In-sample Inference and Forecasting in Misspecified Factor Models

Marine Carrasco*
Université de Montréal
CIREQ, CIRANO

Barbara Rossi†
ICREA-Univ. Pompeu Fabra,
Barcelona GSE and CREI

April 28, 2016

Abstract

This paper considers in-sample prediction and out-of-sample forecasting in regressions with many exogenous predictors. We consider four dimension reduction devices: principal components, Ridge, Landweber Fridman, and Partial Least Squares. We derive rates of convergence for two representative models: an ill-posed model and an approximate factor model. The theory is developed for a large cross-section and a large time-series. As all these methods depend on a tuning parameter to be selected, we also propose data-driven selection methods based on cross-validation and establish their optimality. Monte Carlo simulations and an empirical application to forecasting inflation and output growth in the U.S. show that data-reduction methods outperform conventional methods in several relevant settings, and might effectively guard against instabilities in predictors' forecasting ability.

Keywords: Forecasting, regularization methods, factor models, Ridge, partial least squares, principal components, sparsity, large datasets, variable selection, GDP forecasts, inflation forecasts.

J.E.L. Codes: C22, C52, C53.[1]

*Address: Université de Montréal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada. E-mail: marine.carrasco@umontreal.ca.

†Address: Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: barbara.rossi@upf.edu

1 Introduction

In this paper, we consider a regression with a large dimensional set of exogenous predictors and we are concerned with in-sample prediction and out-of-sample forecasting of the dependent variable. That is, the set of potential predictors X has dimension $(T \times N)$, where N is the cross-section dimension, T is the time-series dimension, and N is large relative to T. We consider four dimension reduction devices: principal components, Ridge, Landweber Fridman and Partial Least Squares. Principal components (PC) go back to Amemiya (1966) and are used in the factor model literature, see e.g. Stock and Watson (2002a) and Bai and Ng (2002). Ridge regularization has its origin in the work by Tikhonov (see Tikhonov and Arsenin, 1977). Landweber Fridman (LF) regularization is a classical method in inverse problem literature, see Kress (1999) and Carrasco, Florens and Renault (2007). Partial Least Squares (PLS) was introduced by Herman and Svante Wold; see Helland (1988) and references therein; it has been used recently in econometrics by Groen and Kapetanios (2008) and Kelly and Pruitt (2015).

All these methods involve a regularization or tuning parameter which needs to be selected. For example, in the principal components approach, the regularization parameter is the number of principal components used for prediction. An appropriate choice of the regularization parameter is crucial. We propose a data-driven selection method based on generalized cross-validation (GCV; Li, 1986, 1987). Note that our criterion differs from that used in traditional factor models, where the number of factors is selected using Bai and Ng's (2002) criteria, for two reasons: first, because we do not impose any factor structure on the data; second, because the GCV criterion minimizes the prediction error in forecasting the target variables as opposed to explaining the variability in the regressors. In fact, we show that our criterion will perform better than Bai and Ng's (2002) criterion when there is no factor model or when the factors which are the most relevant for explaining the variance of the regressors are not the ones which are the most relevant for forecasting the dependent variable.

We study properties of our regularized estimators assuming two representative data generating processes (DGP). The first DGP is referred to as a ill-posed model where the eigenvalues of the matrix $X'X/T$ are bounded and decline to zero gradually. The second DGP is the popular factor

[1] *Acknowledgements:* This paper was prepared for the *Journal of Business & Economic Statistics* invited lecture at the 2016 ASSA-Econometric Society Meetings. We are grateful to X. Cheng, T. Clark, D. Giannone, B. Hansen, J. Stock and N. Swanson for their detailed comments. We would also like to thank S. Pruitt for sharing his codes and G. Kapetanios, J-P. Florens, B. Perron, T. Proietti and seminar participants at the SoFiE Conference on Large-Scale Factor Models in Finance, the 2016 Econometric Society Meetings, the 2015 IAAE Conference, the 2015 Econometric Society World Congress, the Universities of Bocconi, Padova and Navarra, Bank of France and Norges Bank for comments. M. Carrasco gratefully acknowledges partial financial support from SSHRC. B. Rossi gratefully acknowledges financial support from the ERC Grant #615608 and the Spanish Ministry of Economy and Competitiveness Grant ECO2012-33247.

model with a finite number, r, of factors. In this model, the r largest eigenvalues of $X'X/T$ grow with N while the remaining eigenvalues are bounded. In both DGP, the matrix $X'X/T$ is ill-conditioned in the sense that the ratio of the largest over the smallest eigenvalue diverges and regularization is needed to invert $X'X/T$. We found that the convergence rates of our estimators are quite different in the two DGPs. While, in the first DGP, the three main regularization methods (Ridge, principal components, and Landweber Fridman) achieve the same rate in most cases, in the second DGP, principal component and LF have a much faster rate than Ridge, which is due to the factor structure. These are asymptotic results for large N and large T. When comparing PLS and principal components in the first DGP, it is instead unclear which one is faster since there is a bias and variance trade-off. We use our findings to draw useful guidelines for practitioners.

