Forecast comparison of principal component regression and principal covariate regression

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

Econometric Institute, Erasmus University Rotterdam

Econometric Institute Report EI 2005 - 28

Abstract

Forecasting with many predictors is of interest, for instance, in macroeconomics and finance. This paper compares two methods for dealing with many predictors, that is, principal component regression (PCR) and principal covariate regression (PCovR). The forecast performance of these methods is compared by simulating data from factor models and from regression models. The simulations show that, in general, PCR performs better for the first type of data and PCovR performs better for the second type of data. The simulations also clarify the effect of the choice of the PCovR weight on the forecast quality.

Key words: principal components, principal covariates, regression model, factor model, economic forecasting

1 Introduction

In many forecasting applications in macroeconomics and finance, a large number of predictor variables are available that may all help to forecast the variable of interest. In such situations, one should somehow compress the predictor information. For instance, if T observations are available for a set of k predictors, then for $k > T$ it is simply impossible to estimate a multiple regression model that includes all predictors as separate regressors. If $k \leq T$ but k is large, then it is still not advisable to estimate a regression model with all predictors as regressors because the resulting forecasts will have large variance. The forecasts may improve if the information in the predictors is somehow compressed and a forecast equation containing fewer predictors is used.

[*] corresponding author, email address: heij@few.eur.nl

Several methods for forecasting with many predictors have been proposed in the literature. We refer to Stock and Watson (2004) for a survey. For example, in 'Principal Component Regression' (PCR) the predictor information is first summarized by a (small) number of principal components, which are then used as prediction factors in a low-dimensional multiple regression model. This approach is followed, for instance, by Stock and Watson (1999, 2002a,b) within the context of dynamic factor models to forecast key macroeconomic variables like production and inflation from large sets of economic and financial predictor variables. An essential aspect of PCR and similar methods is that they consist of two stages, as first the factors are constructed and then the forecast equation is estimated. The resulting factors need not necessarily be the ones that forecast best, as the construction of the factors in the first stage is not directly related to their use in forecasting in the second stage.

In this paper, we consider an alternative method that combines the two stages of predictor compression and forecasting in a single criterion. This method, called Principal Covariate Regression (PCovR), was proposed by De Jong and Kiers (1992). In contrast to PCR, PCovR is a data-based method that does not employ an explicit underlying statistical model. As the construction of the PCovR factors is directly related to their use in forecasting, this may give better forecasts as compared to two-step methods like PCR.

We compare the forecast performance of PCR and PCovR by means of simulation experiments. We investigate various factors that may affect the forecast performance, including the number of predictors and the correlation of the predictors with the variable to be predicted. The forecast quality is evaluated by means of the root mean squared (one-step-ahead, out of sample) forecast error (RMSE). As the choice of the number of factors is of special interest, we consider the RMSE obtained by applying information-based selection criteria, as in Stock and Watson (1999).

The remainder of this paper is organized as follows. In Section 2, we formulate the forecasting problem with compressed predictors in more detail and we describe the methods PCR and PCovR. Section 3 discusses the general set-up of the simulation experiments, and Section 4 describes the forecast performance of PCR and PCovR for data generated by factor models and regression models. Section 5 concludes with a brief overview and with some suggestions for further research.

2 Forecasting with compressed predictors

2.1 The forecast model

First we introduce some notation. The observations consist of time series of length T on a variable to be predicted (y) and on a set of predictor variables (X). Let k be the number of predictors, then y is a $T \times 1$ vector and X is a $T \times k$ matrix. The idea is to compress the information in the k variables X by means of p factors F, with p (much) smaller than k. Here F is a $T \times p$ matrix consisting of linear combinations of the X variables, so that

$$F = XA$$

for some $k \times p$ matrix A. These factors are used to forecast y by means of a linear regression model. The (one-step-ahead, conditional) forecast equation for y_{T+1} at period T is written as

$$\hat{y}_{T+1} = \alpha + f_{T+1}\beta = \alpha + x_{T+1}A\beta. \tag{1}$$

Here α is the constant term, β is a $p \times 1$ vector, and $f_{T+1} = x_{T+1}A$ where x_{T+1} is the $1 \times k$ vector of values of the predictors at time $T+1$. The forecast of y_{T+1} is conditional, as it depends on x_{T+1} which is assumed to be given. The (multi) h-step-ahead forecast equation, with $h > 1$, has the same structure, replacing $\hat{y}_{T+1}$ and x_{T+1} in (1) respectively by $\hat{y}_{T+h}$ and x_{T+h}. We restrict our attention to $h = 1$ in this paper.

To apply this model in practice, we should estimate the number of factors p and the parameters (A, α, β) of the forecast equation (1). The next two subsections describe two methods to construct the factors F and to estimate (1) for given value of p, that is, principal component regression and principal covariate regression. In Section 4, we evaluate the forecast performance if p is selected by means of an information criterion.

2.2 Principal component regression (PCR)

The method of principal component regression (PCR) consists of two estimation steps. In the first step, A is estimated by means of principal components. That is, the p factors are obtained by minimizing the squared Frobenius norm $||X - \hat{X}||^2$ under the restriction that $\hat{X}$ has rank p. The squared Frobenius norm of a matrix is simply the sum of squares of all elements of the matrix. The X-variables should be standardized to prevent scale effects. For instance, each column (variable) of X is scaled to have zero mean and unit norm. The estimates A can then be obtained from the singular value decomposition (SVD) of X.

For later purposes, it is helpful to describe this first step of PCR in more detail. Let $X = USV'$ be an SVD of X where the singular values in the matrix S are listed in decreasing order. Then $\hat{X} = U_p S_p V_p'$ where U_p and V_p consist respectively of the first p columns of U and V and where S_p is the $p \times p$ diagonal matrix with the p largest singular values of X on the diagonal. If we define $W = V_p V_p'$, then it is easily checked that $\hat{X} = XW = XAB$ for any $k \times p$ matrix A and $p \times k$ matrix B such that $AB = W$. For instance, if we take

$$A = V_p S_p^{-1}$$

and $B = S_p V_p'$, it follows that the factors $F = XA$ satisfy

$$F'F = A'X'XA = I_p$$

so that the p factors in F are scaled and mutually orthogonal. The factors F constructed in this way are the (first p) principal components of X. So, in PCR the parameter matrix A is estimated by minimizing

$$f_X(A, B) = ||X - XAB||^2. \tag{2}$$

In the second step, the parameters α and β in (1) are estimated by ordinary least squares (OLS), for given values of A. Let $F = XA$, then the second step corresponds to minimizing

$$f_y(\alpha, \beta) = ||y - \alpha - F\beta||^2 = ||y - \alpha - XA\beta||^2. \tag{3}$$

Summarizing, PCR consists of an SVD for (2) followed by OLS in (3). The next subsection discusses a method that integrates these two steps by minimizing a single criterion function.

