Journal of Management Science and Engineering 4 (2019) 12–27

Science Press

Contents lists available at ScienceDirect

Journal of Management Science and Engineering

journal homepage: www.keaipublishing.com/en/journals/journal-of-management-science-and-engineering/

Forecasting using supervised factor models

Yundong Tu [a], Tae-Hwy Lee [b,*]

[a] Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China
[b] Department of Economics, University of California, Riverside, CA, 92521, USA

ARTICLE INFO

Article history:
Available online 28 March 2019

JEL classification:
C53
C55
E31

Keywords:
Combining forecasts
Principal components
Supervision matrix
Fixed point
Principal covariate regression
Partial least squares

ABSTRACT

This paper examines the theoretical and empirical properties of a supervised factor model based on combining forecasts using principal components (CFPC), in comparison with two other supervised factor models (partial least squares regression, PLS, and principal covariate regression, PCovR) and with the unsupervised principal component regression, PCR. The supervision refers to training the predictors for a variable to forecast. We compare the performance of the three supervised factor models and the unsupervised factor model in forecasting of U.S. CPI inflation. The main finding is that the predictive ability of the supervised factor models is much better than the unsupervised factor model. The computation of the factors can be doubly supervised together with variable selection, which can further improve the forecasting performance of the supervised factor models. Among the three supervised factor models, the CFPC best performs and is also most stable. While PCovR also performs well and is stable, the performance of PLS is less stable over different out-of-sample forecasting periods. The effect of supervision gets even larger as forecast horizon increases. Supervision helps to reduce the number of factors and lags needed in modelling economic structure, achieving more parsimony.

1. Introduction

High dimensional information in the presence of many predictors brings opportunities to improve the *efficiency* of a forecast by using much richer information than conventionally used and to enhance the *robustness* of a forecast against structural instability which can plague low dimensional forecasting. However, these opportunities come with the challenges. One notable challenge is that the availability of overwhelming information complicates the way we process it to make relevant instruments. To deal with the challenge we “supervise” the high dimensional information. Here, supervision refers to training the predictors for a variable to forecast.

Two types of supervision are considered. The first type of supervision is variable selection or subset selection which refers to selecting variables that are most predictive for a target variable of interest. Do we need to supervise the selection of predictors for a forecast target variable? Bair, Hastie, Paul, and Tibshirani (2006) and Bai and Ng (2008) address this question. They reported that after variable selection (by either hard-threshold method or soft-threshold method), the Principal Component Regression (PCR) performs much better, reducing the mean squared forecast error (MSFE) to a large extent. Various variable selection methods have been proposed such as forward and backward selection, stepwise regression, as

[*] Corresponding author. Department of Economics, University of California, Riverside, CA, 92521, USA.
E-mail address: taelee@ucr.edu (T.-H. Lee).

https://doi.org/10.1016/j.jmse.2019.03.001

presented in e.g. Miller (2002) and Hastie, Tibshirani, and Friedman (2009). Recently the literature is filled with more sophisticated methods such as LASSO (Tibshirani, 1996; Zou, 2006), Elastic Net (Zou & Hastie, 2005; Zou & Zhang, 2009), SCAD (Fan & Li, 2001), Bridge (Huang, Horowitz, & Ma, 2008), Least Angle Regression (Efron, Hastie, Johnstone, & Tibshirani, 2004) and so on. All these methods seek to rank the variables and select a subset of variables based on their ranks.

The second type of supervision can be taken in the process of computing low dimensional latent factors from the high dimensional predictors. If the low dimensional factors are computed only from the predictors X but not using the forecast target y, the factors are not supervised for the forecast target. This approach includes PCR. However, the PCR accounts only for the variation of the selected predictors, but does not directly employ the information about the forecast target. That is, no matter which variable to forecast (whether it is output growth, unemployment, stock returns, bond yields, housing price, interest rate, or inflation), the PCR uses the latent factors of the predictors only. Hence, another question arises. Do we need to supervise the computation of the latent factors for a particular forecast target? If the factors are computed not only from the predictors but also from using the forecast target, the factors are supervised for the forecast target. This paper addresses this question, by considering the supervised factor models. The supervised factor model approach includes Partial Least Square (PLS) regression (de Jong, 1993; Garthwait, 1994), Principal Covariate Regression (PCovR, de Jong, 1992), and Combining Forecast Principal Components (CFPC) which is the PCR on many forecasts constructed in a particular way which will be described shortly.

The question is whether the supervised factors from these supervised factor models are more *efficient* and more *robust* in out-of-sample forecasting than the unsupervised factors from PCR. The first contribution is to examine the properties of these factor models and compare their empirical performance with supervision on the variable selection and/or on the factor computation. The evidence is very clear. These supervisions do substantially improve the prediction. The predictive ability of the three supervised factor models is much better than the unsupervised PCR model. Interestingly, we find that the effect of supervision gets even larger as forecast horizon increases and that the supervision helps a model achieving more parsimonious structure. Among the three supervised factor models, the CFPC performs best and is most stable. While PCovR is nearly as efficient and robust as CFPC, PLS is not as good or stable as CFPC and PCovR. The performance of PLS is not robust over different out-of-sample forecasting periods and over the different forecast horizons. The double supervision of the variable selection and factor computation helps even more in out of sample forecasting.

Since the paper by Boivin and Ng (2006), the proper composition of data used for factor analysis is a widely discussed topic in the literature. In particular for forecasting, it has turned out that simple principal components factors estimated from very large datasets from time to time seem not to be good predictors in macroeconomic applications. In this regard, the second contribution of this paper is to introduce the sequential use of supervised variable selection and supervised factor estimation, which we name as “double supervision” as discussed in Section 5.

The rest of the paper is organized as follows. Section 2 introduces the basic forecasting setup and preliminary material that is needed for the understanding of factor models. Section 3 presents the unsupervised factor model, PCR. Section 4 examines the supervised factor models, PLS, PCovR and CFPC. Section 5 looks into ways to supervise the factor computation together with variable selection. In Section 6, forecasting exercises are carried out to compare the performance of these forecasting models for monthly CPI inflation in U.S. Section 7 concludes.

2. General framework: linear factor model

Consider the linear regression model,

$$y = X\beta + e, \tag{1}$$

where y is a $T \times 1$ vector, X is a $T \times N$ matrix of explanatory variables and β is the true but unknown parameter. In case of $N \gg T$, or when columns of X are highly correlated, the OLS estimation of the regression coefficient β is not feasible. Hence, for the purpose of forecasting, we consider the following factor model,

$$F = XR, \tag{2}$$

$$XB = FP' + E, \tag{3}$$

$$y = UQ' + G. \tag{4}$$

Here F is a $T \times r$ factor matrix. Equation (2) says that the factor is linear in X. Each column of F is a factor, which is a linear combination of rows of X. The $N \times r$ matrix R is the weight matrix attaching to X. U is the factor matrix for y, which is usually assumed to be the same as F. However, the estimation of U varies as we take different estimation approaches and it can be far different from F as in PLS. P and Q are corresponding factor loading for X and y. The $N \times N$ matrix B is called “supervision matrix”. Note that the factor structure (3) contains that of Stock and Watson (2002a) and Bai (2003) as a special case, with B being the identity matrix. As it is formulated, (4) is a linear factor model due to the linearity in both the construction of F in (2) and the prediction equation (4). E and G are the error terms. While it appears that B and P may not be identified as presented, we will demonstrate special cases of this general framework in Sections 3 and 4 with respective specifications of F, R, B, P, U and. Q.

In the case that the number of factors used in (4) is less than or equal to the number of observations, T, the coefficient Q can be estimated by OLS estimator $\widehat{Q}$, with U being estimated by $\widehat{U}$. The forecast is formed as

$$\widehat{y} = \widehat{U}\widehat{Q}'. \tag{5}$$

The factor models, PCR, PLS, PCovR and CFPC, that we consider in this paper all fall into this general framework of (2), (3) and (4), with different ways of specifying R, U and B. For example, as will be seen in the next section, PCR takes B as the identity matrix, and then forms the weight matrix R to be the matrix of eigenvectors of $X'X$, with U being F.