Simulations show that the dimension reduction methods that we consider outperform traditional factor models in many relevant settings. We find substantial forecasting gains, in particular, when there are a large number of relevant factors or when the predictors have a factor structure but the factors are not related to the variable to be predicted. In both cases, the assumptions underlying the usual factor models break down, either because the number of factors is not "small" or because the factors are irrelevant for predicting the target variable (even though they can effectively summarize the information contained in the predictors).

Our paper is related to several contributions in the literature on forecasting using a large dataset of predictors. When forecasting with a large dataset of predictors, estimating the parameters by OLS has several drawbacks. A first issue is that, even if all predictors are relevant, the variance of the mean-square forecast error is increasing in the dimension of the predictors; e.g., under Gaussianity assumptions, Stock and Watson (2006) show that the distribution of the forecast given the predictors is Normal, with a variance proportional to the size of the cross section dimension divided by the size of the time series dimension: as the number of predictors grows, so does the variance. Another issue is that most predictors are highly correlated, which makes inverting the matrix of the second moments of the predictors very imprecise. A final issue is that not every potentially important predictor is actually relevant: keeping weak predictors can introduce unnecessary sampling variability to the forecast.

To improve upon OLS while continuing to enjoy the benefits of extracting information from a large dataset of predictors, the literature has moved in the direction of either summarizing the information from the large dataset of predictors into a low-dimensional vector of latent factors, or using all the variables but imposing some kind of shrinkage.

The former route (that is, summarizing the information in a few latent factors) has led to the popular factor models (Stock and Watson, 2002b; Forni et al., 2005), where the number of factors has been typically estimated by Bai and Ng's (2002) information criteria (IC). Stock and Watson (2002a) perform Monte Carlo simulations to compare the forecasting performance of factor models

where the number of factors is selected via AIC and BIC with that of factor models where the number of factors is selected via Bai and Ng's (2002) criteria. They find that the relative performance depends on the DGP. Our contribution to this literature is to consider information criteria and data reduction models other than Stock and Watson (2002a), as well as their performance in more general DGPs; in addition, we study their theoretical properties and empirical performance. Alternative ways to select factors for forecasting have been proposed by Bai and Ng (2008) and Cheng and Hansen (2015). Bai and Ng (2008) propose a targeted predictors approach which first selects subsets of regressors based on individual t-tests, then extract principal components from this subset.[2] In this paper we also consider factor models, but estimating the number of factors using generalized cross-validation; also, differently from the factor model literature, we consider methodologies that extract information from all predictors using dimension reduction techniques. Cheng and Hansen (2015) propose to select factors and lag structures for forecasting using Mallows' (1973) IC and leave-one-out cross-validation. Relative to Cheng and Hansen (2015), we do not assume a factor structure and compare the GCV criterion with Mallows' criterion in Monte Carlo simulations.[3]

The second route (that is, using all the predictors but imposing some shrinkage) has led to considering several distinct procedures. One avenue is to use Bayesian VARs; De Mol et al. (2008) show that, with specific choice of priors, Bayesian VARs reduce to penalized Ridge (when using a Gaussian prior, which gives decreasing weights to ordered eigenvalues of PC) or Lasso (when using a double exponential prior, which gives zero weight to variables whose coefficients are small, thus enforcing sparsity). We contribute to this strand of the literature by considering an alternative method to select the penalization in Ridge (via generalized cross-validation). We also include Lasso in some of our Monte Carlo simulation results.

A second avenue is to consider forecast combinations, either using equal weight or Bayesian model averaging (Wright, 2009). Cheng and Hansen (2015) also propose forecast combinations for factor models based on Mallows' criterion. Differently from these contributions, we investigate generalized cross-validation criteria for factor model selection as well as other dimension reduction techniques applicable in more general models. We also compare our techniques with BMA and equal weight forecast combinations in the empirical analysis.[4]

[2] Giovannelli and Proietti (2015) propose an alternative way to choose the factors taking into account their correlation with the target variable, and controlling for the error rate of the selection. See also Huang and Lee (2010) for supervised factor models.

[3] Djogbenou (2015) also shows that leave-d-out cross-validation and a bootstrap method he proposes consistently select the number of factors in factor-augmented models. Mao and Stevanovic (2014) show that the selection of factors is very sensitive to the presence of structural instability. Gonçalves, McCracken, and Perron (2015) investigate the predictive ability of factor-augmented models.

[4] Other techniques that have been considered in the literature include Bagging (e.g. Inoue and Kilian, 2008), Boosting (e.g. Bai and Ng, 2009), combinations of OLS and PC (e.g. Hillebrand and Lee, 2012), independent

An alternative avenue is to consider dimension reduction techniques, like we do. A few recent contributions have investigated PLS. Groen and Kapetanios (2008) compare three regularization methods (principal components, Ridge and PLS) and study their properties when the model has a weak factor structure. Kelly and Pruitt (2015) propose a new forecasting method based on the use of proxies; when the choice of proxies is automatic, their method boils down to PLS. The main contributions of our paper relative to these works are that: (a) we consider a broader class of methods including Landweber Fridman which has never been applied to forecasting so far; (b) while these papers do not give any indication on how to select the tuning parameters in practice, we provide a data-driven method for selecting the regularization parameter, making our approach easily applicable by practitioners. Moreover, the methods we consider do not require proxies. Finally, the data reduction methods we consider, like Kelly and Pruitt's (2015) method, have superior performance when "the factors dominating the forecast target's variation contribute only weakly to variance among the predictors."[5]

