, 25, 50), the number of predictors N (50, 100, 250, 500), the number of observations T (25, 50, 100, 250), the number of latent factors k (5, 10, 20, 40), the number of factor lags q (0, 1, 2), the forecast R-squared R_f^2 (0.1, 0.5, 0.9), the correlation parameters ρ and γ ((0,0), (0.9,0), (0, 0.9), (0.45, 0.45)), the amount of time variation in the factor loadings c (0, 10), and the fraction of irrelevant predictors π (0, 0.25). This gives a total of 36864 designs, but the restriction (2) (with $m = q + 1$) rules out some of them. Further, because of computing time considerations, we limit the analysis of designs with $T = 250$ and $N = 250$ or 500. We will analyze the simulation outcomes for three sets of designs, that is, (i) all designs with $T \leq 100$, (ii) 'simple' designs, and (iii) 'complex' designs. Here we define the design to be simple if it is close to the assumptions of principal component analysis, and we define this set by the conditions that $R_f^2 = 0.5$ and $\rho = \gamma = c = \pi = 0$. This gives 768 simple designs in total, some of which are eliminated because of (2). An equally large set of complex designs is defined by the conditions that $R_f^2 = 0.5$ and $\rho = \gamma = 0.45$, $c = 10$ and $\pi = 0.25$, in which case the predictors are correlated across time and across variables, a considerable part of them is irrelevant, and the factor loadings are time varying. This set may be of most interest for practical applications, as many data sets contain correlated predictors and time varying characteristics.

For each specific configuration of the design parameters, we perform 1000 replications with fixed time T to forecast (at estimation time $T - h$). The number k of factors is given, but the number of lags q is selected by BIC. For reasons of computational efficiency, the number of factors is not selected from the data, as the considered designs have up to forty latent factors. However, we performed also simulations where k and q are jointly determined by BIC, and the outcomes are similar to the ones with given value of k. This finding is in line with the results in Stock and Watson (2002b, Table 1).

Finally, we mention that we also considered h-step-ahead forecasting with $y^h_{t+h} = y_{t+h}$. The results are qualitatively very much the same as for the case considered in this paper, that is, the relative forecast performance of MPCR as compared to PCR is affected in the same way by the various design parameters. The effects are even quite comparable in quantitative terms, but to save space we will not discuss this any further (details are available on request).

3.3 Forecast results

We compare the forecast accuracy of MPCR with that of PCR in terms of the relative mean squared prediction error (MSE). For each specific design, the MSE is defined by $\sum_i e^2_{m,i} / \sum_i e^2_i$, where $e_{m,i}$ is the forecast error of MPCR in the i-th replication and e_i is the corresponding forecast error of PCR ($i = 1, \ldots, 1000$). Tables 1-3 show mean MSE values over different sets of designs, controlling for one or several of the design parameters. The MSE is expressed in percentage form, so that values less than 100 indicate better performance of MPCR as compared to PCR.

<< TABLES 1-3 to be inserted around here. >>

MPCR performs, on average, better than PCR for every considered design parameter, as all average MSE values in Table 1 are smaller than 100. The gain is larger for complex designs

(23.6% on average) than for simple designs (10.2%). Further, MPCR gains more if the observation interval (T) is short, the number of predictor variables (N) is large, the number of latent factors and lags (k, q) is large, the factor loadings are time varying ($c \neq 0$), and some of the predictors are irrelevant ($\pi \neq 0$). The forecast R-squared (R_f^2) has hardly any effect. The same is true for the predictor correlations (ρ, γ), except for the case $(\rho, \gamma) = (0, 0.9)$ where the gains are relatively smaller. The results for the forecast horizon (h) in Table 1 seem to suggest that MPCR gains more for small horizons. However, this is misleading, because the restriction (2), with $m = q + 1$, implies that some designs with large values for (k, q) are ruled out if h is large and T is small. This restriction causes a positive correlation between h and T over the considered sets of simulation designs, as small values of T rule out large values of h. The correlation between h and T is 35% for the set of all designs and 23% for the sets of simple and complex designs. Therefore, the joint effect of the design parameters (h, T, N) is analyzed in more detail in Table 3. For $T = 250$, (2) imposes no restriction on the considered designs, and in this case the gains of MPCR tend to be larger for larger h, as expected. Note that, e.g. for $T = 100$, designs with $(k, q) = (40, 1)$ give $k(q + 1) = 80$, so that estimation is possible only for $h \leq 10$ and not for $h \geq 25$. As a consequence, the MSE values for $T = 100$ in Table 3 tend to decrease from $h = 5$ to $h = 10$ (as expected), but then increase for $h \geq 25$, as a set of designs is ruled out where the differences between MPCR and PCR are more prominent. Similar arguments explain the results for $T = 250$, 50 and 25. Table 3 clearly illustrates that the gains of MPCR tend to be larger for a small observation interval and for a large number of predictors.