The choice of the weight matrix and number of factors is the focus of factor modelling. To choose the number of factors, the usual information criterion such as AIC or BIC can be used. In the empirical section (Section 6), we will look into this aspect in further details. We focus on the choice of weight matrix in the next two sections. Section 3 will present a popular (unsupervised) factor model, PCR, which has been extensively used in economic forecasting as well as in other social sciences. See Stock and Watson (2002a). PCR is unsupervised and methods of supervising it will be presented in Section 4.

For matrix decomposition used later, we adopt the following convention: for a $T \times N$ matrix C, it is decomposed into two blocks C_1 and C_2, with C_1 containing its first r columns $c_1, \ldots, c_r$ and C_2 containing the rest. That is, $C \equiv [C_1, C_2]$, where $C_1 = [c_1, \ldots, c_r]$ and $C_2 = [c_{r+1}, \ldots, c_N]$ Also '$a := b$' means that a is defined by b, while '$a =: b$' means that b is defined by a.

3. PCR

In this section we review how PCR can be used in forecasting. First we begin by using the eigenvalue decomposition, and then in Section 4.1 we show PCR in an alternative framework for the principal component analysis (the NIPALS algorithm for PCR, with details in Appendix A). The purpose of presenting these two alternative framework is that we will use the former to introduce a supervised factor model called CFPC in Section 4.3 and we will use the latter to introduce another supervised factor model called PLS in Section 4.1.

Note that, PCR is when $P = R, B = I$ and $U = F$ in the framework in Section 2, namely:

$$F = XR,$$

$$X = FR' + E,$$

$$y = FQ' + G,$$

where R is the matrix of eigenvectors of $X'X$. Stock and Watson (2002a) considered the case when y is one variable with (X, y) admitting the factor representation of (3) and (4). Equation (4) specifies the forecast equation while (3) gives the factor structure. The factor F in (2) is estimated using principal components and then it is used to form the prediction from (4) for y.

Let the $N \times r$ $(r \leq \min(T, N))$ matrix R_1 be the first r eigenvectors, corresponding to the largest r eigenvalues $\Lambda_1 = \text{diag}\ (\lambda_1, \ldots, \lambda_r)$ of $X'X$. The eigenvalue decomposition of $X'X$ is

$$X'X = R\Lambda R' = R_1\Lambda_1 R_1' + R_2\Lambda_2 R_2', \tag{6}$$

where $\Lambda = \text{diag}\ (\Lambda_1, \Lambda_2)$ is the eigenvalue matrix and $R = [R_1, R_2]$ is the eigenvector matrix corresponding to Λ. As R is orthonormal with $R'R = I$, we have

$$R_1'X'XR_1 = \Lambda_1. \tag{7}$$

Stock and Watson (2002a,b) has shown that the true factors can be consistently estimated by the first r principal components of X. Therefore, we adopt that $\widehat{F} = XR_1$. With $\widehat{U} = \widehat{F}$, the OLS estimator of the coefficient Q, $r \times 1$ vector, is given as

$$\begin{aligned} \widehat{Q} &= \left(\widehat{F}'\widehat{F}\right)^{-1}\widehat{F}'y \\ &= (R_1'X'XR_1)^{-1}R_1'X'y \\ &= \Lambda_1^{-1}R_1'X'y. \end{aligned} \tag{8}$$

Therefore, PCR forecast is formed as

$$\widehat{y}_{\text{PCR}} = \widehat{F}\widehat{Q} = XR_1\Lambda_1^{-1}R_1'X'y. \tag{9}$$

The main criticism on PCR goes as follows. In the choice of the weight matrix R, PCR imposes only the factor structure for X. This is naive since it does not take into account the dependent variable y. That is, no matter what y to forecast, PCR uses the same fixed combination of X to form the prediction equation. Ignoring the target information of y in the computation of the factors leads to inefficient forecast of the target y. Therefore, a supervision on the choice of weight matrix and thus supervised factor models will be called for to make more efficient predictions. This issue is to be addressed in the next section.

4. Supervised factor models

In this section we consider three supervised factor models, the partial least square, principal covariate regression, and the combining forecast-principal component. The analysis here is based on the factor framework in Section 2. The three models are generalization of the PCR in different ways to supervise the factors for the forecast target. y.

4.1. PLS

Although originally proposed by Wold (1966) in the field of econometrics, the partial least square (PLS) regression has rarely been used in economics but rather popular in chemometrics. Empirical results in chemistry show that PLS is a good alternative to multiple linear regression and PCR methods. See Wold, Ruhe, Wold, and Dunn (1984), Otto and Wegscheider (1985), and Garthwait (1994) for more details. Since PLS also supervises the factor computation process, it raises the possibility that it can outperform PCR, which is a reason that we include the PLS in this paper.[1]

There have been several algorithms designed for PLS, among which NIPALS (Nonlinear Iterative PArtial Least Square) is the most notable one. In the next subsection we review the NIPALS algorithm briefly to show PLS in the general framework of factor models, (2), (3) and (4), in Section 2. The purpose of the next subsection is to show that PLS can be viewed as a generalization of PCR. However, the readers uninterested in details can skip the following subsection and Appendix A-B.

4.1.1. NIPALS algorithm for PCR and PLS

Alternative to the eigenvalue decomposition used in Section 3 for PCR, we can use the Nonlinear Iterative PArtial Least Square (NIPALS) algorithm developed by Wold (1966, 1975) to perform the principal component analysis, which decomposes matrix X of rank r as a sum of r matrices of rank 1 as

$$\begin{aligned} X &= M_1 + M_2 + \ldots + M_r + E_r \\ &= f_1p_1' + f_2p_2' + \ldots + f_rp_r' + E_r =: FP' + E_r, \end{aligned} \tag{10}$$

where the second line uses the fact that the rank 1 matrices M_h can be written as outer products of two vectors, f_h (score) and p_h' (loading), and $F = [f_1, f_2, \ldots, f_r]$, $P' = [p_1', p_2', \ldots, p_r']$. NIPALS does not compute all the principal components F at once. But it calculates f_1 and p_1' from X, then the outer product f_1p_1' is subtracted from X, and the residual E_1 is calculated. This residual is used to calculate f_2 and p_2', and so on. The formal NIPALS algorithm for PCR is stated in Appendix A, where it is shown that, on convergence, the NIPALS algorithm gives the same principal components as derived by the eigenvalue decomposition of Section 3 for PCR. The algorithm does converge in practice.

Now, to see how this algorithm can be extended from PCR to PLS, let us turn back to the regression problem (4). The NIPALS algorithm can work for both X and y separately to extract factors as in (10). That is,

$$X = FP' + E_r = \sum_{h=1}^{r} f_hp_h' + E_r, \tag{11}$$

$$y = UQ' + G_r = \sum_{h=1}^{r} u_hq_h' + G_r. \tag{12}$$

Thus we can form an inner relationship between x-score, f, and y-score, u as

$$u_h = \gamma_h f_h + \varepsilon_h, \tag{13}$$

for each pair of components. OLS estimation can be used for (13) thus we could use (12) to form a prediction with x-scores, f, extracted with newly observed x.

However, note that the decomposition process in (11) and (12) still does not incorporate the valuable information of y when forming the x-scores. Thus we consider the modification of the decomposition of X and y, using NIPALS, as stated in Appendix B. Note that in the special case of $y = X$, x-factors extracted by NIPALS gives exactly the principal components of X as one might have already conjectured. Thus in this case, NIPALS for PLS is the same as NIPALS for PCR. See Geladi and Kowalski (1986) and Mardia, Kent, and Bibby (1980) for an excellent discussion for NIPALS algorithm and its adaptations for PCR and

[1] We note that the statistics literature is quite sceptical concerning the theoretical properties of PLS, see Butler and Denham (2000) and Lingjaerde and Christophersen (2000). In particular, PLS is known to have strange shrinkage properties. Both papers derive the PLS estimator in a similar way as a shrinkage estimator in a multivariate regression problem. However, both papers find that PLS does not shrink but rather expands some of the coefficients. Recently Kelly and Pruitt (2015) propose the three-pass regression filter method which is related to PLS.