2.3 Principal covariate regression (PCovR)

Principal covariate regression (PCovR) combines the two stages of compressing the predictors and estimating the parameters of the forecast equation by optimizing a single criterion. This method was proposed by De Jong and Kiers (1992). In PCovR, the parameters (A, B, α, β) are estimated simultaneously by minimizing a weighted average of the forecast errors (3) and the predictor compression errors (2). For given weights $w_1 > 0$ and $w_2 > 0$ and for given number of factors p, the criterion to be minimized is

$$f(A, B, \alpha, \beta) = w_1||y - \alpha - XA\beta||^2 + w_2||X - XAB||^2. \tag{4}$$

Here the $T \times p$ matrix $F = XA$ consists of p factors that compress the predictor information in the $T \times k$ matrix X, as p is always chosen to be (much) smaller than $\mathrm{rank}(X)$, and hence also (much) smaller than k. As the choice of the factors F is based partly on their quality in fitting y, this may lead to better forecasts as compared to two-step methods like PCR. As before, A is a $k \times p$ matrix of rank p, B a $p \times k$ matrix, α is a scalar and β a $p \times 1$ vector. Clearly, if (A, B, α, β) is an optimal set of coefficients then $(AR, R^{-1}B, \alpha, R^{-1}\beta)$ is also optimal for every invertible $p \times p$ matrix R. Therefore, A may be chosen such that $F'F = A'X'XA = I_p$. With this restriction, the parameters are identified up to an orthogonal transformation R, that is, with $R'R = I_p$.

The vector norm in (4) is the Euclidean norm and the matrix norm is the Frobenius norm. To prevent scaling effects of the variables y and X, and because only the relative weight w_1/w_2 is of importance, we will consider weights of the form

$$w_1 = \frac{w}{||y||^2}, \qquad w_2 = \frac{1-w}{||X||^2}, \tag{5}$$

with $0 \leq w \leq 1$. The user has to choose the weight w, balancing the objectives of good predictor compression for X (for w small) and good (in-sample) fit for y (for w large). The parameter w should be chosen between 0 and 1, because otherwise the criterion (4) becomes unbounded from below and has no optimal solution. The limiting case where the weight w approaches zero gives PCR, and if w approaches one then this gives OLS.

The minimization of (4) is a nonlinear—in fact, bilinear—optimization problem, because of the product terms $A\beta$ and AB. The optimal estimates of (A, B, α, β) can be computed by means of two SVD's, as explained in the Appendix.

3 Design of the simulation experiments

3.1 Data generating process

In Section 4, we will compare the forecast performance of PCR and PCovR by means of various simulation experiments. The specification of the data generating process (DGP) and of the employed forecast model are varied to investigate the forecast performance under different conditions. Therefore, we discuss in this section the general set-up of the experiments.

The three simulation experiments in Section 4 can all be seen as instances of dynamic factor models. For more background on this kind of models we refer to Stock and Watson (2002a), see also Boivin and Ng (2003). In a stationary dynamic factor model, the observed data (y, X) are related to unobserved underlying factors F that evolve dynamically over time. Let f_t be the $1 \times p$ vector with factor scores at times t, then a first-order model for the factors is given by $f_t = f_{t-1}\Phi + u_t$ where Φ is a $p \times p$ stable matrix and u_t is uncorrelated with f_{t-1}. The DGP is described by

$$y_t = f_t\beta + \varepsilon_t, \qquad x_t = f_t\Lambda + v_t, \qquad f_t = f_{t-1}\Phi + u_t. \tag{6}$$

We write the model in terms of row vectors of observations and error terms at time t, to conform with the notation that the t-th row x_t of the $T \times k$ matrix X contains the observations at time t for the k predictors. In (6), β is a $p \times 1$ vector and Λ is a $p \times k$ matrix of factor loadings.

The three experiments in Section 4 correspond to three different choices for the factor loading matrix Λ. Section 4.1 considers a 'factor DGP' with $p = 2$ factors, where each factor loads on a set of $k/2$ predictor variables. Section 4.2 considers a 'regression DGP' with $p = k$ factors, where each factor loads on a single predictor. Finally, Section 4.3 discusses a 'dyadic factor DGP' that lies in between the two foregoing cases. For each experiment, we consider different specifications for the number of predictors k, for the parameters β, and for the squared correlations ρ^2_{yf} and ρ^2_{xf}, that is, for the amount of information that the predictors X carry on the factors F and for the extent to which y can be predicted from F.

The variables in the simulations are normalized, as follows. The matrix Φ is diagonal, with coefficient $-1 < \phi < 1$ on the diagonal, and the error terms u_t are mutually independent white noise processes with mean zero and variance $(1 - \phi^2)$. Therefore, all p factors are mutually independent and the covariance matrix is $\text{var}(f_t) = I_p$. The errors $v_t = (v_{1,t}, \ldots, v_{k,t})$ are mutually independent white noise processes with mean zero and with variance $\sigma^2_{v_i}$, and v_t is independent of u_t and hence also of f_t. All predictors are normalized to have mean zero and variance one. Let x_i, λ_i and v_i denote respectively the i-th predictor variable, the i-th column of Λ and the i-th component of v, so that $x_{i,t} = f_t\lambda_i + v_{i,t}$. Then $\text{var}(x_{i,t}) = 1 = ||\lambda_i||^2 + \sigma^2_{v_i}$ and $\rho^2_{x_if} = ||\lambda_i||^2$, so that a desired squared correlation $\rho^2_{x_if}$ is achieved by scaling λ_i so that $||\lambda_i||^2 = \rho^2_{x_if}$ and by taking

$$\sigma^2_{v_i} = 1 - \rho^2_{x_if}.$$

The errors ε_t are white noise, independent from (u_t, v_t) and with mean zero and variance σ^2_ε. As $\rho^2_{yf} = ||\beta||^2/(||\beta||^2 + \sigma^2_\varepsilon)$, it follows that a desired squared correlation ρ^2_{yf} is achieved by taking

$$\sigma^2_\varepsilon = ||\beta||^2 \frac{1 - \rho^2_{yf}}{\rho^2_{yf}}. \tag{7}$$

3.2 Forecast models

The purpose is to forecast the dependent variable y one-step-ahead on the basis of observed past data on y and current and past data on a set of k predictors X. The forecast equation given in (1) is $\hat{y}_{T+1} = \alpha + f_{T+1}\beta$, and we compare different methods to estimate the factor $f_{T+1} = x_{T+1}A$ and the parameters (α, β) from observations on the dependent variable y (for times $t \leq T$) and on the k predictor variables X (for times $t \leq T+1$). The forecast of y_{T+1} is conditional, as x_{T+1} is assumed to be given. In Section 4, we will report detailed results obtained for simulations with $T = 100$, and results for $T = 400$ will be discussed in more general terms.

We consider forecast models with various possible values for the number of factors p. To condense the information, we do not report the results for all values of p. Instead, we employ information criteria to choose p and report the resulting outcomes. In the simulations, we used five criteria, that is, the Bayes information criterion (BIC), the Akaike information criterion, and three information criteria developed specifically for choosing the number of factors by Bai and Ng (2002). It turned out that BIC performs best in all simulations, as on average it provides more accurate forecasts with fewer factors as compared to the other four criteria. Therefore, in what follows we will only report the results obtained by BIC[1]. Let $\hat{y}_t = a + \hat{f}_t b$ be the fitted values of y and let $s_p^2 = ||y - \hat{y}||^2/T$ be the residual variance of y, obtained by using p factors. Then the number of factors p is selected by minimizing the BIC criterion

$$\text{BIC}(p) = \log(s_p^2) + (p+1)\frac{\log(T)}{T}.$$

Note that, although the PCR factors depend only on X and not on y, the BIC criterion depends on the fit for y, so that the (past) forecast quality plays a role in selecting the number of factors and the corresponding forecast model. For PCR, sometimes relatively large values of p are required because some of the DGP's of Section 4 do not have a parsimonious dynamic factor structure. For instance, the DGP in Section 4.3 has $p = 7$ factors if $k = 10$, $p = 31$ if $k = 40$, and $p = 63$ if $k = 100$. For PCR, the number of factors is therefore only restricted to $p \leq 0.8k$, so that the compression obtained by replacing X by F should be at least a modest 20%. For simplicity, only a grid of values for p is considered, that is, for $k = 100$ the considered values for p are $(1, 2, 3, 4, 5, 6, 7, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80)$, and for $k = 10$ and $k = 40$ the values $p \leq 0.8k$ from this grid are considered. So the total number of PCR forecast models is (at most) eighteen.