Table 2 shows further statistics, where we consider also the subsets of designs where MPCR performs worse than PCR and where it performs at least twice as well. MPCR performs better than PCR in the far majority of cases (98.3% for the set of all designs and for the set of complex designs, and 91.7% for the set of simple designs). In the few cases where MPCR performs worse, the loss is often very small, with a median of around 1% for all three design classes. The worst case is a loss of 43.6%, which occurs for a design with $h = 5$, $T = 25$, $N = 500$, $k = 5$ and $q = 2$. In this case, the forecast equation (7), with a constant included, contains $1 + k(q + 1) = 16$ parameters, and the effective sample size is $T = 25$ for PCR and $T - h = 20$ for MPCR. So the number of degrees of freedom is 9 for PCR and only 4 for MPCR, which explains the relatively bad forecast performance of MPCR in this case. Table 2 shows further that large forecast gains (with factor two or larger) are mostly obtained within the classes of all designs and complex designs.

The results in Tables 1-3 concern average effects. It is also of interest to investigate partial effects. Because of the large number of design parameters, we use response surfaces as in Boivin and Ng (2006). A response surface is obtained by regressing the relative MSE on the set of ten design parameters. The resulting regression coefficients measure the partial effect of each design parameter on the relative MSE. Table 4 shows estimation results for several response surfaces (the table contains also some other results that will be discussed in the next section). The regression coefficients are displayed only if they are significant at the 0.01% level, using robust standard errors. Apart from linear surfaces, we consider also a second-order one. This surface has in principle ten linear and fifty-five second-order terms, but for simplicity we present only the signs of the coefficients of a restricted specification, including interactions between the 'data' characteristics (h, N, T) and between the 'estimation' characteristics (h, T, k, q) in the restriction (2), with $m = q + 1$. The linear response surface for the set of all designs shows that MPCR gains more for larger values of (h, N, k, q, c, π) and for smaller values of (T, γ), whereas the effects of R_f^2 and ρ are not significant. These results are in line with the ones in Tables 1-3. The quadratic response surface has linear terms that mostly have the same sign as in the linear specification. The signs of (h^2, T^2, N^2) are opposite to those of (h, T, N), indicating that the marginal effects of these parameters level off. The negative sign of the interaction term kq means that the effects of k and q enforce each other, which is because the number of coefficients in the forecast equation (7) is $k(q + 1)$. The positive sign of the interaction terms Tk and Tq means that the effects of k and q are smaller for larger T, which is because the number of degrees of freedom in (7) is $T - k(q + 1)$ for PCR and $T - h - k(q + 1)$ for MPCR.

3.4 Other evaluation criteria

Apart from the MSE, Table 4 contains also results of linear response surfaces for some other criteria. For each criterion, the surface is obtained by regressing the relevant performance index on the set of ten design parameters. For a given design configuration and criterion, the performance index is computed as follows. The criterion value is computed, for both PCR and MPCR, in 1000 replications, and the performance index is obtained by dividing the resulting mean for MPCR over the 1000 replications by the mean for PCR.

We use three criteria to evaluate the estimation accuracy of the forecast equation (7). We include a constant term in estimating this equation, so that the number of regressors is $d = k(q + 1) + 1$. The regression (7) can be written in matrix form as $y = Z\beta + \varepsilon$, where β is the $d \times 1$ vector of parameters and Z has d columns. The variance of the OLS estimate b of β

is $\text{var}(b) = s^2(Z'Z)^{-1}$, and the first estimation criterion is the (size-adjusted) determinant of this variance, $\text{varb} = (\det(\text{var}(b)))^{1/d}$. The determinant is raised to the power $(1/d)$ to remove size effects, as the determinant is the product of the d eigenvalues of $\text{var}(b)$. The other two estimation criteria are the two components of the variance, that is, the residual variance s^2 and the predictor contribution $\text{pdet} = (\det(Z'Z)^{-1})^{1/d}$ to the variance.