PLS. The PLS is to find a linear combination of the columns of X leading to a maximal covariance with the forecast target variable y. See Groen and Kapetanios (2016) where the relationship between PCR and PLS is analyzed.

4.2. Principal covariate regression

Principal Covariate Regression (PCovR) is a prediction method proposed by de Jong (1992). “Covariate” was termed to stress that, apart from PCR, the components should vary with the dependent variable y. The attractiveness of PCovR features its combination of PCR on X and a regression on y by minimizing an appropriately defined least square loss function as follows,

$$l(\alpha_1, \alpha_2, R, P, Q) = \alpha_1 ||X - XRP'||^2 + \alpha_2 \|y - XRQ'\|^2, \tag{14}$$

where α_1 and α_2 are the (non-negative) weights attached to PCR on X and regression of y, respectively. That is, the choice of the factor weight matrix R depends not only on the PCR of X, but also on the regression equation (4). Then the factor is computed from $F = XR$ as in (2).

Note that, PCovR is when $B = I$ and $U = F$ in the framework in Section 2. Unlike PCR, $P \neq R$. While PCR takes (3) and (4) in two separate steps, PCovR puts them together in one step as shown in (14). Two special cases of PCovR need to be pointed out here. For $\alpha_1 = 0$, the loss (14) emphasizes completely on fitting y. Another extreme is when $\alpha_2 = 0$. In this case, (14) emphasizes completely on the principal component analysis on X or PCR as described in Section 3.

Note that the minimization of (14) is nonlinear in nature due to the product terms RP and RQ. An algorithm for the estimation of the unknown parameters (R, P, Q) is given in de Jong (1993). Or see Heij, Groenen, and van Dijk (2007) for an explicit SVD based algorithm.[2]

Although supervision is incorporated in PCovR by allocating weight to the regression (4), there is no guidance regarding the optimal choice of the weight attached. Thus choice of α_1 and α_2 can only be done on rather arbitrary grounds. In practice, one may consider a set of specifications for α_1 and α_2, as did in Heij et al. (2007).

For prediction purpose, we propose an estimation of optimal weights by a grid search algorithm, with the exploit of information available. Note that only the relative weights attached matter here. We consider a normalization of the weights by the norm of the data matrix, that is,

$$\alpha_1 = w \Big/ \|X\|^2, \text{ and } \alpha_2 = (1-w) \Big/ \|y\|^2.$$

Therefore, we need to consider a choice of w instead of choices of α_1 and α_2 simultaneously. In Section 6, we choose the value of w from $\{10^{-6}, 10^{-4}, 0.1, 0.5, 0.9\}$ together with the determination of other model parameters, such as number of factors and lags, by the Bayesian information criterion.

4.3. CFPC

This subsection discusses another form of supervision on the computation of factors. This is a method quite different from those examined earlier in this section. The two previous supervised models directly compute the factors, while CFPC first computes forecasts and then computes the principal components of the forecasts as a tool to combining forecasts.

Consider a linear projection of y on x_i for each $i = 1, 2, \ldots, N$,

$$y = x_i b_{0i} + u_i, \tag{15}$$

where b_{0i} is estimated by OLS,

$$b_i := (x_i' x_i)^{-1} x_i' y. \tag{16}$$

Thus the prediction could be formed as

$$\hat{y}_i := x_i b_i. \tag{17}$$

To write (17) in compact form,

$$\hat{Y} := [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N] = [x_1 b_1, x_2 b_2, \ldots, x_N b_N] =: XB, \tag{18}$$

where $B := \text{diag }(b) = \text{diag }(b_1 \ldots b_N)$ is the diagonal matrix with $b_1, b_2, \ldots, b_N$ sitting on the diagonal. Let L_1 be the $N \times r$ eigenvectors corresponding to the r largest eigenvalues $\Omega_1 = \text{diag }(\omega_1, \omega_2, \ldots, \omega_r)$ of $\hat{Y}'\hat{Y}$. Parallel to (6), we also have its eigenvalue decomposition of $\hat{Y}'\hat{Y}$ as follows,

[2] We would like to thank Christiaan Heij and Dick van Dijk for kindly sharing their Matlab code for PCovR.

$$\widehat{Y}'\widehat{Y} = L\Omega L' = L_1'\Omega_1 L_1 + L_2\Omega_2 L_2'.$$

The principal component estimator of F which is the first r principal components of $\widehat{Y}$, is therefore given as $\widehat{F} = \widehat{Y}L_1 = XBL_1$. Then consider the following regression,

$$y = \widehat{F}Q + \varepsilon = \widehat{Y}L_1 Q + \varepsilon. \tag{19}$$

The OLS estimation of the coefficient Q, $r \times 1$ vector, in (19) is given as

$$\widehat{Q} = \left(\widehat{F}'\widehat{F}\right)^{-1}\widehat{F}'y = (L_1'BX'XBL_1)^{-1}L_1'BX'y = \Omega_1^{-1}L_1'BX'y. \tag{20}$$

Therefore, our CFPC forecast is formed as

$$\widehat{y}_{\text{CFPC}} = \widehat{Y}L_1\widehat{Q} = XBL_1\Omega_1^{-1}L_1'BX'y. \tag{21}$$

Remark 1. (Combining forecasts with many forecasts)**:** Although the above analysis is explicitly stated for $\widehat{Y} = XB$, the result is useful when we observe only $\widehat{Y}$ but not X (e.g., Survey of Professional Forecasters). The CFPC forecast, $\widehat{y}_{\text{CFPC}} = \widehat{Y}L_1\Omega_1^{-1}L_1'\widehat{Y}'y$, would then produce a method of combining N forecasts in $\widehat{Y}$ when $N \to \infty$.

Remark 2. The biggest difference between CFPC and PCR lies in the set of variables we use to extract the principal components. In PCR, the principal components are computed from x's directly, without accounting for their relationship with the forecast target variable y. This problem with PCR leads Bai and Ng (2008) to consider first selecting a subset of predictors ("targeted predictors") of x's that are informative in forecasting y, then using the subset to extract factors. In contrast, since CFPC combines forecasts not the predictors, the principal components in CFPC are computed from the set of individual forecasts $(\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N)$ that contain both information on x's and on all past values of y. This actually provides us further intuitions on why CFPC may be more successful than PCR.

Remark 3. Forecasting combination using principal components has been proposed in Chan, Stock, and Watson (1999) and in Stock and Watson (2004), there labeled "principal component forecast combination." We will refer to this approach as CFPC (combining forecasts principal components). The specifications for individual forecasts in CFPC, however, differ from those in Chan et al. (1999) and Stock and Watson (2004) in that individual forecasting models considered here use different and non-overlapping information sets (one regressor at a time), not a common total information set as assumed in Chan et al. (1999) and Stock and Watson (2004).

Remark 4. (Comparison of PCR and CFPC when X has full column rank): Instead of using original predictors X to form principal components, CFPC uses the predicted matrix of y, $\widehat{Y}$. This is where supervision is incorporated. It is interesting to note that there are cases that PCR and CFPC give the same prediction. Note that in case of $N \leq T$ and when X has full column rank, and each column of X is predictive for y ($b_i \neq 0$ for all $i = 1, \ldots, N$), we could exhaust all principal components of X and those of $\widehat{Y}$. Thus we have, from (7),

$$R_1\Lambda_1^{-1}R_1' = (X'X)^{-1}. \tag{22}$$

And also

$$BL_1\Omega_1^{-1}L_1'B = B\left(\widehat{Y}'\widehat{Y}\right)^{-1}B = B(BX'XB)^{-1}B = (X'X)^{-1}, \tag{23}$$

where the last equality follows from the fact that B is also a full rank diagonal matrix. Thus, combining (9), (21), (22) and (23) gives

$$\widehat{y}_{\text{PCR}} = \widehat{y}_{\text{CFPC}}.$$

Therefore, PCR and CFPC are equivalent in this case when X has a full column rank. When X does not have a full column rank, the principal components of the forecasts in CFPC and the principal components of predictors in PCR will differ from each other, because the linear combinations maximizing covariances of forecasts (for which the supervision operates for the relationship between y and X) and the linear combinations maximizing the covariances of predictors (for which there is no supervision) will be different.