On the other hand, for PCovR the number of factors is always restricted

[1] The results for the other criteria are available upon request.

to $p \leq 3$. In a sense, all DGP's in Section 4 correspond to $p = 1$ relevant prediction factor, that is, $F\beta$, so that even $p = 1$ could be a reasonable choice for PCovR. We analyzed the consequences of allowing values larger than three for p. The far majority of PCovR models chosen by BIC have $p = 1$ or $p = 2$ factors, except in cases of severe overfitting, as will be discussed at the end of Section 4.1. For ease of comparison of PCovR with PCR, we also report the results for PCR if the number of factors is restricted to $p \leq 3$, and we indicate this method by PCR3.

To apply PCovR, we further have to specify the weight factor w in (4) and (5). We consider a grid of five values for w, that is, $(0.0001, 0.01, 0.1, 0.5, 0.9)$. For $w = 0.0001$, most weight is assigned to approximating X, in which case PCovR will be close to PCR. At the other extreme, for $w = 0.9$ most weight is assigned to fitting y. The consequences of the choice of the weight w on the forecast quality is discussed at the end of Section 4.1. In particular, it turns out that large weights w should be excluded in some situations in order to prevent overfitting.

3.3 Forecast evaluation

For each DGP, we perform one thousand simulation runs with $T = 100$. The data of each run are used to compute a single one-step-ahead forecast of y_{T+1}. As the number of factors p is chosen by BIC, this gives two PCR forecasts (one with p selected from a large grid and another with p selected from $\{1, 2, 3\}$) and five PCovR forecasts (one for each choice of the weight factor w). The forecast quality of the resulting seven methods is compared by the root mean squared forecast error (RMSE) of y_{T+1} over the thousand simulation runs. The RMSE is defined as

$$\text{RMSE}_j = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\frac{(y_{T+1,i} - \hat{y}_{T+1,ij})^2}{\sigma_\varepsilon^2}}, \tag{8}$$

where j denotes the employed forecast method, $y_{T+1,i}$ is the actual value of y at the forecast time $T + 1 = 101$ in the i-th simulation run, and $\hat{y}_{T+1,ij}$ is the value forecasted by method j in the i-th simulation run ($i = 1, \ldots, 1000$, $j = 1, \ldots, 7$). The squared forecast error $(y_{T+1,i} - \hat{y}_{T+1,ij})^2$ is divided by the error variance σ_ε^2, as this provides a natural benchmark for the forecast errors that would be obtained if the DGP (6) were estimated perfectly.

The variance of the dependent variable y depends on the DGP. Therefore, to facilitate the interpretation of the reported RMSE values, we also report the RMSE for the model-free 'zero-prediction' $\hat{y}_{T+1} = 0$, which is equal to the square root of $\text{var}(y_t)/\sigma_\varepsilon^2$. We call this the relative standard deviation of

y, denoted by rsd(y). The variance of y_t in (6) is $(||\beta||^2 + \sigma_\varepsilon^2)$, and with the expression in (7) for σ_ε^2 this gives

$$\text{rsd}(y) = \sqrt{\frac{\text{var}(y)}{\sigma_\varepsilon^2}} = \sqrt{\frac{||\beta||^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2}} = \sqrt{1 + \frac{\rho_{yf}^2}{1-\rho_{yf}^2}} = \sqrt{\frac{1}{1-\rho_{yf}^2}}.$$

4 Three simulation experiments

4.1 Simulation with factor DGP

In the first simulation experiment, the data are generated by a dynamic factor model with $p = 2$ factors. The parameters in the model (6) are chosen as follows. The 2×2 matrix Φ is a diagonal matrix with value $\phi = 0.7$ on the diagonal, and the errors u_t are mutually independent white noise processes with mean zero and variance 0.51, so that $\text{var}(f_t) = I_2$. For the case of k predictors, the first factor loads with coefficient 0.9 on the first $k/2$ predictors and the second factor loads with coefficient 0.6 on the remaining $k/2$ predictors. Let $1_{k/2}$ and $0_{k/2}$ denote the $1 \times (k/2)$ row vector with all elements equal to 1 and 0 respectively, then the $2 \times k$ factor loading matrix is

$$\Lambda = \begin{pmatrix} 0.9(1_{k/2}) & 0_{k/2} \\ 0_{k/2} & 0.6(1_{k/2}) \end{pmatrix}.$$

The predictors are normalized to have variance 1, so that the error variance of v in (6) is $\sigma_v^2 = 0.19$ for predictors loaded by the first factor and $\sigma_v^2 = 0.64$ for predictors loaded by the second factor. The corresponding squared correlations $\rho_{x_i f}^2$ between the predictors and the factors are respectively 0.81 and 0.36. The 2×1 parameter vector β is either $(1 \;\; 0)'$, $(0 \;\; 1)'$ or $(1 \;\; 1)'$, so that y depends respectively on the first factor, on the second factor, and on both factors. The number of predictors k is 10, 40 or 100. The squared correlation ρ_{yf}^2 between y_t and the factor f_t is either 0.1, 0.5 or 0.9.

With three options for the number k of predictors, three options for the parameter vector β and three options for the correlation between y and the factors, the total number of DGP's is 27. As the DGP's have $p = 2$ factors, we estimated PCR and PCovR models with $p = 1, 2,$ or 3 factors, giving in total 18 models. That is, we actually consider PCR3 with $p \leq 3$ instead of PCR, as allowing for more than three principal components does not improve the forecast performance for this DGP. For each of the six methods (PCR3 and PCovR with five possible weights), the number of factors p for each data set is selected by BIC.

Table 1 shows the RMSE's and the mean of the selected number p of factors over the thousand simulation runs. The reported values are rounded to two decimals. As the RMSE is measured relative to the best (DGP) predictor, all RMSE values are larger than 1 and measure the loss in forecast quality as compared to the optimal forecast based on exact knowledge of the DGP. A ** in Table 1 stands for RMSE values between 10 and 100 and *** stands for values between 100 and 1000.

We summarize the main results in Table 1. First we consider PCR3. This method performs rather well, as expected. On average, the RMSE decreases for larger k as more predictors contain information on the relevant prediction factor $f_t\beta$. Response related to the first factor (the case $\beta = (1 \;\; 0)'$) is easiest to forecast, and response related to the second factor (the case $\beta = (0 \;\; 1)'$) is the hardest. The RMSE mostly increases if ρ^2_{yf} increases, but the gain as compared to the zero-prediction also increases. Note that larger values of ρ^2_{yf} correspond to a smaller error variance σ^2_ε in (7), so that it becomes harder to get close to the 'perfect prediction' benchmark used in the definition of the RMSE in (8). Finally, the number of principal component factors is mostly close to $p = 2$ if $\rho^2_{yf} = 0.5$ or 0.9, but somewhat lower if $\rho^2_{yf} = 0.1$ and also if the response is related to the first factor (the case $\beta = (1 \;\; 0)'$).