Another criterion is related to the fact that the aim of principal components is to maximize the 'variance accounted for' (VAF), that is, the amount of variance of the original predictor variables that is captured by the factors. The VAF is defined as the sum, over all predictor variables, of the explained sum of squares obtained by regressing each predictor variable on the constructed set of factors. As criterion we consider the relative VAF, that is, the multivariate R-squared defined by the total explained sum of squares divided by the total sum of squares of all predictors. Further, we use criteria to evaluate the accuracy of the constructed factors, both at the forecast time T and on the estimation interval $[1, T-h]$. We define the 'forecast R-squared' R_F^2 at time T by the squared correlation between the $k(q+1) \times 1$ vector of true factors and their lags $(f_T, \ldots, f_{T-q})$ and the corresponding vector of estimated factors and their lags. Further, let F be the $k \times (T-h)$ matrix of true factors f over the period $[1, T-h]$, and let $\hat{F}$ be the corresponding matrix of estimated factors. We define the 'estimation R-squared' R_E^2 as the total explained sum of squares of the regression of each estimated factor (column of $\hat{F}$) on the set of true factors (all columns of F), divided by the total sum of squares of all elements of $\hat{F}$.

<< **TABLE 4 to be inserted around here.** >>

Table 4 shows some characteristics of the linear response surfaces of the six discussed criteria.

The reported mean value is the relative performance of MPCR with respect to PCR, averaged over the set of considered simulation designs. For simplicity, we report only the sign of the coefficients that are significant at the 0.01% level, using robust standard errors. The estimation criteria show that MPCR achieves a substantial reduction, of 40.1% on average, in the variance of the estimated coefficients of the forecast equation (7). Most of this gain is due to a reduction (of 38.2%) in the contribution of the factors to this variance. This gain is larger for larger values of $(h, N, k, q, R_f^2, \rho, c, \pi)$ and for smaller values of (T, γ). This reflects the fact that MPCR differs relatively more from PCR for larger (h, k, q) and smaller T. The factor criteria show that MPCR has a slightly better (3.5%) 'variance accounted for', slightly better (3.4%) forecast factors, and somewhat worse (8.3%) factors on the estimation interval. This loss is due to the fact that the factor loadings are time-invariant in 50% of the designs, and PCR uses h more observations than MPCR in estimating the factors. Further, the case of time varying loadings does not help MPCR, because of the negative coefficient of c in the response surface for R_E^2.

4 EMPIRICAL APPLICATION

4.1 Data and forecast design

We use the data set of Stock and Watson (2002a). Here we mention only the most relevant aspects, and we refer to Stock and Watson (2002a) for further practical aspects, for instance, on data vintages, data transformations, and the treatment of outliers.

We apply PCR and MPCR to forecast eight macroeconomic variables. Four of these variables are real, that is, in the notation of Stock and Watson (2002a, Appendix B): industrial production (ip), personal income (gmyxpq), manufacturing and trade sales (msmtq), and nonagricultural employment (lpnag). The other four variables are prices: the consumer price index (punew), the consumer price deflator (gmdc), the consumer price index excluding food and energy (puxx), and the producer price index (pwfsa). The forecasts are based on diffusion index models using a set of $N = 146$ macroeconomic predictor variables (their 'balanced panel'). Monthly observations are available over the period 1959:01 till 1998:12, with missing values for some of the variables in the first two months. Therefore, the data are considered over the interval 1959:03 to 1998:12, giving a total of 478 observations.

The models are estimated, selected, and used in simulated out-of-sample forecasting, as follows. The considered forecast horizons are $h = 6$, 12, and 24 months. For a given time instant T, forecasts of the h-period average $\hat{y}^h_{T+h}$ are constructed using the DI, DI-AR and DI-AR-Lag models of equation (1), where the number of factors k, the number of lagged factors $(m - 1)$, and the number of autoregressive terms p are selected using BIC, as discussed in Section 2.1. Following Stock and Watson (2002a), we take $K = 4$, $M = 3$, and $P = 6$ for DI-AR-Lag, $K = 12$ and $P = 6$ for DI-AR, and $K = 12$ for DI. At time T, the time interval used to construct the principal components in PCR runs from 1959:03 to T, and the forecast model (1) is estimated over the sample period running from 1960:01 to $T - h$. In MPCR, the factors are constructed from the predictors on the interval running from 1959:11 (as $M = 3$) to $T - h$. The forecast procedure is applied sequentially, starting at 1970:01 and running until 1998:12 $- h$, and the forecast quality is evaluated by means of the mean squared forecast error (MSE) of the resulting $348 - h$ forecasts.

4.2 Forecast results

Table 5 reports the percentage gains in MSE of MPCR as compared to PCR, which is summarized by means of two boxplots in Figure 2. MPCR achieves positive gains in the majority of cases, and on average the gain is larger for the real variables than for the price variables. The gains are, in general, larger for longer forecast horizons, which is in line with the fact that the modification of MPCR becomes more substantial for longer horizons.