Remark 5. (Regression one-at-a-time): CFPC described here employs the regression of y on x_i one-at-a-time to formulate the prediction matrix $\widehat{Y}$. It is simple to implement and computationally appealing. Nevertheless, it can be generalized in various ways.

CFPC also enjoys some theoretical justification as presented in Proposition 1 below. From comparing (1) and (21), CFPC estimates β by $\widehat{\beta}(b) = BL_1\Omega_1^{-1}L_1'BX'y$, which depends on an initial estimator b. To proceed, we define a function $f(\bullet)$ such that

$$\begin{aligned}\widehat{\beta} &= BL_1\Omega_1^{-1}L_1'BX'y \\ &= \text{diag}(b)L_1\Omega_1^{-1}L_1'\text{diag}(b)X'y =: f(b),\end{aligned} \tag{24}$$

where $B = \text{diag}\ (b_1, \dots, b_N)$ and $b = (b_1, \dots, b_N)'$. Note that $\beta\ = (\beta_1, \dots, \beta_N)'$.

We now show that the true parameter β is an asymptotic fixed point for $f(\bullet)$. We first state assumptions:

Assumption 1. (a) The process $\{X_t, y_t\}$ is jointly stationary and ergodic. (b) $E[X_t'(y_t - X_t\beta)] = 0$. (c) β is an interior point of parameter space Θ. (d) Assumptions A-F of Bai (2003) are satisfied for the factor structure (3) with $B = \text{diag}\ (\beta)$. (e) $||\Sigma_{XB}^{-1}\Sigma_E|| = O_p(N/T)$, where Σ_ξ denotes the variance-covariance matrix of ξ, and $||\bullet||$ denotes a matrix norm. (f) $N^2/T \to 0$, as N, $T \to \infty$.

Proposition 1. Under *Assumption 1*, the true parameter β in (1) is an asymptotic fixed point for $f(\bullet)$ defined in (24), that is,

$$(f(\beta) - \beta)_i = O_p\left(\max\left\{\frac{N}{\sqrt{T}}, \frac{N^2}{T}\right\}\right) = o_p(1)\ \textit{for all } i,$$

where a_i denotes the i-th element of a.

The proof is in Appendix C.

Remark 6. (Fixed point): Proposition 1 justifies the construction of the supervision matrix $B = \text{diag}\ (b)$. When we start with $B = \text{diag}\ (b)$ such that b is close to β, CFPC would give an estimate of $\beta, f(b)$, which is close enough to the true value β in the sense of Proposition 1. If one formulates the CFPC alternatively following that of Chan et al. (1999) and Stock and Watson (2004) (as noted in Remark 3), such a fixed point result may not be available. Note further that if $N > T$, i.e., when condition (f) in Assumption 1 fails, the β in regression (1) become unidentified. Therefore, the above fixed point result may not be available and the theoretical property of CFPC remains unknown.

5. Double supervision: supervising factors with variable selection

The previous section looks into supervised factor models from the perspective of supervising the formation of latent factors for a given set of original predictors. Before that step, we can consider selecting a subset of the predictor variables. Boivin and Ng (2006) raise the concern of the quality of data when researchers are ambitious to employ all data available from large panels. Through simulation and application examples, they show that factors extracted from a smaller set of variables are likely to perform no worse, and in many cases even better, in forecasting than those extracted from a lager set of series.

To forecast using a subset of variables when there is too much information has been a popular research topic and many methods have been developed to tackle the issue − see Miller (2002) and Hastie et al. (2009). Variable selection in forecasting in the presence of many predictors is not as simple as in an AR model for which the lags have a natural order. Predictors are not in a natural order. Thus we can not determine which variables should be included and which are not unless we find ways to rank them. We rank the predictors in two ways: hard-thresholding and soft-thresholding.

5.1. Hard-thresholding variable selection

The method of hard-thresholding is to use a statistical test to determine if a particular predictor is significant in forecasting, without considering the effect of other predictors. Bair et al. (2006) take this approach. (Although their model is termed as supervised principal component model, their supervision is in selection of predictors but not in computation of the principal components. Supervision there is only performed via variable selection, but not directly through the factor computation process.) In this paper, lags of y_t are included as regressors with each individual x_{it} to get the individual t-statistic as an indicator of the marginal predictive power of x_{it}, following Bai and Ng (2008). It involves the following steps: For each $i\ = 1,\dots,N$, run the regression of y_{t+h} on a constant, four lags of $\{y_{t-j}\}_{j=0}^{3}$ and x_{it}. Let t_i denote the t-statistic associated with the i-th predictor x_{it}. Select those variables with t_i larger than a threshold value at a given significance level and apply factor models to them. As we show in the empirical application of Section 6, the hard-threshold variable selection plays a critical role in forecasting in the sense that it can substantially reduce MSFE.

5.2. Soft-thresholding variable selection

Hard-thresholding variable selection is highly likely to choose variables similar to each other (so called the "group effect"). In this sense, important information may be lost during the selection process. In contrast to the hard-thresholding which uses a single index to separate qualified predictors from others, soft-thresholding employs more flexible indices to select variables. There are several variable selection methods of this kind, see Tibshirani (1996), Efron et al. (2004), and Zou and Hastie (2005) among many others.

In this paper, we use the least angle regression or LARS of Efron et al. (2004). LARS has gained its popularity in forecasting literature due to its comparative advantages. First, it gives relative ranking of predictors unlike hard-thresholding which gives the marginal predictive power of each predictor. Second, it avoids the group effect. Third, it is very fast and has the same order of computation complexity as OLS.

The LARS (Efron et al., 2004) algorithm proceeds roughly as follows. Like classical forward selection we first find the predictor, say x_{j_1} which is most correlated to the response y. However, instead of taking the largest step in the direction of x_{j_1} as in forward selection, we stop at the point where some other predictor, say x_{j_2}, has as much correlation with the current residual. Instead of continuing along x_{j_1}, LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j_3} makes its way into the "most correlated" set. LARS then proceeds equiangularly between x_{j_1}, x_{j_2} and x_{j_3}, that is, along the "least angle direction," until a fourth variable enters, and so on. Readers interested in LARS are referred to Efron et al. (2004) for detailed description of the algorithm and its satisfactory properties.

In the next section, we apply the LARS algorithm to first select 30 variables, as in Bai and Ng (2008), from the 131 predictors. Then we use the four factor methods, PCR, PLS, PCovR, CFPC, to the 30 variables in forecasting the monthly CPI inflation of U.S.

6. Empirical applications

This section compares the methods described in the previous two sections. Variable of interest to forecast is the logarithm of PUNEW, i.e., CPI all items, using some or all of the 132 monthly time series predictors. Data used are available on Mark Watson's website: http//www.princeton.edu/mwatson. The data range from 1960:1 to 2003:12, with 528 monthly observations in total. These data are transformed by taking logs, first or second differences as suggested in Stock and Watson (2004). Following Stock and Watson (2002b), define

$$y^h_{t+h} := \frac{1200}{h} \cdot (y_{t+h} - y_t) - 1200 \cdot (y_t - y_{t-1}), \tag{25}$$

and

$$z_t := 1200 \cdot (y_t - y_{t-1}) - 1200 \cdot (y_{t-1} - y_{t-2}).$$

For $h = 1, 3, 6, 12, 18, 24, 30$ and 36, we form the factor-augmented forecast as, given information at time t,

$$\widehat{y}_{t+h|t} = \widehat{\alpha}_0 + \widehat{\alpha}'_1(L)z_t + \widehat{\beta}'_1(L)\widehat{f}_t,$$

Here, z_t is the set of lagged variables and $\widehat{f}_t$ latent factors. The number of lags of z_t and $\widehat{f}_t$ are determined by the BIC with the maximum number of lags set to six when the sample size permits, and is reduced to four otherwise. Although we are forecasting the change in inflation, we will continue to refer to the forecasts as inflation forecasts.