Next we consider the results for the PCovR method. PCovR with weight $w = 0.0001$ gives results that are nearly identical to those of PCR3, as expected. For $k = 10$, the RMSE is comparable to that of PCR3 for all weights, but the number of factors for PCovR is consistently lower than for PCR3, and more distinctively so for larger weights. For $k = 40$, the RMSE is almost identical to that of PCR3 for $w = 0.0001$ and $w = 0.01$, but larger weights perform worse. For $k = 40$ and $\rho^2_{yf} = 0.1$ or 0.5, the average number of factors p is the largest for $w = 0.1$ (ranging between 2.6 and 3), the smallest for $w = 0.9$ (with $p = 1$ always), and roughly between 1 and 2 for $w = 0.0001$. An intuitive explanation of this pattern is that for $w \approx 1$ it suffices to construct a single factor that fits y well and that for $w \approx 0$ it is best to construct the two DGP factors to approximate X well. For more moderate values of the weight w, it may require three factors to achieve a good approximation of both y and X. However, for $k = 40$ and $\rho^2_{yf} = 0.9$, the average of p in most cases decreases if w increases, from $p \approx 2$ for $w \leq 0.1$ to $p = 1$ for $w = 0.9$.

We discuss three further issues of interest, that is, overfitting in the case of $k = 100$ predictors, the performance in larger samples with $T = 400$, and misspecification of the forecast model.

For $k = 100$, PCovR with weight $w = 0.01$ or larger does not perform acceptably anymore. This is due to *overfitting*, which can be explained intuitively, as follows. Consider a PCovR model with $p = 2$ factors based on $k = 100$ predictors to forecast y on the basis of $T = 100$ past observations. Generically,

Table 1. RMSE and mean number of factors p in simulation experiment with factor DGP

DGP				RMSE						mean p					
					PCovR with weight w						PCovR with weight w				
k	β'	ρ^2_{yf}	rsd(y)	PCR3	1E−4	0.01	0.10	0.50	0.90	PCR3	1E−4	0.01	0.10	0.50	0.90
10	(1 0)	0.1	1.05	*1.02*	*1.02*	*1.02*	*1.02*	1.03	1.07	1.09	1.10	1.10	1.14	1.16	1.00
10	(1 0)	0.5	1.41	*1.04*	*1.04*	*1.04*	*1.04*	1.05	1.08	1.36	1.35	1.34	1.25	1.00	1.00
10	(1 0)	0.9	3.16	*1.18*	1.19	1.19	*1.18*	1.20	1.22	1.72	1.74	1.72	1.54	1.02	1.00
10	(0 1)	0.1	1.05	*1.02*	*1.02*	*1.02*	*1.02*	*1.02*	1.03	1.62	1.62	1.63	1.68	1.73	1.00
10	(0 1)	0.5	1.41	*1.13*	*1.13*	*1.13*	*1.13*	1.15	1.16	2.04	2.04	2.03	2.02	1.42	1.00
10	(0 1)	0.9	3.16	*1.90*	*1.90*	*1.90*	*1.90*	1.95	1.96	2.08	2.08	2.07	2.02	1.29	1.00
10	(1 1)	0.1	1.05	*1.03*	*1.03*	*1.03*	1.04	1.04	1.06	1.42	1.42	1.42	1.49	1.43	1.00
10	(1 1)	0.5	1.41	*1.07*	*1.07*	*1.07*	*1.07*	1.09	1.10	1.99	1.99	1.99	1.95	1.27	1.00
10	(1 1)	0.9	3.16	*1.65*	*1.65*	*1.65*	*1.65*	1.69	1.71	2.06	2.06	2.05	1.98	1.45	1.00
40	(1 0)	0.1	1.05	1.03	*1.02*	1.03	1.27	1.33	1.33	1.07	1.07	1.11	2.99	2.28	1.00
40	(1 0)	0.5	1.41	1.03	1.03	*1.02*	1.12	1.21	1.26	1.23	1.23	1.25	2.58	1.83	1.00
40	(1 0)	0.9	3.16	*1.03*	1.04	1.04	1.06	1.22	1.35	1.62	1.61	1.61	1.53	1.06	1.00
40	(0 1)	0.1	1.05	*1.04*	*1.04*	*1.04*	1.25	1.28	1.32	1.78	1.78	1.85	3.00	1.78	1.00
40	(0 1)	0.5	1.41	*1.07*	*1.07*	*1.07*	1.18	1.26	1.33	2.03	2.03	2.06	2.78	1.39	1.00
40	(0 1)	0.9	3.16	*1.38*	*1.38*	*1.38*	1.44	1.67	1.76	2.05	2.05	2.05	2.06	1.47	1.00
40	(1 1)	0.1	1.05	*1.04*	*1.04*	*1.04*	1.25	1.30	1.33	1.47	1.47	1.54	3.00	2.09	1.00
40	(1 1)	0.5	1.41	*1.01*	*1.01*	*1.01*	1.13	1.24	1.30	2.01	2.01	2.04	2.79	1.86	1.00
40	(1 1)	0.9	3.16	*1.22*	*1.22*	*1.22*	1.26	1.46	1.56	2.04	2.04	2.05	2.07	1.84	1.00
100	(1 0)	0.1	1.05	*1.03*	*1.03*	**	***	***	***	1.09	1.09	2.16	3.00	2.97	2.95
100	(1 0)	0.5	1.41	*1.04*	*1.04*	3.00	***	***	***	1.23	1.23	1.45	3.00	3.00	2.99
100	(1 0)	0.9	3.16	*1.06*	*1.06*	4.21	**	***	***	1.62	1.62	1.63	2.46	3.00	3.00
100	(0 1)	0.1	1.05	*1.01*	*1.01*	**	***	***	***	1.81	1.81	2.58	3.00	3.00	3.00
100	(0 1)	0.5	1.41	*1.03*	*1.03*	5.90	***	***	***	2.03	2.03	2.25	3.00	3.00	3.00
100	(0 1)	0.9	3.16	*1.16*	*1.16*	6.28	**	***	***	2.03	2.04	2.07	2.71	3.00	3.00
100	(1 1)	0.1	1.05	*1.03*	*1.03*	**	***	***	***	1.53	1.53	2.40	3.00	3.00	3.00
100	(1 1)	0.5	1.41	*1.03*	*1.03*	6.39	***	***	***	2.02	2.03	2.25	3.00	3.00	3.00
100	(1 1)	0.9	3.16	*1.09*	*1.09*	9.46	***	***	***	2.04	2.05	2.08	2.69	3.00	3.00

The four DGP columns show the number of predictors k, the parameter vector β in $y_t = f_t\beta + \varepsilon_t$, the squared correlation ρ^2_{yf} between y and f, and the relative standard deviation of y, rsd(y).

The six RMSE columns show the RMSE defined in (8) for the methods PCR3 and PCovR (the last one for five weight factors). For each DGP row, values in italics show the methods with the smallest RMSE (although the differences are small in some cases). A ** stands for RMSE values between 10 and 100 and *** for values between 100 and 1000.

The six columns for the mean of p show the average over the thousand simulations of the number of factors selected by BIC for PCR3 and for PCovR for the different weight factors. In all cases, the maximal considered number of factors is $p = 3$.

the 100×100 predictor matrix X will be invertible. In this case, one factor f^* can be used to fit y in (4) perfectly, leaving the second factor to approximate X. The perfect fit for y is obtained by defining $f^* = Xa^*$, where the 100×1 vector a^* is defined by $a^* = X^{-1}y$. Clearly, this choice of a^* gives a perfect fit for y, which is particularly attractive if in the PCovR criterion (4) the weight w is large, but the forecasts generated in this way will be unreliable.