<< **TABLE 5 and FIGURE 2 to be inserted around here.** >>

The main results for the real variables are as follows. Averaged over these four variables, the gains for horizon $h = 24$ are around 5% for the DI-AR-Lag model and around 10% for the DI-AR and DI models. Note that the maximum number of factors is $K = 4$ for DI-AR-Lag, whereas $K = 12$ for the other two models. This distinction may partly explain the relatively smaller differences between the two methods for DI-AR-Lag. The average gain ranges for one-year ahead prediction ($h = 12$) between 3.5% and 7.4%, and for half-a-year ahead prediction ($h = 6$) between 5.9% and 9.3%. In some cases, larger gains are obtained, up to more than 15%. As compared to PCR, MPCR gives a reduction in MSE for thirty-five of the thirty-six real cases, and in a single case there is an increase in MSE (of 2.3%). Overall, MPCR clearly provides better forecasts than PCR for the four real variables.

The gains are considerably smaller for the four price series, around 1% in most cases. In some cases, the gain is substantial, up to 14.1%, but in other cases the MSE increases by up to 5.6%. We mention two possible causes for these results. First, the DI model is not appropriate for the price series, as it neglects the considerable amount of autocorrelation that is present in these series. Stock and Watson (2002a, Tables 3 and 4) report losses in MSE for PCR (as compared to the AR benchmark) for all four price series, ranging from a loss of 30% for the consumer price index up to a loss of 144% for the producer price index. Second, the AR benchmark performs relatively well for the price series, so that the gains of the DI-AR and DI-AR-Lag models are relatively small, see Stock and Watson (2002a, Tables 1 to 4). Therefore, the method to construct the factors is less important for these series.

We mention some further results. The reduction of the forecast MSE is largely due to a smaller forecast variance, whereas the bias is not much affected. For instance, MPCR has a smaller forecast variance than PCR for all thirty-six forecasts for the real variables, with an average reduction of around 10%. This result supports the motivation for MPCR, namely, to reduce the forecast variance by increasing the factor variance over the estimation interval.

We performed the test of Diebold and Mariano (1995), with robust standard errors, to examine whether MPCR provides a significantly lower MSE than PCR. For the thirty-six series of forecasts of the real variables, seventeen are significantly better at the 10% level and eight at the 5% level. For the thirty-six series of forecasts of the price variables, seven are significantly better at the 10% level and four at the 5% level. Of the twelve cases (out of seventy-two) where PCR has a smaller MSE than MPCR, none is significant at the 10% level. Summarizing the results for the considered empirical data, MPCR is significantly better in forecasting than PCR in some cases, it is better in the far majority of cases, and it is never significantly worse.

We also compared the number of factors k, factor lags $(m-1)$ and autoregressive lags $(p-1)$ of the forecast model (1) that are selected by BIC. As was mentioned before, the AR terms are much more important for the price series than for the real series, with an overall average of

$(p-1)$ of around 5 for the price series and below 0 for the real series. The models selected for PCR and MPCR hardly differ, although MPCR tends to select slightly smaller values for k, m, and p. This means that the improved forecast performance of MPCR is not due to differences in the forecast model selected by BIC, but to differences in the constructed factors.

4.3 Comparison of factor spaces

In comparing the factor spaces constructed by PCR and MPCR, we focus on the use of the factor components in estimating the forecast model (1). For simplicity we consider the model DI, that is, with $m = 1$ (no lagged factors) and with $p = 0$ (no autoregressive terms). At time T, the forecast model (1) is estimated using the factors over the period $[T_0, T - h]$, where the initial time T_0 is 1960:01 in our application. Let $T_e = T - h - T_0 + 1$ and let F and F_m denote the corresponding $T_e \times k$ factor matrices of PCR and MPCR respectively. We compare these two factor matrices by means of some of the criteria discussed in Section 3.4.

MPCR has a larger variance accounted for (VAF) than PCR for each time T, with average gains of about .5% for horizon $h = 6$ months, 1% for $h = 12$ and 2% for $h = 24$. So MPCR performs consistently better in this respect, although the differences are relatively small.

The standard errors of the regression coefficients in the DI model are proportional to the inverse of the $(k \times k)$ matrices $F'F$ (for PCR) and $F_m'F_m$ (for MPCR). As measured by the determinant of these inverse matrices, the gains of MPCR as compared to PCR increase consistently for longer forecast horizon and for larger number of factors. This finding is in line with the fact that MPCR differs more from PCR in these cases. Even for a single factor $(k = 1)$, the differences are substantial, with gains in the determinant (that is, in this case, a reduction of the estimation variance) of 12% for $h = 6$, 15% for $h = 12$ and 20% for $h = 24$.