As parameter instability is salient in economic time series, we employ two ways to tackle this difficulty in evaluating different forecasting schemes. First, note that for each time period t, the predictors are selected and the forecasting equation is re-estimated after new factors are estimated. We do not restrict the optimal predictors to be the same for every time period. Second, we consider 9 forecast subsamples: 1970:1–1979:12, 1980:1–1989:12, 1990:1–1999:12, 1970:1–1989:12, 1980:1–1999:12, 1970:1–1999:12, 1970:1–2003:12, 1980:1–2003:12, 1990:1–2003:12. For all the forecast subsamples, estimation sample starts at 1960:3 and ends with the data available. For example, for subsample 1970:1–1979:12, the first h-step forecast of 1970:1 is based on estimation up to 1960:3–1970:1-h. Here, 1970:1-h is meant for 1969:2 when $h = 3$, as an example. The last forecast is for 1979:12, and it employs parameters estimated for the sample 1960:3–1979:12-h. That is, recursive scheme is used here, as in Bai and Ng (2008).

MSFE are used to examine the performance of different forecasting procedures. We denote RMSFE as the ratio of the MSFE for a given method relative to the MSFE of PCR model. Therefore, RMSFE less than one means that the specified method outperforms the PCR model in the forecasting practice considered.

Tables 1–8 report RMSFE for each of forecast horizons $h = 1, 3, 6, 12, 18, 24, 30, 36$. Column 1 lists the 9 out-of-sample forecasting subsamples. We report three panels of the RMSFE results depending on whether or how we conduct the variable selection prior to applying the factor models. The first panel of the results reported in Columns 2–5 is for factor models without variable selection, where we use the all 131 predictors to estimate the latent factors for PCR, CFPC, PLS, PCovR. Columns 6–9 and Columns 10–13 present RMSFE for the factor models after selecting the predictors. The second panel reported in Columns 6–9 uses the hard-threshold variable selection at 5% level with the critical value 1.65 for t statistics. To keep in line with Bai and Ng (2008), the third panel of the results reported in Columns 10–13 uses the soft-thresholding variable selection via the LARS algorithm to select 30 variables. Note that PCR without variable selection is used as a benchmark (in each row) in computing the relative MSFEs and thus the values for PCR in Column 2 is 1.000 for all cases.

6.1. *Supervision on computation of factors*

One of the main objectives of this paper is to examine the effect of supervision on the computation of latent factors. The main conclusion over this topic is summarized as below.

(1). Although not reported in the table, the performance of AR(4) is generally as good as AR model with number of lags selected by BIC. However, the predictive ability of the univariate AR model decreases as forecasting horizon increases, reporting larger MSFE as horizon getting larger.
(2). CFPC is better than PCR, no matter variable selection is performed or not. Looking at Column 3 for CFPC from Tables 1–8, for 62 out of 72 cases without variable selection, CFPC reports a RMSFE less than 1. In the case of hard threshold variable selection, 63 out of 72 cases favor CFPC (Column 7 for CFPC from Tables 1–8). Also, the LARS variable selection reports 64 out of 72 cases that are in favor of CFPC over PCR (Column 11 for CFPC from Tables 1–8).
(3). PLS is not doing as well as one might expected. From Table 1, we can see that supervision on factor computation does not make PLS much better than PCR. And it is seen in Tables 1 and 2 that PLS could be very bad and unstable, reporting RMSFE larger than 2. However, as horizon increases, as in Tables 6–8, PLS indeed improves over PCR a lot, reducing RMSFE even below 70%. Variable selection also improves the performance of PLS over PCR, as can be seen in the last two panels of Tables 1–8
(4). PCovR performs better than PCR most of the cases, with 64 out of 72 cases reporting RMSFE lower than 1 without variable selection. Its better predictability is also revealed after variable selection. For example, for $h = 36$, the subsample 90:1–99:12 reports RMSFE of PCovR as 0.187 while that of PCR is 0.834, with hard-thresholding variable selection.

By comparing the RMSFEs in each of the three panels from the tables, we conclude that, the supervision on the computation of factors does improve the predictability of the naive principal component. This improvement is quite substantial as noted above.

6.2. *Supervision on predictors*

Next, let us take a look at the effects of variable selection on the predictability of factor models.

(1). One notable observation from Tables 1–8 is that, variable selection does not make much difference for PCR, with RMSFE closely around 1 most of the cases. This finding is consistent with that reported in Stock and Watson (2002b).
(2). Hard threshold variable selection can make CFPC even better. Most of the cases, hard threshold variable selection reports RMSFE smaller than that without variable selection. To the contrary, more than often, the soft-thresholding LARS variable selection worsens the predictive ability of CFPC.
(3). PLS generally reports lower MSFE when variable selection is carried out in the first step. Hard threshold even makes PLS the best method for several cases. See Table 8 for the second and fourth subsamples for example.

Table 1
RMSFE, $h = 1$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	1.138	2.057	1.024	1.085	1.018	1.167	1.093	1.005	1.078	1.060	**0.985**
80.1–89.12	1.000	**0.863**	1.109	0.931	1.046	0.922	1.060	0.969	1.001	0.927	0.911	0.926
90.1–99.12	1.000	0.994	1.291	1.003	1.054	**0.949**	1.079	0.985	1.008	0.986	1.040	0.993
70.1–89.12	1.000	0.988	1.538	0.973	1.064	0.966	1.108	1.025	1.003	0.996	0.978	**0.953**
80.1–99.12	1.000	**0.898**	1.158	0.951	1.048	0.929	1.065	0.973	1.003	0.943	0.946	0.944
70.1–99.12	1.000	0.989	1.497	0.978	1.062	0.963	1.103	1.018	1.004	0.994	0.989	**0.960**
70.1–03.12	1.000	1.007	1.497	0.979	1.066	0.985	1.136	1.018	1.003	0.998	0.992	**0.964**
80.1–03.12	1.000	**0.943**	1.224	0.957	1.056	0.969	1.121	0.981	1.002	0.959	0.959	0.954
90.1–03.12	1.000	1.057	1.389	**0.993**	1.071	1.035	1.209	0.999	1.004	1.004	1.028	0.994

Table 2
RMSFE, $h = 3$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	1.016	1.492	**0.900**	1.002	1.000	1.027	0.945	1.012	0.943	0.960	0.907
80.1–89.12	1.000	0.904	0.984	0.886	0.973	**0.871**	0.921	0.904	**0.990**	0.892	0.924	0.925
90.1–99.12	1.000	1.055	1.735	1.072	1.126	1.044	1.321	1.066	0.990	1.010	0.991	1.058
70.1–89.12	1.000	0.946	1.176	**0.891**	0.984	0.920	0.961	0.919	0.998	0.911	0.937	0.919
80.1–99.12	1.000	0.934	1.132	0.923	1.003	**0.905**	1.000	0.936	0.990	0.915	0.937	0.952
70.1–99.12	1.000	0.961	1.251	**0.915**	1.003	0.937	1.009	0.939	0.997	0.924	0.944	0.937
70.1–03.12	1.000	0.973	1.256	**0.928**	1.003	0.953	1.021	0.944	0.997	0.931	0.950	0.944
80.1–03.12	1.000	0.955	1.156	0.940	1.003	0.933	1.019	0.943	0.991	**0.926**	0.946	0.959
90.1–03.12	1.000	1.070	1.541	1.060	1.072	1.072	1.238	1.032	**0.993**	1.003	0.997	1.035