The foregoing arguments indicate that large weights w should be excluded if the number of predictors k is large relative to the number of observations T. We investigated the performance of PCovR in *larger samples* by increasing the number of observations from $T = 100$ (as reported in Table 1) to $T = 400$. The considered DGP's are the same as before, so that the number of observations is now considerably larger than the number of predictors, as $k \leq 100$. The RMSE of all models decreases, and very substantially so for PCovR with large weights. For instance, for $k = 40$ and $w = 0.9$, the RMSE ranges for $T = 100$ between 1.26 and 1.76 (see Table 1) and for $T = 400$ between 1.00 and 1.37. The improvements are even more substantial for $k = 100$; for $w = 0.01$, the RMSE ranges from 3.00 to over 10 for $T = 100$ and from 1.00 to 1.14 for $T = 400$, and for $w = 0.9$ the RMSE is always over 100 for $T = 100$ whereas it ranges between 1.12 and 1.27 for $T = 400$. The best forecasts for this DGP are still obtained for small weights. For $k = 10$ and $k = 40$, the RMSE's of PCR3 and PCovR with $w \leq 0.10$ are nearly identical, and for $k = 100$ this holds true for $w \leq 0.01$.

Finally, we mention some consequences of *misspecification* of the forecast model, that is, models where p is chosen either too small or too large as compared to the DGP with $p = 2$. The price of over-specification in models with $p = 3$ factors is in general small, as the RMSE of PCR3 and PCovR increases in most cases only by at most 0.01. Exceptions are PCovR for $k = 100$ with $w \geq 0.01$, as the previously discussed problem of overfitting in this situation gets worse for $p = 3$, and PCovR for $k = 40$ and $w = 0.1$ or 0.5, where the increase in RMSE is of the order 0.1 to 0.2. The consequences of under-specification in models with $p = 1$ factor depend on the DGP and on the employed model. For PCR3, the RMSE increases for all DGP's, least so for $\beta = (1 \;\; 0)'$ and most for $(0 \;\; 1)'$. These results are as expected, as for $p = 1$ PCR3 will select the first factor, whereas for $\beta = (0 \;\; 1)'$ the relevant predictor is the second factor. For instance, for the DGP with $k = 40$, $\beta = (0 \;\; 1)'$ and $\rho^2_{yf} = 0.9$, the RMSE for PCR3 with $p = 1$ is 3.07, as compared to 1.38 if $p = 2$ and 1.39 if $p = 3$. For PCovR with $p = 1$ factor, the RMSE also increases as compared to $p = 2$ for small weights, but in some cases the loss is very small. For instance, again for the DGP with $k = 40$, $\beta = (0 \;\; 1)'$ and $\rho^2_{yf} = 0.9$, the RMSE of PCovR with $w = 0.5$ is 1.68 for $p = 1$, as compared to 1.67 for $p = 2$ and 1.68 for $p = 3$. Further, for $w = 0.9$ the RMSE hardly depends on p for all DGP's, with differences of at most 0.01, except in overfitting cases with $k = 100$ and $w \geq 0.01$. So PCovR is less sensitive to under-specification than

PCovR in some situations, which is due to the fact that the relevant forecast factor in (1) can in principle be modelled by a single factor.

4.2 Simulation with regression DGP

In the second simulation, the data are generated by a regression model, that is, $y_t = x_t\beta + \varepsilon_t$. More precisely, in terms of dynamic factor models, the data are generated by (6) with $p = k$ factors and with the following parameters. The factor loading matrix is $\Lambda = I_k$ and $v_t = 0$, so that all factors are observed without error and there are no 'common' factors, but only independent factors that generate the predictor variables X. Further, Φ is a $k \times k$ diagonal matrix with value $\phi = 0.7$ on the diagonal, and $\sigma_u^2 = 0.51$. So the k predictors are mutually independent autoregressive processes of order one with covariance matrix $\mathrm{var}(x_t) = I_k$. We consider again the cases $k = 10$, 40, and 100. The $k \times 1$ vector β is either $(1, 0, \ldots, 0)'$ or $(1/\sqrt{k})(1, 1, \ldots, 1)'$, so that the relevant predictor is respectively one of the observed predictors and the average of all observed predictors. The vector β is scaled so that in both cases $\mathrm{var}(x_t\beta) = 1$. The squared correlation $\rho_{yf}^2 = \rho_{yx}^2$ between y_t and the relevant predictor $x_t\beta$ is either 0.1, 0.5 or 0.9.

With three options for the number k of predictors, two options for the parameter vector β and three options for the correlation between y and X, the total number of DGP's is eighteen. As the DGP has $p = k$ factors, we estimated PCR models with p ranging in a grid from $p = 1$ to (maximally, for $k = 100$) $p = 80$, as discussed in Section 3.2. For PCovR, we estimate models with $p = 1, 2$, or 3 factors, as the DGP has a single relevant prediction factor, that is, $F\beta$. For ease of comparison we also report the results of PCR3, that is, PCR with at most three factors.

As before, to limit the output tables, the number of factors for the PCR and PCovR models are selected by BIC. Table 2 shows the RMSE and the mean of p for the considered DGP's (in rows) and models (in columns), and we summarize the main results. PCovR performs better than PCR in most cases. The difference is the largest for $k = 10$, $\rho_{yx}^2 = 0.9$, and $w = 0.5$ or $w = 0.9$. In these cases, PCR uses on average $p = 7.5$ factors with an RMSE of around 1.5, whereas PCovR uses on average $p = 1$ factor with an RMSE of around 1.1. The higher RMSE of PCR is not due to allowing a larger number of factors, as PCR3 for these cases uses on average $p = 2.5$ factors with an RMSE of around 2.5, which is considerably worse than the RMSE of PCR. The results in Table 2 show that the RMSE of PCR is consistently lower than that of PCR3, and most notably so for $k = 10$ and $k = 40$ with $\rho_{yx}^2 = 0.5$ or 0.9. For $k = 40$, PCovR is still better than PCR, and hence also better than PCR3, but the gain in RMSE is more modest than for $k = 10$. For instance, for $k = 40$,

Table 2. RMSE and mean number of factors p in simulation experiment with regression DGP

DGP				RMSE							mean p						
						PCovR with weight w							PCovR with weight w				
k	β	ρ^2_{yx}	rsd(y)	PCR	PCR3	1E−4	0.01	0.10	0.50	0.90	PCR	PCR3	1E−4	0.01	0.10	0.50	0.90
10	one	0.1	1.05	1.09	*1.08*	*1.08*	*1.08*	*1.08*	1.10	1.12	1.55	1.35	1.35	1.37	1.50	1.07	1.00
10	one	0.5	1.41	1.13	1.28	1.30	1.28	1.14	*1.05*	*1.05*	5.74	2.18	2.20	2.23	2.31	1.00	1.00
10	one	0.9	3.16	1.50	2.54	2.62	2.50	1.47	1.08	*1.07*	7.51	2.51	2.52	2.58	2.66	1.05	1.00
10	all	0.1	1.05	1.08	1.08	*1.07*	1.08	1.08	1.09	1.10	1.53	1.34	1.35	1.36	1.48	1.06	1.00
10	all	0.5	1.41	1.14	1.29	1.31	1.30	1.17	*1.08*	*1.08*	5.70	2.15	2.15	2.18	2.32	1.00	1.00
10	all	0.9	3.16	1.52	2.52	2.59	2.47	1.51	*1.08*	*1.08*	7.54	2.48	2.48	2.53	2.63	1.05	1.00
40	one	0.1	1.05	*1.09*	1.10	*1.09*	*1.09*	1.22	1.41	1.43	1.31	1.20	1.20	1.28	2.02	1.00	1.00
40	one	0.5	1.41	1.35	1.42	1.43	1.40	*1.31*	1.42	1.44	5.01	1.73	1.74	1.89	1.91	1.00	1.00
40	one	0.9	3.16	1.63	2.98	3.01	2.88	1.60	*1.38*	1.39	25.43	2.13	2.10	2.28	2.26	1.03	1.00
40	all	0.1	1.05	*1.04*	*1.04*	*1.04*	*1.04*	1.14	1.32	1.33	1.25	1.17	1.18	1.24	2.03	1.00	1.00
40	all	0.5	1.41	1.32	1.37	1.38	1.36	*1.27*	1.38	1.40	4.97	1.72	1.71	1.93	1.92	1.00	1.00
40	all	0.9	3.16	1.68	2.98	3.01	2.89	1.64	*1.36*	*1.36*	25.81	2.11	2.11	2.32	2.30	1.03	1.00
100	one	0.1	1.05	1.08	1.08	*1.07*	1.27	**	**	**	1.19	1.13	1.13	1.37	1.59	2.17	1.53
100	one	0.5	1.41	*1.45*	*1.45*	*1.45*	2.13	**	**	**	3.61	1.61	1.59	1.96	1.63	2.08	1.76
100	one	0.9	3.16	*2.60*	3.08	3.10	3.29	**	**	**	32.21	1.95	1.95	2.28	1.85	2.04	1.96
100	all	0.1	1.05	*1.06*	*1.06*	*1.06*	1.34	**	**	**	1.22	1.14	1.14	1.40	1.53	2.16	1.55
100	all	0.5	1.41	*1.33*	1.36	1.36	1.69	**	**	**	3.65	1.66	1.65	2.02	1.64	2.04	1.79
100	all	0.9	3.16	*2.45*	2.96	2.98	3.16	**	**	**	32.93	1.99	1.99	2.32	1.85	2.03	1.98