5 CONCLUSION

In this article, we proposed an improved method for the construction of principal components in forecasting with diffusion index models. The forecast quality of such models is affected by the predictor variance on the estimation interval, and our method maximizes this variance. Simulation experiments and an empirical application to eight macroeconomic variables both indicate that this modification leads, in general, to better forecasts. The simulations show that the gains are larger in situations with larger forecast horizon, smaller observation interval, more predictor variables, more latent factors and factor lags, more time variation in the factor loadings, and more irrelevant predictors. In the empirical application, the forecast gains are most notable for the real variables, with a reduction of the mean squared forecast error by about 5% to 10%, whereas this gain is rather small for the price variables.

As topics for further research, we are interested in developing alternative methods to construct the diffusion indexes. The methods considered in this article are two-step methods, as the diffusion indexes are constructed without taking the forecast purpose into account. As the quality of the models is evaluated in terms of their forecast accuracy, it could pay to take this criterion explicitly into account in constructing the indexes. Another point of interest is the employed model selection method. Although BIC turns out to work quite well in empirical comparisons, this criterion is not directly related to the purpose of forecasting and forecast-based selection criteria might give better results.

6 Appendix

In this appendix, we derive the approximation $\alpha \approx \sqrt{hR_f^2/(1+hR_f^2)}$ mentioned at the end of Section 3.1 to stabilize the predictability for different horizons of data generated by (3-7). We assume for simplicity that $q = 0$ in (7). Define the $k \times 1$ vector $\beta = (1, \ldots, 1)'$, then $||\beta||^2 = k$. It follows from (5) and (7) that

$$\begin{aligned} y_{t+i} &= \beta'\alpha^i f_t + \beta' \sum_{j=0}^{i-1} \alpha^j u_{t+i-j} + \varepsilon_{t+i}, \\ \sum_{i=1}^{h} y_{t+i} &= \sum_{i=1}^{h} \alpha^i \beta' f_t + \beta' U_{t+h} + \sum_{i=1}^{h} \varepsilon_{t+i}, \end{aligned}$$

where $U_{t+h} = (\sum_{i=1}^{h-1} \alpha^i)u_{t+1} + (\sum_{i=1}^{h-2} \alpha^i)u_{t+2} + \ldots + u_{t+h}$. The explained variance of $\sum_{i=1}^{h} y_{t+i}$ (using the optimal prediction $\sum_{i=1}^{h} \alpha^i \beta' f_t$) is

$$(\sum_{i=1}^{h} \alpha^i)^2 ||\beta||^2/(1-\alpha^2) = k\alpha^2 (\sum_{i=0}^{h-1} \alpha^i)^2/(1-\alpha^2),$$

and the error variance (of $\beta' U_{t+h} + \sum_{i=1}^{h} \varepsilon_{t+i}$) is $h + k\sum_{j=1}^{h}(\sum_{i=0}^{j-1} \alpha^i)^2$. Define $S_j = (\sum_{i=0}^{j-1} \alpha^i)^2 = (1-\alpha^j)^2/(1-\alpha)^2$, then the above results imply that the forecast R^2 for the predicted h-period average $y_{t+h}^h = \frac{1}{h}\sum_{i=1}^{h} y_{t+i}$ is

$$R_f^2 = \frac{k\alpha^2 S_h}{k\alpha^2 S_h + h(1-\alpha^2) + k(1-\alpha^2)\sum_{j=1}^{h} S_j}.$$

We now make various rough approximations to derive an approximate expression of α in terms of h and R_f^2. Here, we use that for sufficiently large h there holds $S_h \approx 1/(1-\alpha)^2$ and $(1-\alpha)^2 \sum_{j=1}^{h} S_j = \sum_{j=1}^{h}(1-\alpha^j)^2 \approx \int_0^h (1-\alpha^x)^2 dx = \int_0^h (1+\alpha^{2x}-2\alpha^x)dx \approx h - \frac{1}{2\log(\alpha)} + \frac{2}{\log(\alpha)} = h + \frac{3}{2\log(\alpha)}$, as $0 < \alpha < 1$. Combining these results, we get for k and h sufficiently large the approximation

$$R_f^2 \approx \frac{k\alpha^2}{k\alpha^2 + h(1-\alpha^2)(1-\alpha)^2 + kh(1-\alpha^2) + \frac{3k(1-\alpha^2)}{2\log(\alpha)}} \approx \frac{\alpha^2}{h(1-\alpha^2)},$$

where we neglected all terms in the denominator except the one with the product kh (note that $(1-\alpha^2)/\log(\alpha) \to -2$ for $\alpha \uparrow 1$, so that the last term in the denominator is bounded). Solving this for α in terms of R_f^2, we get $\alpha \approx \sqrt{hR_f^2/(1+hR_f^2)}$. The simulation results in Section 3.3 show that, with this choice of α, the value of R_f^2 does not have a significant effect on the relative MSE of MPCR as compared to PCR, see Tables 1 and 4. This result is an indication of the appropriateness of this choice of α as function of the horizon h.