Table 3
RMSFE, $h = 6$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	0.902	1.267	0.908	0.922	0.824	**0.806**	0.847	0.983	0.842	0.824	0.829
80.1–89.12	1.000	0.895	1.119	0.937	0.971	**0.842**	0.929	0.886	1.001	1.005	1.056	0.953
90.1–99.12	1.000	0.964	1.656	0.996	1.068	0.975	1.374	**0.955**	0.985	0.971	1.027	1.006
70.1–89.12	1.000	0.898	1.180	0.925	0.951	**0.835**	0.878	0.870	0.994	0.938	0.960	0.902
80.1–99.12	1.000	0.910	1.229	0.949	0.991	**0.870**	1.021	0.900	0.998	0.998	1.050	0.964
70.1–99.12	1.000	0.907	1.243	0.935	0.966	**0.853**	0.944	0.881	0.993	0.942	0.969	0.916
70.1–03.12	1.000	0.927	1.261	0.959	0.972	**0.874**	0.977	0.893	0.993	0.954	0.984	0.932
80.1–03.12	1.000	0.939	1.258	0.984	0.998	**0.899**	1.062	0.915	0.998	1.010	1.064	0.983
90.1–03.12	1.000	1.045	1.600	1.098	1.063	1.036	1.387	**0.986**	0.989	1.023	1.085	1.055

Table 4
RMSFE, $h = 12$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	0.762	1.281	0.807	0.960	0.738	0.886	0.734	1.014	**0.716**	0.757	0.727
80.1–89.12	1.000	0.797	1.008	0.843	1.042	**0.787**	0.845	0.868	1.005	0.967	1.114	0.867
90.1–99.12	1.000	**0.891**	1.415	0.980	1.101	0.906	1.381	0.924	0.976	0.936	0.979	1.015
70.1–89.12	1.000	0.782	1.128	0.827	1.006	**0.765**	0.863	0.808	1.009	0.856	0.956	0.805
80.1–99.12	1.000	0.817	1.095	0.872	1.054	**0.812**	0.959	0.880	0.998	0.961	1.085	0.899
70.1–99.12	1.000	0.796	1.166	0.847	1.018	**0.784**	0.931	0.824	1.004	0.867	0.959	0.833
70.1–03.12	1.000	0.822	1.205	0.883	1.034	**0.811**	0.980	0.855	1.004	0.883	0.979	0.859
80.1–03.12	1.000	0.857	1.162	0.926	1.076	**0.854**	1.033	0.924	0.998	0.978	1.106	0.934
90.1–03.12	1.000	1.013	1.567	1.143	1.167	1.029	1.528	1.072	**0.979**	1.007	1.084	1.111

(4). For PCovR, the LARS variable selection makes it the best for several subsample when $h = 1$. For all other cases, hard threshold works better as a variable selection procedure to improve the performance of PCovR.

6.3. *Effects of "double supervision"*

Double supervision includes supervision for the variable-selection as in Bair et al. (2006), and also the supervision of the factor computation. It would be interesting to see that the above two parts on supervision leads to the essence of this paper. The RMSFE reported for factor models after supervision on the computation of factors and also the selection of variable are generally lower than 1, as can be seen in the last two panels of Tables 1–8 Exception to this conclusion is for PLS with short forecasting horizons. In most of the cases, the reduction of MSFE relative to PCR is clearly noticeable. After variable selection, CFPC reports RMSFE as low as 40% in a lot of cases. PCovR can reduce RMSFE to be as low as 18.7%. The findings affirm the conjecture raised in Section 1 that the *double* supervision in selection of predictors and formation of latent factors should be carried out in forecasting practice.

Table 5
RMSFE, $h = 18$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	**0.596**	1.083	0.673	0.895	0.629	0.793	0.656	1.021	0.630	0.633	0.688
80.1–89.12	1.000	0.707	0.878	0.701	1.026	**0.686**	0.765	0.701	1.028	0.760	0.895	0.710
90.1–99.12	1.000	1.084	1.527	1.012	1.313	1.055	1.509	0.991	0.974	**0.907**	1.007	1.047
70.1–89.12	1.000	**0.657**	0.971	0.688	0.966	0.660	0.778	0.681	1.025	0.701	0.776	0.700
80.1–99.12	1.000	0.767	0.981	0.750	1.071	**0.744**	0.883	0.747	1.019	0.783	0.913	0.763
70.1–99.12	1.000	**0.696**	1.023	0.718	0.998	0.697	0.846	0.710	1.020	0.720	0.797	0.732
70.1–03.12	1.000	**0.722**	1.056	0.750	1.023	**0.722**	0.886	0.735	1.018	0.746	0.820	0.765
80.1–03.12	1.000	0.804	1.039	0.800	1.106	**0.783**	0.947	0.786	1.016	0.821	0.942	0.816
90.1–03.12	1.000	1.145	1.605	1.147	1.389	1.126	1.588	1.086	**0.973**	1.038	1.108	1.187

Table 6
RMSFE, $h = 24$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	**0.524**	0.951	0.625	0.890	0.561	0.632	0.610	0.975	0.623	0.597	0.672
80.1–89.12	1.000	0.794	0.867	0.834	1.078	0.773	**0.718**	0.837	1.014	0.849	0.850	0.849
90.1–99.12	1.000	0.450	0.881	0.347	0.868	0.444	0.775	0.324	0.957	0.346	0.933	**0.316**
70.1–89.12	1.000	**0.667**	0.907	0.735	0.989	0.673	0.678	0.730	0.995	0.743	0.731	0.766
80.1–99.12	1.000	0.684	0.871	0.678	1.011	**0.668**	0.736	0.674	0.996	0.689	0.876	0.679
70.1–99.12	1.000	**0.624**	0.901	0.658	0.965	0.628	0.697	0.650	0.988	0.664	0.771	0.676
70.1–03.12	1.000	**0.644**	0.927	0.675	0.973	0.647	0.730	0.668	0.991	0.687	0.784	0.700
80.1–03.12	1.000	0.711	0.914	0.704	1.019	**0.695**	0.785	0.700	1.000	0.723	0.889	0.716
90.1–03.12	1.000	0.569	0.994	0.482	0.919	0.563	0.898	**0.466**	0.976	0.507	0.957	0.488

Table 7
RMSFE, $h = 30$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	**0.513**	0.934	0.672	0.911	0.539	0.619	0.579	0.986	0.612	0.592	0.625
80.1–89.12	1.000	0.761	0.809	0.844	1.029	0.760	**0.704**	0.832	1.002	0.842	0.862	0.822
90.1–99.12	1.000	0.343	0.765	0.285	0.831	0.343	0.654	**0.240**	1.020	0.308	0.707	0.290
70.1–89.12	1.000	**0.658**	0.861	0.773	0.980	0.668	0.669	0.727	0.995	0.746	0.750	0.740
80.1–99.12	1.000	0.620	0.794	0.656	0.962	**0.619**	0.687	0.632	1.008	0.662	0.810	0.643
70.1–99.12	1.000	**0.586**	0.839	0.661	0.946	0.593	0.666	0.615	1.001	0.646	0.740	0.637
70.1–03.12	1.000	**0.600**	0.864	0.673	0.953	0.607	0.684	0.629	1.001	0.668	0.753	0.663
80.1–03.12	1.000	0.639	0.833	0.673	0.971	**0.637**	0.713	0.651	1.008	0.693	0.824	0.680
90.1–03.12	1.000	0.435	0.874	0.389	0.875	0.434	0.727	**0.351**	1.018	0.445	0.761	0.444

6.4. Supervision and forecasting horizon

The effect of supervision over forecasting horizons h is very clear, which can be seen by comparing the results across the eight tables. From examining the RMSFE numbers as a function of the eight values of forecast horizons $h = 1, 3, 6, 12, 18, 24, 30, 36$, it can be clearly seen that the RMSFEs are generally decreasing with h for the three supervised factor models. That is, the superiority of supervised factor models is getting more and more significant as the forecasting horizon increases. On the other hand, the unsupervised factor model, PCR, has RMSFEs moving up and down over the horizons with no pattern over forecasting horizon h.