The four DGP columns show the number of predictors k, the parameter vector β in $y_t = x_t\beta + \varepsilon_t$, the squared correlation ρ^2_{yx} between y and x, and the relative standard deviation of y, rsd(y). For β, 'one' stands for $\beta = (1, 0, \ldots, 0)'$ and 'all' stands for $\beta = (1/\sqrt{k})(1, 1, \ldots, 1)')$.

The seven RMSE columns show the RMSE defined in (8) for the methods PCR, PCR3 (with the restriction $p \leq 3$), and PCovR (for five weight factors). For each DGP row, values in italics show the methods with the smallest RMSE (although the differences are small in some cases). A ** stands for RMSE values between 10 and 100.

The seven columns for the mean of p show the average over the thousand simulations of the number of factors selected by BIC for PCR and PCR3 and for PCovR for the different weight factors. For PCR3 and PCovR, the maximal considered number of factors is $p = 3$, whereas for PCR p is selected from a grid of values with $p \leq 0.8k$.

$\rho^2_{yx} = 0.9$, and $w = 0.5$ or $w = 0.9$, PCovR has an RMSE of around 1.36 to 1.39, as compared to 1.63 to 1.68 for PCR and 2.98 for PCR3. For $k = 40$, the average number of factors for PCovR (roughly 1 to 2) is again much smaller than for PCR (with a mean of more than 25 if $\rho^2_{yx} = 0.9$). For $k = 100$, PCovR only provides acceptable predictions for $w = 0.0001$, with RMSE's comparable to PCR but with much fewer factors. For instance, if $k = 100$ and $\rho^2_{yx} = 0.9$ then PCovR uses on average roughly 2 factors, and PCR uses on average more than 32 factors. Larger weights do not give acceptable forecasts due to overfitting, as explained at the end of Section 4.1. Further, for all DGP's with $\rho^2_{yx} = 0.1$, the RMSE of the zero-prediction $\hat{y}_{T+1} = 0$—shown in the column rsd(y) in Table 2—is in most cases smaller than that of all PCR and PCovR methods. This result means that the forecast models do not perform well in case of low correlation between the predictor x and the forecasted variable y.

Again, if the number of observations is increased from $T = 100$ to $T = 400$ then the RMSE of PCovR decreases substantially for larger weights w. The improvements are largely comparable to those discussed at the end of Section 4.1.

The results are as expected. PCovR is a flexible method to construct factors that predict well. On the other hand, PCR constructs the factors in a preliminary step, without regarding the forecast objective. The BIC criterion is based on the (within-sample) fit, so that the forecast objective plays a role in selecting p, but it requires many principal components in order to have a reasonable chance to incorporate a significant part of the relevant predictor $F\beta$. Stated otherwise, PCR is more suited for DGP's with common factors than for regression-type DGP's as in this simulation.

4.3 Simulation with dyadic factor DGP

In the third simulation experiment, the DGP in a sense lies in between the 'extremes' discussed in Section 4.1, with very few factors that load equally much on large sets of predictors, and Section 4.2, with many factors that each load only on a single predictor. We consider a factor structure that we call dyadic, as the factors load on dyadic parts of the predictor set. We describe the model in detail for $k = 10$. In this case there are $p = 7$ factors. The first factor loads on predictors 1-8, the second factor on predictors 1-4, the third on 5-8, the fourth on 1-2, the fifth on 3-4, the sixth on 5-6 and the seventh on 7-8. The remaining two predictors (9 and 10) are white noise, not related to the factors. Therefore, the 7×10 loading matrix is as follows, where c is a scaling constant.

$$\Lambda = c \begin{pmatrix} 1\,1\,1\,1\,1\,1\,1\,1\,0\,0 \\ 1\,1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,1\,1\,1\,0\,0 \\ 1\,1\,0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1\,0\,0 \end{pmatrix}.$$

The dyadic expansion for $k = 10$ is for levels $d = 0, 1, 2$, and on level d there are 2^d factors each loading on sets of 2^{3-d} predictors. In a similar way, for $k = 40$ there are $2^5 - 1 = 31$ factors, with 2^d factors loading on sets of 2^{5-d} predictors, for $d = 0, 1, 2, 3, 4$, and for $k = 100$ there are $2^6 - 1 = 63$ factors, with 2^d factors loading on sets of 2^{6-d} predictors, for $d = 0, 1, 2, 3, 4, 5$.

For $k = 10$, the DGP is further specified as follows, with obvious modifications for $k = 40$ and $k = 100$. The 7×7 matrix Φ is diagonal with value $\phi = 0.8$ on the diagonal, and the variance of all components of the errors u_t is 0.36 so that the factors have covariance matrix $\text{var}(f_t) = I_7$. The predictors are normalized to have mean zero and variance one. Let x_i, λ_i and v_i denote respectively the i-th predictor variable, the i-th column of Λ and the i-th component of v, so that $x_{i,t} = f_t\lambda_i + v_{i,t}$. Then $\text{var}(x_{i,t}) = 1 = ||\lambda_i||^2 + \sigma^2_{v_i}$ and $\rho^2_{x_if} = ||\lambda_i||^2$. Therefore, a desired level of correlation ρ^2_{xf} is achieved by choosing the scaling constant c in the loading matrix Λ such that $3c^2 = \rho^2_{xf}$, that is, $c = \sqrt{\rho^2_{xf}/3}$. Three cases for β are considered. For $\beta = (1, 0, \ldots, 0)'$ the relevant predictor is the first, highest level dyadic factor that loads on the largest set of predictors, for $\beta = (0, 0, 0, 1, 0, 0, 0)'$ the relevant predictor is the first lowest level dyadic factor that loads only on the first two predictors, and for $\beta = (1/\sqrt{3})(1, 1, 0, 1, 0, 0, 0)'$ the relevant predictor is the average of all factors that load on the first two predictors. In all cases, β is normalized to have norm 1. Finally, the squared correlations ρ^2_{yf} and ρ^2_{xf} are either 0.5 or 0.9. With three options for k, three options for β and four options for $(\rho^2_{yf}, \rho^2_{xf})$ this gives in total thirty-six DGP's. However, as the results for $(\rho^2_{yf}, \rho^2_{xf}) = (0.5, 0.5)$ are quite close to those for $(0.5, 0.9)$, we will not report the results for $(0.5, 0.5)$ to limit the size of the output table to 27 DGP rows. The considered PCR and PCovR models are the same as in Section 4.2.