References

[1] Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2-nd ed. New York: John Wiley.

[2] Boivin, J., and Ng, S. (2006), "Are more data always better for factor analysis?," *Journal of Econometrics*, 132, 169-194.

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

[4] Jolliffe, I.T. (2002), *Principal Component Analysis*, 2-nd ed. Berlin: Springer.

[5] Lin, J.L., and Tsay, R.S. (2005), "Comparisons of Forecasting Methods with Many Predictors," *Working Paper.*

[6] Ng, S., and Perron, P. (2005), "A Note on the Selection of Time Series Models," *Oxford Bulletin of Economics and Statistics*, 67, 115-134.

[7] Stock, J.H., and Watson, M.W. (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293-335.

[8] Stock, J.H., and Watson, M.W. (2002a), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147-162.

[9] Stock, J.H., and Watson, M.W. (2002b), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

[10] Stock, J.H., and Watson, M.W. (2005), "An Empirical Comparison of Methods for Forecasting Using Many Predictors," *Working Paper.*

[11] Stock, J.H., and Watson, M.W. (in press), "Forecasting with Many Predictors," in *Handbook of Economic Forecasting*, eds. G. Elliott, C.W.J. Granger and A. Timmermann, Amsterdam: North-Holland.

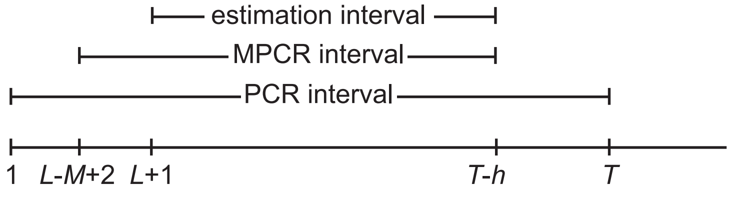

Figure 1: DATA WINDOWS Time intervals for estimation and for the construction of the diffusion indexes by means of PCR and MPCR; T is current time, h is the forecast horizon, $M-1$ is the maximal lag of the diffusion indexes in the forecast model, and L is the maximal lag of the diffusion indexes and of the autoregressive terms in the forecast model.

Table 1: AVERAGE MSE PER DESIGN PARAMETER Marginal average forecast MSE of MPCR relative to PCR (in percentages), for each design parameter and over three design sets: 'all' ($T \leq 100$, 13440 designs with average MSE 77.2; the table shows the number of designs in this set for each fixed parameter value), 'simple' ($R_f^2 = 0.5$ and $\rho = \gamma = c = \pi = 0$, 472 designs with average MSE 89.8), and 'complex' ($R_f^2 = 0.5$, $\rho = \gamma = 0.45$, $c = 10$ and $\pi = 0.25$, 472 designs with average MSE 76.4).

		#designs	MSE					#designs	MSE		
		all	all	simple	complex			all	all	simple	complex
h	5	4608	73.5	87.1	72.7	R_f^2	.1	4480	78.1	·	·
	10	4032	74.9	88.6	71.5		.5	4480	76.2	89.8	76.4
	25	3072	80.6	90.7	78.3		.9	4480	77.2	·	·
	50	1728	86.3	95.1	87.6	(ρ, γ)	(0,0)	3360	74.4	89.8	·
T	25	1344	56.8	64.0	43.9		(.9,0)	3360	74.4	·	·
	50	4224	69.1	77.1	58.4		(0,.9)	3360	83.2	·	·
	100	7872	85.0	91.3	77.0		(.45,.45)	3360	76.7	·	76.4
	250	0	·	98.1	88.8	c	0	6720	82.6	89.8	·
N	50	3360	82.8	91.9	82.3		10	6720	71.7	·	76.4
	100	3360	77.8	90.1	77.1	π	0	6720	79.7	89.8	·
	250	3360	75.2	88.9	73.9		.25	6720	74.6	·	76.4
	500	3360	73.9	88.3	72.1						
k	5	4992	79.0	90.8	74.2						
	10	4224	78.2	90.0	76.0						
	20	2880	76.4	89.3	78.0						
	40	1344	69.0	88.1	78.9						
q	0	5760	82.8	91.6	80.7						
	1	4416	74.0	88.5	74.1						
	2	3264	71.5	88.8	73.0						