Note that the forecast target variable y^h_{t+h} defined in (25) is the average monthly changes over the h months, and it may be easier to forecast when forecasting horizon h is longer as it becomes smoother. The three supervised factor models are able to capture this feature in y^h_{t+h} while PCR fails to do so. We also observe (although not reported for space) that neither AR(4) or AR models with number of lags selected by BIC capture this feature. This is seen from the RMSFE values for these univariate models, which are generally increasing over the forecasting horizons. Hence, it seems that richer information from

Table 8
RMSFE, $h = 36$.

SAMPLE	NO				HARD THRESHOLD				LARS(30)			
	VARIABLE SELECTION				VARIABLE SELECTION				VARIABLE SELECTION			
	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR	PCR	CFPC	PLS	PCovR
70.1–79.12	1.000	0.546	0.847	0.715	0.938	**0.501**	0.505	0.616	1.039	0.668	0.594	0.667
80.1–89.12	1.000	0.713	0.781	0.834	1.049	0.692	**0.634**	0.830	0.995	0.782	0.845	0.819
90.1–99.12	1.000	0.280	0.612	0.238	0.834	0.279	0.523	**0.187**	1.021	0.261	0.506	0.287
70.1–89.12	1.000	0.643	0.808	0.784	1.003	0.613	**0.580**	0.741	1.013	0.735	0.740	0.756
80.1–99.12	1.000	0.547	0.716	0.605	0.967	**0.534**	0.591	0.583	1.005	0.582	0.715	0.614
70.1–99.12	1.000	0.546	0.756	0.639	0.958	**0.524**	0.565	0.593	1.015	0.608	0.678	0.631
70.1–03.12	1.000	0.561	0.768	0.650	0.962	**0.539**	0.581	0.605	1.015	0.628	0.690	0.645
80.1–03.12	1.000	0.567	0.735	0.622	0.972	**0.555**	0.613	0.600	1.005	0.610	0.731	0.636
90.1–03.12	1.000	0.361	0.670	0.324	0.863	0.360	0.583	**0.275**	1.019	0.368	0.570	0.377

multivariate environment benefits the factor models even more especially for longer forecast horizons when they are supervised on the selection of the variables and on the computation of their latent factors.

6.5. Supervision and number of factors

Another important finding of this paper (not reported) is that, for supervised factor models, the number of factors selected by BIC is less than that of PCR. This finding also favors the previous result that, with supervision, factor models tend to form better latent variables and thus need less indices to describe “the state of the economy”, as termed in Heij et al. (2007). They report the result for PCovR and this paper validates their conclusion for PLS and CFPC.

7. Conclusions

In this paper we emphasize the importance of supervision in choosing and estimating the factors, such that they are supervised to target the variable to be forecast rather than simply represent the set of regressors without regard for the variable to be forecast. We discuss the construction of forecasts when factors are extracted such that they are targeting the forecasting variable. Three possible methods are discussed: partial least square regression, principal covariate regression, and combining forecast principal components.

In exploiting high dimensional information from large number of predictors we wish to improve efficiency of a forecast and to enhance the robustness of a forecast. This paper compares the forecasting performance of factor models in such data-rich environment. Our findings suggest that one can profit from supervising the computation of factors.

Computation of latent factors may be doubly supervised with variable selection. Variable selection is generally useful for the supervised factor models. Interestingly, the effect of supervision gets even larger as forecast horizon increases and the supervision also helps a factor model achieving more parsimonious factor structure. Among the supervised factor models compared in this paper, CFPC stands out for its superiority in predictive ability and its stability in performance. In general, the CFPC model generates most efficient and robust forecasts.

Acknowledgements

The authors thank the Editor and two anonymous referees for helpful comments. Tu would like to thank support from National Natural Science Foundation of China (Grant 71301004, 71472007, 71532001, 71671002), China’s National Key Research Special Program (2016YFC0207705), the Center for Statistical Science at Peking University, and Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education.

Conflicts of interest

The authors declare no conflict of interest.

Appendix

A NIPALS Algorithm for PCR

The intuition behind the working of the nonlinear iterative algorithm for PCR goes as follows. Formally,

$$\begin{aligned} &E_1 = X - f_1 p_1',\ \ E_2 = E_1 - f_2 p_2',\ \ \ldots \\ &E_h = E_{h-1} - f_h p_h',\ \ \ldots\ \ E_r = E_{r-1} - f_r p_r'. \end{aligned} \tag{A.1}$$

The NIPALS follows the steps for the computation of f_h:

1. Take a vector x_J from X and call it f_h:
2. Normalize f_h: $f_h' = f_h' / \|f_h'\|$
3. Calculate p_h':

$$p_h' = f_h' X \tag{A.2}$$

4. Normalize p_h' : $p_h' = p_h' / ||p_h'||$
5. Calculate f_h:

$$f_h = X p_h \tag{A.3}$$

6. Compare f_h in step 2 with that obtained in step 5. If they are the same, stop. Otherwise go to step 2.

Note that the evolution of p_h' and f_h are described by (A.2) and (A.3). Substitute (A.3) into (A.2), we have

$$c p_h' = (X p_h)' X, \tag{A.4}$$

where c is a constant that accounts for the normalization in step 4. This is equivalent to

$$0 = (X'X - c I_r) p_h. \tag{A.5}$$

This is exactly the eigenvalue/eigenvector equation for $X'X$ in PCR. Hence, the NIPALS algorithm gives the same principal components as derived by eigenvalue decomposition.

B NIPALS Algorithm for PLS

For the X block: (1) take u_{start} = some y_J (instead of some x_J); (2) normalize u: $u = u/||u||$; (3) $p' = u'X$; (4) normalize p': $p' = p'/||p'||$; (5) $f = Xp$.

For the y block: (6) $q = f$ (instead of some y_S); (7) normalize q: $q = q/||q||$; (8) $u' = y'q$; (9) normalize u': $u' = u'/||u'||$; (10) compare f in step 5 with that in the preceding iteration step. If they are equal (up to a tolerance level) then stop; otherwise go to step 2.

By exchanging scores in step 1 and 6, the above algorithm supervises the computation of the x-score thus should improve the predictability of PLS over PCR. For the purpose of prediction, we can rewrite (12) as

$$E_h = E_{h-1} - f_h p_h';\ \ X = E_0,$$

$$G_h = G_{h-1} - u_h q_h';\ \ y = G_0,$$

and a mixed relation is available as

$$G_h = G_{h-1} - \gamma_h f_h q_h',$$

where $\gamma_h = (u_h' f_h)/(f_h{}' f_h)$. Therefore,

$$\hat{y} = \sum \hat{u}_h q_h' = \sum \gamma_h f_h q_h' = F \Gamma Q', \tag{B.1}$$

where $\Gamma = \text{diag}\ (\gamma_1, \ldots, \gamma_r)$.

Note that the x-score extracted in the h^{th} iteration, f_h, is a linear combination of E_{h-1}, instead of as a direct function of original data matrix X. de Jong (1993) gives a direct relationship as $F = XR \equiv XW(P'W)^{-1}$, where $P = [p_1, \ldots, p_h]$ and $W = [E_0 u_1, \ldots, E_{h-1} u_h]$. Thus, (B.1) can be used for prediction as

$$\widehat{y}_{\text{PLS}} = XR\Gamma Q'. \tag{B.2}$$

That is, we have $R = W(P'W)^{-1}$, $U = F\Gamma$ for the linear factor model framework, while β in (1) is estimated by $\widehat{\beta} = R\Gamma Q'$.