In Table 3, we present the RMSE's and the mean value of the number of factors p if BIC is used for PCR, PCR3 and the five PCovR methods with w equal to 0.0001, 0.01, 0.1, 0.5 or 0.9. We summarize the main results. Overall, PCR and PCovR with $w = 0.0001$ perform best. These two methods have comparable RMSE's, although PCR is on average somewhat better than PCovR, in particular if $\rho^2_{yf} = \rho^2_{xf} = 0.9$. One notable exception is the DGP with $k = 40$ predictors, with β 'low', so that the DGP prediction factor is at the lowest dyadic level, and with $\rho^2_{yf} = \rho^2_{xf} = 0.9$. For this DGP, PCovR with weight w equal to 0.1, 0.5 or 0.9 has an RMSE of around 2.75, as compared to 2.91

Table 3. RMSE and mean number of factors p in simulation experiment with dyadic factor DGP

DGP					RMSE							mean p						
							PCovR with weight w							PCovR with weight w				
k	β	ρ^2_{yf}	ρ^2_{xf}	rsd(y)	PCR	PCR3	1E−4	0.01	0.10	0.50	0.90	PCR	PCR3	1E−4	0.01	0.10	0.50	0.90
10	high	0.5	0.9	1.41	*1.20*	*1.20*	*1.20*	*1.20*	*1.20*	*1.20*	1.23	1.30	1.24	1.25	1.25	1.23	1.07	1.00
10	high	0.9	0.5	3.16	*2.44*	*2.44*	*2.44*	*2.44*	*2.44*	2.46	2.50	1.49	1.40	1.40	1.40	1.27	1.00	1.00
10	high	0.9	0.9	3.16	2.29	2.29	2.29	2.29	*2.28*	*2.28*	2.32	1.78	1.46	1.47	1.46	1.39	1.05	1.00
10	low	0.5	0.9	1.41	1.21	1.34	1.35	1.35	1.31	1.22	*1.20*	4.31	1.73	1.72	1.76	2.12	1.88	1.00
10	low	0.9	0.5	3.16	2.88	2.98	2.99	2.98	2.93	2.82	*2.81*	3.61	1.87	1.89	1.92	2.16	1.37	1.00
10	low	0.9	0.9	3.16	*2.25*	2.90	2.93	2.92	2.63	2.33	2.31	5.68	2.00	2.00	2.04	2.50	1.78	1.00
10	mix	0.5	0.9	1.41	1.10	1.12	1.12	1.12	1.10	*1.09*	*1.09*	3.01	2.01	2.01	2.01	1.99	1.44	1.00
10	mix	0.9	0.5	3.16	1.93	1.99	1.99	1.98	1.92	*1.90*	1.93	3.70	2.23	2.23	2.22	2.04	1.06	1.00
10	mix	0.9	0.9	3.16	*1.25*	1.58	1.59	1.58	1.43	1.27	1.27	5.75	2.38	2.38	2.38	2.36	1.79	1.00
40	high	0.5	0.9	1.41	1.23	1.23	1.23	*1.22*	1.24	1.47	1.54	1.57	1.40	1.41	1.44	1.92	1.48	1.00
40	high	0.9	0.5	3.16	*2.46*	2.48	2.48	2.47	2.51	2.78	2.90	2.02	1.68	1.68	1.71	1.90	1.03	1.00
40	high	0.9	0.9	3.16	2.35	2.37	2.37	2.37	*2.34*	2.61	2.73	2.66	1.82	1.83	1.85	2.09	1.36	1.00
40	low	0.5	0.9	1.41	*1.44*	*1.44*	*1.44*	*1.44*	1.45	1.57	1.59	2.54	1.58	1.57	1.68	2.74	1.43	1.00
40	low	0.9	0.5	3.16	3.11	3.11	3.11	*3.10*	3.21	3.49	3.55	2.71	1.73	1.74	1.89	2.77	1.07	1.00
40	low	0.9	0.9	3.16	2.91	3.10	3.10	3.08	2.76	*2.75*	2.77	8.32	1.91	1.90	2.02	2.85	1.47	1.00
40	mix	0.5	0.9	1.41	*1.13*	1.16	1.17	1.16	1.14	1.27	1.31	3.45	2.30	2.31	2.35	2.64	1.51	1.00
40	mix	0.9	0.5	3.16	2.03	2.13	2.15	2.10	*1.99*	2.18	2.24	4.63	2.56	2.56	2.60	2.60	1.15	1.00
40	mix	0.9	0.9	3.16	1.54	1.98	1.99	1.93	1.55	*1.50*	1.55	10.72	2.65	2.64	2.67	2.80	1.96	1.00
100	high	0.5	0.9	1.41	1.25	*1.24*	1.25	**	****	****	****	1.76	1.45	1.45	1.59	2.89	2.94	2.88
100	high	0.9	0.5	3.16	2.39	*2.37*	*2.37*	**	***	***	***	2.47	1.76	1.75	1.92	2.80	2.90	2.89
100	high	0.9	0.9	3.16	*2.30*	*2.30*	2.39	**	****	****	****	3.22	1.85	1.84	1.96	2.72	2.91	2.92
100	low	0.5	0.9	1.41	1.42	*1.41*	*1.41*	5.00	****	****	****	2.34	1.50	1.49	1.72	2.91	2.88	2.82
100	low	0.9	0.5	3.16	*3.14*	3.16	3.15	**	***	***	***	2.73	1.70	1.70	2.10	2.82	2.82	2.75
100	low	0.9	0.9	3.16	*3.12*	3.16	3.15	**	****	***	****	4.39	1.79	1.80	2.05	2.89	2.89	2.81
100	mix	0.5	0.9	1.41	*1.18*	1.20	1.21	**	****	***	***	3.50	2.17	2.17	2.31	2.87	2.90	2.88
100	mix	0.9	0.5	3.16	*2.07*	2.21	2.22	**	***	***	***	5.03	2.49	2.49	2.60	2.82	2.86	2.85
100	mix	0.9	0.9	3.16	*1.78*	2.17	2.21	**	***	****	***	7.49	2.54	2.54	2.61	2.84	2.87	2.87

The five DGP columns show the number of predictors k, the type of parameter vector β in $y_t = f_t\beta + \varepsilon_t$, the squared correlations ρ^2_{yf} between y and f and ρ^2_{xf} between x and f, and the relative standard deviation of y, rsd(y). For β, 'high' and 'low' stand respectively for the highest level and first lowest level dyadic factor, and 'mix' stands for the case where the DGP prediction factor is the average of all dyadic factors that load on the first two predictors.
The seven RMSE columns show the RMSE defined in (8) for the methods PCR, PCR3 (with the restriction $p \leq 3$), and PCovR (for five weight factors). For each DGP row, values in italics show the methods with the smallest RMSE (although the differences are small in some cases). A ** stands for RMSE values between 10 and 100, *** for values between 100 and 1000, and **** for values larger than 1000.
The seven columns for the mean of p show the average over the thousand simulations of the number of factors selected by BIC for PCR and PCR3 and for PCovR for the different weight factors. For PCR3 and PCovR, the maximal considered number of factors is $p = 3$, whereas for PCR p is selected from a grid of values with $p \leq 0.8k$.

for PCR and 3.10 for PCR3. That is, PCovR is better in detecting the relevant (lowest dyadic level) prediction factor that loads only on the first two predictors out of the observed set of forty predictors. Further, for all DGP's, the number of factors is consistently larger for PCR than for PCovR, up to a factor four for DGP's with $\rho^2_{yf} = \rho^2_{xf} = 0.9$. PCR3 has a larger RMSE than PCR for nearly all DGP's, and the increase is the largest—up to over 30%—for $k = 10$ and $k = 40$ with $\rho^2_{yf} = \rho^2_{xf} = 0.9$ and with β of type 'low' or 'mix'. For $k = 100$, PCovR only works well for small weights because larger weights lead to overfitting, as explained at the end of Section 4.1. If the number of observations is increased from $T = 100$ to $T = 400$ then the RMSE of PCovR decreases substantially for larger weights w, with improvements comparable to those discussed at the end of Section 4.1.