Table 2: STATISTICS OF DESIGNS AND MSE Mean values of design parameters (h, T, k, q) and statistics (mean, median, maximum, minimum and standard deviation) of MSE (in percentages) for nine designs. The sets of 'all', 'simple' and 'complex' designs are each considered for three cases: all designs in the set, designs where MPCR performs worse than PCR (MSE > 100), and designs where MPCR performs at least twice as well as PCR (MSE < 50).

				mean				MSE				
MSE	design	#designs	%designs	h	T	k	q	mean	med	max	min	std
all	all	13440	100	17	77	13	0.8	77.2	83.1	143.6	1.7	20.3
	simple	472	100	19	147	16	0.9	89.8	95.5	101.2	11.8	14.3
	complex	472	100	19	147	16	0.9	76.4	81.6	106.0	9.1	20.4
> 100	all	226	1.7	12	99	10	1.0	102.3	100.9	143.6	100.0	4.3
	simple	39	8.3	10	238	11	1.2	100.3	100.1	101.2	100.0	0.4
	complex	8	1.7	50	175	28	0.3	102.3	101.3	106.0	100.2	2.4
< 50	all	1593	11.9	9	52	17	1.2	34.8	36.7	50.0	1.7	10.9
	simple	13	2.8	6	44	22	1.0	36.1	41.5	47.5	11.8	10.6
	complex	53	11.2	8	53	15	1.3	32.6	34.9	49.6	9.1	11.3

Table 3: AVERAGE MSE PER (h, T, N) DESIGN Marginal average forecast MSE for each (h, T, N) design. Some cells are empty: for $h \geq T$ as forecasting is not possible, and for some designs with $T = 250$ and $N \geq 250$ due to long computation times.

		all				simple				complex			
N	h	$T = 25$	50	100	250	$T = 25$	50	100	250	$T = 25$	50	100	250
50	5	58.2	71.2	92.1	99.2	58.1	79.8	95.4	100.2	61.0	58.8	86.3	97.7
	10	62.2	75.9	87.9	97.7	66.2	82.6	94.3	99.3	51.9	62.0	77.2	95.4
	25	·	76.5	88.8	95.6	·	80.4	94.2	98.4	·	76.0	79.4	89.5
	50	·	·	89.7	96.0	·	·	94.2	98.9	·	·	91.0	91.2
100	5	55.1	66.5	87.0	97.3	65.6	75.2	92.6	99.5	35.0	52.4	78.0	95.1
	10	59.7	72.6	81.9	94.8	62.7	81.4	90.2	98.1	48.9	57.0	67.5	91.7
	25	·	73.4	85.8	92.9	·	74.8	91.9	97.4	·	77.0	75.0	85.4
	50	·	·	87.2	94.2	·	·	92.3	98.5	·	·	89.5	87.8
250	5	54.0	62.6	84.4	·	63.8	72.8	92.0	98.8	35.3	48.1	74.7	92.9
	10	57.4	69.2	78.8	·	66.0	78.8	88.1	97.3	48.6	51.5	63.9	89.1
	25	·	69.7	84.6	·	·	76.6	89.7	97.0	·	71.6	72.4	81.4
	50	·	·	85.1	·	·	·	89.7	97.2	·	·	89.9	83.1
500	5	53.6	61.4	83.1	·	64.4	71.5	90.4	98.6	32.4	45.6	72.6	92.5
	10	56.0	68.2	77.4	·	66.3	79.1	86.9	97.1	41.6	49.0	61.8	87.0
	25	·	68.2	83.9	·	·	74.7	88.9	96.6	·	68.4	71.0	80.2
	50	·	·	83.1	·	·	·	89.3	97.0	·	·	89.4	81.5

Table 4: RESPONSE SURFACES Performance criteria (columns, MPCR as percentage of PCR) and simulation design (rows) are related by linear regression (except a quadratic one for MSE, with quadratic terms shown in column '(qdr)'). The row 'MPCR' shows which sign of the coefficients corresponds to better performance of MPCR. The row 'mean' shows the percentage average score for MPCR as compared to PCR for the MSE and for the other six performance criteria. The row '$100R^2$' shows the R-squared (multiplied by 100) of the regression of the corresponding response surface. Table values + (-, 0) stand for positive (negative, insignificant) coefficients.