C Proof of Proposition 1

Rewrite (3) as

$$XB = C + E, \tag{C.1}$$

where $C = FP'$, is the common component of XB. Note that C is estimated using principal component method as

$$\begin{aligned}\tilde{C} &= \widehat{F}\widehat{P}' \\ &= XBL_1(L_1'BX'XBL_1)^{-1}L_1'BX'XB \\ &= XBL_1\Omega_1^{-1}L_1'BX'XB.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\tilde{C}'\tilde{C} &= BX'XBL_1\Omega_1^{-1}L_1'BX'XBL_1\Omega_1^{-1}L_1'BX'XB \\ &= BX'XBL_1\Omega_1^{-1}L_1'BX'XB,\end{aligned}$$

which leads to

$$\begin{aligned}J &\equiv B(X'X/T)BL_1\Omega_1^{-1}L_1'BX'XB - (B\Sigma_X B - \Sigma_E) \\ &= B(X'X/T)BL_1\Omega_1^{-1}L_1'BX'XB - (\Sigma_{XB} - \Sigma_E) \\ &= B(X'X/T)BL_1\Omega_1^{-1}L_1'BX'XB - \Sigma_C \\ &= \left(\tilde{C}'\tilde{C}/T\right) - \Sigma_C \\ &= \frac{1}{T}\left(\tilde{C}'\tilde{C} - \tilde{C}'C + \tilde{C}'C - C'C\right) + \left(\frac{1}{T}C'C - \Sigma_C\right) \\ &= \frac{1}{T}\tilde{C}'\left(\tilde{C} - C\right) + \frac{1}{T}\left(\tilde{C} - C\right)'C + \left(\frac{1}{T}C'C - \Sigma_C\right) \\ &\equiv \psi^1 + \psi^2 + \psi^3.\end{aligned} \tag{C.2}$$

Note that

$$\psi_{ij}^1 = \frac{1}{T}\sum_{t=1}^{T}\tilde{C}_{ti}\left(\tilde{C}_{tj} - C_{tj}\right) = O_p\left(\frac{1}{\sqrt{NT}}\right). \tag{C.3}$$

which follows from Theorem 3 of Bai (2003) that $\tilde{C}_{it} - C_{it} = O_p(1/\sqrt{N})$ under Assumption 1.d. Similarly, we have

$$\psi_{ij}^2 = O_p\left(\frac{1}{\sqrt{NT}}\right). \tag{C.4}$$

By Assumption 1.a, we have

$$\psi_{ij}^3 = O_p\left(\frac{1}{\sqrt{T}}\right). \tag{C.5}$$

(C.3), (C.4) and (C.5) lead to

$$J_{ij} = O_p\left(\frac{1}{\sqrt{NT}}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right) + O_p\left(\frac{1}{\sqrt{T}}\right) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Note that (C.2) is equivalent to

$$J = B\Sigma_X\Big[BL_1\Omega_1^{-1}L_1'B'X - I_N - \tag{C.6}$$

$$(\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}\Big]B + o_p(1)$$

$$\equiv B\Sigma_X HB + o_p(1)$$

where

$$H = \left(BL_1\Omega_1^{-1}L_1'BX'X - I_N - (\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}\right)$$

where $\varepsilon_{NT}, \zeta_{NT}$ are sequences of small positive numbers such that $\varepsilon_{NT}, \zeta_{NT} \to 0$ as $N, T \to \infty$. ε_{NT} and ζ_{NT} are introduced to guarantee the matrix inverse exists. To see (C.6), note that the last term in the bracket of the right hand

$$\begin{aligned}
&B\Sigma_X(\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B \\
&= B(\Sigma_X + \varepsilon_{NT}I_N)(\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B \ - B\varepsilon_{NT}(\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B \\
&= B(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B + o_p(1) \\
&= (B + \zeta_{NT}I_N)(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B \quad - \zeta_{NT}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1}B + o_p(1) \\
&= \Sigma_E(B + \zeta_{NT}I_N)^{-1}B + o_p(1) \\
&= \Sigma_E(B + \zeta_{NT}I_N)^{-1}(B + \zeta_{NT}I_N) \quad - \Sigma_E(B + \zeta_{NT}I_N)^{-1}\zeta_{NT}I_N + o_p(1) \\
&= \Sigma_E + o_p(1) \quad \text{as } \varepsilon_{NT} = o_p(1) \text{ as } N, T \to \infty.
\end{aligned}$$

Note that (C.6) is true for all values of β. Therefore, it must be the case that

$$H_{ij} = O\left(J_{ij}\right) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Define

$$K = BL_1\Omega_1^{-1}L_1'BX'X - I_N.$$

We have

$$\begin{aligned}
\mathrm{K} &= H + (\Sigma_X + \varepsilon_{NT}I_N)^{-1}(B + \zeta_{NT}I_N)^{-1}\Sigma_E(B + \zeta_{NT}I_N)^{-1} \\
&= H + \Sigma_{XB}^{-1}\Sigma_E + o_p(1),
\end{aligned}$$

i.e.,

$$\begin{aligned}
K_{ij} &= H_{ij} + O_p(N/T) \quad (\text{by} \quad \text{Assumption} \quad 1.\text{e}) \\
&= O_p\left(\max\left\{\frac{1}{\sqrt{T}}, \frac{N}{T}\right\}\right).
\end{aligned} \tag{C.7}$$

By definition of $f(\beta)$,

$$\begin{aligned}
f(\beta) - \beta &= \text{diag}(\beta)L_1\Omega_1^{-1}L_1'\text{diag}(\beta)X'y - \beta \\
&= BL_1\Omega_1^{-1}L_1'BX'y - \beta \\
&= BL_1\Omega_1^{-1}L_1'BX'(X\beta + e) - \beta \\
&= \left(BL_1\Omega_1^{-1}L_1'BX'X - I_N\right)\beta + o_p(1) \quad (\text{by} \quad \text{Assumption} \quad 1.\text{b}) \\
&= K\beta + o_p(1),
\end{aligned}$$

and it follows from (C.7) that

$$(f(\beta) - \beta)_i = O_p\left(\max\left\{\frac{N}{\sqrt{T}}, \frac{N^2}{T}\right\}\right). \tag{C.8}$$

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica, 71*(1), 135–171.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics, 146*, 304–317.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association, 101*(473), 119–137.

Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis. *Journal of Econometrics, 132*, 169–194.

Butler, N. A., & Denham, M. C. (2000). The peculiar shrinkage properties of PLS. *Journal of the Royal Statistical Society B, 62*, 585–593.

Chan, Y. L., Stock, J. H., & Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review, 1*, 91–121.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics, 32*(2), 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Garthwait, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association, 89*(425), 122–127.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17.

Groen, J. J. J., & Kapetanios, G. (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis, 100*, 221–239.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning – data mining, inference, and prediction* (2nd ed.). Springer.

Heij, C., Groenen, P. J. F., & van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariate regression. *Computational Statistics & Data Analysis, 51*, 3612–3625.

Huang, J., Horowitz, J. L., & Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics, 36*, 587–613.

de Jong, S. (1992). Principal covariate regression part I. theory. *Chemometrics and Intelligent Laboratory Systems, 14*, 155–164.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems, 18*, 251–261.

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics, 186*(2), 294–316.

Lingjaerde, O. C., & Christophersen, N. (2000). Shrinkage structure of partial least squares. *Scandinavian Journal of Statistics, 27*(3), 459–473.

Mardia, K., Kent, J., & Bibby, J. (1980). *Multivariate analysis*. London: Academic Press.

Miller, A. (2002). *Subset selection in regression*. Chapman & Hall/CRC.

Otto, M., & Wegscheider, W. (1985). Spectrophotometric multicomponent applied to trace metal determinations. *Analytical Chemistry, 57*, 63–69.

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

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics, 20*, 147–162.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*, 405–430.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiaah (Ed.), *Multivariate analysis* (pp. 391–420). New York: Academic Press.

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares approach. In M. S. Bartlett, & J. Gani (Eds.), *Perspectives in probability and statistics*. London: Academic Press. Papers in Honour of.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J., III (1984). The collinearity problem in linear regression, the partial least squares approach to generalized inverse. *SIAM Journal on Scientific and Statistical Computing, 5*, 735–743.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics, 37*(4), 1773-1751.