5 Conclusion

In this paper, we compared Principal Component Regression (PCR) and Principal Covariate Regression (PCovR). In PCovR, the factors are estimated by minimizing a criterion that consists of a weighted average of the squared errors for the dependent variable y and those for the predictors X. Simulation experiments show that the PCovR forecasts may outperform the two-step PCR forecasts if the data are generated by many underlying factors. PCR performs better for relatively low-dimensional factor models, but comparable forecast accuracy can be obtained by PCovR models with fewer factors if the weight factor w is chosen sufficiently small.

We conclude by mentioning some possible extensions that we are currently working on. The number of factors was chosen by BIC, and it is of interest to consider forecast-based selection methods, for instance, cross-validation methods. Further, if the number of predictors is large then the PCovR weights should be small to prevent overfitting, and it is of interest to provide bounds on the weight w in terms of the data. Another option is to apply some kind of regularization to prevent overfitting. Finally, it is of practical interest in time series forecasting to extend PCovR by including lagged factors and variables, as was done for PCR, for instance, by Stock and Watson (1999).

A Algorithm for PCovR

An algorithm for PCovR in Section 2.3 is described in De Jong and Kiers (1992). Nonetheless, for clarification we present an explicit SVD based algorithm. We prove that the minimization of (4) can be solved by means of two SVD's. We use the notation of Section 2.3, and for simplicity we assume that all variables are scaled to have sample mean zero. Then the best estimate of α is $\alpha = 0$, so that we can discard this parameter in what follows

Let $\tilde{y} = \sqrt{w_1}y$, $\tilde{\beta} = \sqrt{w_1}\beta$, $\tilde{X} = \sqrt{w_2}X$ and $\tilde{B} = \sqrt{w_2}B$, and let $D = [\tilde{y}, \tilde{X}]$ be the $T \times (k+1)$ weighted data matrix and $C = [\tilde{\beta}, \tilde{B}]$ the $p \times (k+1)$ matrix of coefficients. Then the PCovR criterion (4) can be written as the minimization (in the sense of the Frobenius norm) of

$$f(G) = ||\tilde{y} - XA\tilde{\beta}||^2 + ||\tilde{X} - XA\tilde{B}||^2 = ||D - XAC||^2 = ||D - XG||^2.$$

Here D and X are known data matrices—as the weights w_1 and w_2 are known—and $G = AC$ is a $k \times (k+1)$ matrix of reduced rank p that should be chosen to minimize $f(G)$. This can be solved as follows.

(1) Let $m = \text{rank}(X)$ and let $X = USV'$ be an SVD of X, with S an $m \times m$ diagonal matrix with the (positive) singular values of X—in decreasing order—on the diagonal and with U $(T \times m)$ and V $(k \times m)$ such that $U'U = V'V = I_m$.

(2) The minimization of $f(G)$ is equivalent to minimization of $\tilde{f}(G) = ||U'D - SV'G||^2$, which can be seen as follows.

$$\begin{aligned} f(G) &= \text{trace}[(D - XG)'(D - XG)] \\ &= \text{trace}(D'D) - 2\text{trace}(D'XG) + \text{trace}(G'X'XG) \\ &= \text{trace}(D'D) - 2\text{trace}(D'USV'G) + \text{trace}(G'VS^2V'G), \\ \tilde{f}(G) &= \text{trace}[(U'D - SV'G)'(U'D - SV'G)] \\ &= \text{trace}(D'UU'D) - 2\text{trace}(D'USV'G) + \text{trace}(G'VS^2V'G). \end{aligned}$$

The terms trace$(D'D)$ and trace$(D'UU'D)$ are constant, that is, independent of the choice of G), as D is defined in terms of the data (y, X) and the weights (w_1, w_2) and U is defined in terms of the data X. The above result shows that $f(G)$ and $\tilde{f}(G)$ attain their minimum for the same G.

(3) The optimal rank p approximation of $U'D$ is obtained by SVD—using the first p singular values and vectors—which we write as $(U'D)_p = U_pS_pV_p'$, where U_p $(m \times p)$, S_p $(p \times p)$ and V_p $(k+1) \times p$ with $U_p'U_p = V_p'V_p = I_p$. The optimal choice of G is then given by $G = VS^{-1}(U'D)_p = VS^{-1}U_pS_pV_p'$.

(4) The optimal rank p approximation of D is therefore $XG = U(U'D)_p$. The first column of the $T \times (k+1)$ matrix XG is the corresponding approximation $\tilde{y}_p$ of $\tilde{y}$, and the last k columns of XG give the approximation $\tilde{X}_p$ of $\tilde{X}$. The corresponding approximation of y is $(1/\sqrt{w_1})\tilde{y}_p$ and that of X is $(1/\sqrt{w_2})\tilde{X}_p$, which gives a rank p approximation of the $T \times k$ regressor matrix X.

(5) Explicit expressions for the optimal parameters (A, B, b) are obtained as follows. Let $A = GV_p = VS^{-1}U_pS_p$, $b = (1/\sqrt{w_1})(V_p')_1$ (the scaled first column of V_p'), $B = (1/\sqrt{w_2})(V_p')_{2-(k+1)}$ (the scaled columns 2 to $(k+1)$ of V_p') and $F = XA$. Then $XG = XVS^{-1}U_pS_pV_p' = XAV_p' = FV_p'$, so

that we can write the minimal value of the criterion function as

$$\begin{aligned}||D - XG||^2 = ||D - FV_p'||^2 &= ||[\sqrt{w_1}y \quad \sqrt{w_2}X] - F[\sqrt{w_1}b \quad \sqrt{w_2}B]||^2 \\ &= w_1||y - XAb||^2 + w_2||X - XAB||^2.\end{aligned}$$

So the above expressions define optimal values for A, B and β. The factors are $F = XA$, and the optimal approximation of y is $\hat{y} = XAb$ and that of X is $\hat{X} = XAB$.

References

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

[2] Boivin, J, and S. Ng (2003), Are more data always better for factor analysis?, *NBER Working paper no 9829*, forthcoming in *Journal of Econometrics.*

[3] De Jong, S., and H.A.L. Kiers (1992), Principal covariate regression, *Chemometrics and Intelligent Laboratory Systems* 14, pp. 155-164.

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

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

[6] Stock, J.H., and M.W. Watson (2002b), Macroeconomic forecasting using diffusion indixes, *Journal of Business and Economic Statistics* 20, pp. 147-162.

[7] Stock, J.H., and M.W. Watson (2004), Forecasting with many predictors, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, North-Holland, Amsterdam (to appear).