criterion							estimation			factors		
		MSE					varb	s^2	pdet	VAF	R_F^2	R_E^2
design		all	simple	complex	all		all	all	all	all	all	all
surface		lin	lin	lin	qdr		lin	lin	lin	lin	lin	lin
MPCR		-	-	-	-	(-)	-	-	-	+	+	+
mean		77.2	89.8	76.4	77.2		59.6	95.6	61.8	103.5	103.4	91.7
$100R^2$		60.3	45.1	47.8	76.8		88.1	32.6	88.4	53.1	21.6	77.7
1	(qdr)	70.27	82.26	64.49								
$h/10$	(h^2)	-0.37	0	1.82	-	(+)	-	-	-	+	0	-
$T/10$	(T^2)	4.71	1.09	1.46	+	(-)	+	+	+	-	-	+
$N/10$	(N^2)	-0.15	0	-0.19	-	(+)	-	0	-	-	-	0
$k/10$	(k^2)	-8.30	-3.09	-1.97	-	(-)	-	+	-	-	+	+
q	(q^2)	-10.45	-3.51	-6.42	-	(0)	-	+	-	+	+	-
R_f^2	(hN)	0			-	(0)	-	0	-	+	+	-
ρ	(hT)	0			0	(0)	-	-	-	-	-	-
γ	(NT)	8.95			+	(0)	+	0	+	-	-	+
c	(hk)	-1.09			-	(+)	-	-	-	-	-	-
π	(hq)	-20.24			-	(-)	-	-	-	+	0	-
	(Tk)					(+)						
	(Tq)					(+)						
	(kq)					(-)						

Table 5: EMPIRICAL FORECAST GAINS   Percentage gains in MSE of MPCR as compared to PCR for eight economic variables (four real variables and four price variables), with averages (columns ‘av.’).

	Real variables					Price variables				
Model	ip	gmyxpq	msmtq	lpnag	av.	punew	gmdc	puxx	pwfsa	av.
Horizon $h = 6$										
DI-AR-Lag	6.41	4.08	1.90	11.38	5.94	-0.06	0.97	1.03	2.22	1.04
DI-AR	9.35	2.27	9.22	16.54	9.34	-0.06	1.95	1.03	0.47	0.85
DI	9.37	2.28	1.02	15.65	7.08	-1.02	1.30	0.78	3.19	1.06
Horizon $h = 12$										
DI-AR-Lag	2.70	6.63	3.08	1.63	3.51	-2.53	3.34	6.99	-5.36	0.61
DI-AR	5.41	11.42	4.87	8.08	7.44	-4.22	2.46	6.58	-1.62	0.80
DI	2.02	11.42	0.23	5.59	4.81	1.68	-1.72	1.50	1.71	0.79
Horizon $h = 24$										
DI-AR-Lag	5.97	2.00	13.19	-2.30	4.71	-5.57	-4.29	14.10	1.12	1.34
DI-AR	8.03	8.01	15.94	7.26	9.81	1.98	3.04	12.35	5.25	5.66
DI	9.83	8.01	15.75	7.94	10.38	-0.59	0.20	2.80	2.98	1.34

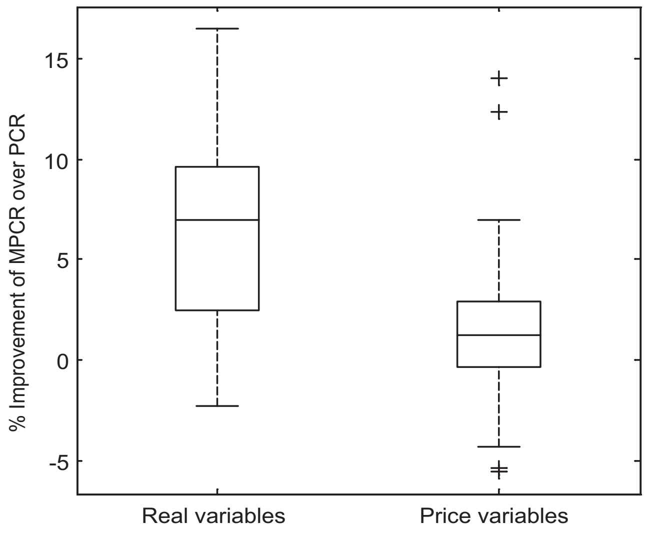

Figure 2: EMPIRICAL FORECAST GAINS   Boxplots of the percentage gain in MSE of MPCR as compared to PCR, for real variables and for price variables; each boxplot contains the thirty-six (real or price) MSE values in Table 5.