The remainder of the paper is organized as follows. Section 2 presents the four estimation methods. Section 3 discusses the rates of convergence of the estimators in two different models: a ill-posed model and a factor model. Section 4 presents data-driven methods for selecting the tuning parameter involved in the regularization methods we consider, and establishes their optimality. Section 5 presents Monte Carlo experiments results and Section 6 presents the empirical results. Section 7 concludes. The proofs are collected in the Not-for-Publication Appendix (Carrasco and Rossi, 2016).

2 Estimation Methods

The model is:

$$y_t = x_t'\delta + \varepsilon_t,\ t = 1, 2, ..., T,$$

where y_t is a scalar, x_t is a $(N \times 1)$ vector of predictors, and δ is $(N \times 1)$ vector of unknown parameters. In matrix notation, let y be a $(T \times 1)$ vector, X be a $(T \times N)$ matrix and ε be a $(T \times 1)$

component analysis and sparse PCA (Kim and Swanson, 2014b), combining forecast PC (Huang and Lee, 2010), principal covariate regression (e.g. Tu and Lee, 2013). Stock and Watson (2012) show that IC, BMA and Bagging produce forecasts equal to a weighted average of the predictors, where the weights are the OLS coefficients times a shrinkage factor that depends on t-statistic of that coefficient. Kim and Swanson (2014a) provide a comprehensive empirical analysis of the performance of several of these methods. Kim and Swanson (2015) discuss factor MIDAS approaches for forecasting.

[5] Tu and Lee (2013), Barbarino (2014) and Fuentes, Poncela and Rodrigues (2015) also empirically investigate the forecasting performance of PLS. Kelly and Pruitt (2013) show that a single-factor PLS is successful to forecast stock market returns.

vector such that:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \; X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

Let $\Sigma_{xx} = E(x_t x_t')$, $\Sigma_{xy} = E(x_t y_t)$, $S_{xx} = X'X/T$, and $S_{xy} = X'y/T$.

The estimation of δ can be viewed as solving the equation $y = X\delta$, i.e. $S_{xy} = S_{xx}\delta$. When S_{xx} is invertible, the OLS estimator of δ is:

$$\hat{\delta} = (S_{xx})^{-1} S_{xy}. \tag{1}$$

This estimate involves the inversion of the $(N \times N)$ matrix S_{xx}. However, when N is large, $X'X$ may be ill-conditioned so that the resulting $\hat{\delta}$ may have a large variance. Moreover, if $N > T$, this estimate is not implementable. The population OLS estimator is $\delta = \Sigma_{xx}^{-1}\Sigma_{xy}$. A good estimator of δ should rely on a good estimator of Σ_{xx}^{-1}. The data-reduction estimators we consider in this paper involve a regularized inverse of S_{xx} denoted $(S_{xx}^{\alpha})^{-1}$, which depends on the regularization parameter α.

Before introducing the various regularization techniques, it is useful to recast the problem as an inverse problem. Let $\mathcal{H}$ be the Hilbert space corresponding to $\mathbb{R}^N$ endowed with the norm $\|u\|^2 = u'u$ with associated inner product $\langle u_1, u_2 \rangle = u_1' u_2$. Let $\mathcal{E}$ be the Hilbert space corresponding to $\mathbb{R}^T$ endowed with the norm $\|v\|_T^2 = v'v/T$ with associated inner product $\langle v_1, v_2 \rangle = v_1' v_2/T$. Let H be the operator from $\mathcal{H}$ to $\mathcal{E}$ defined by $Hu = Xu$ for all $u \in \mathbb{R}^N$. Let H^* be the operator from $\mathcal{E}$ to $\mathcal{H}$ which is the adjoint of H. H^* is such that $H^*v = X'v/T$ for all $v \in \mathbb{R}^T$. Observe that the operator $H^*H = X'X/T = S_{xx}$. $\hat{\psi}_j$ are the $(T \times 1)$ orthonormalized eigenvectors of the $(T \times T)$ matrix XX'/T such that[6]

$$\frac{XX'}{T}\hat{\psi}_j = \hat{\lambda}_j^2 \hat{\psi}_j.$$

By normalization, $\hat{\psi}_j'\hat{\psi}_j/T = 1$ The system $\left\{\hat{\lambda}_j, \widehat{\phi}_j, \hat{\psi}_j\right\}$ $j = 1, 2, ...$ is the singular value decomposition of the operator H. We have $H\widehat{\phi}_j = X\widehat{\phi}_j = \hat{\lambda}_j\hat{\psi}_j$ and $H^*\hat{\psi}_j = X'\hat{\psi}_j/T = \hat{\lambda}_j\widehat{\phi}_j$. The eigenfunctions $\hat{\psi}_j$, $j = 1, 2, ...T$ can be computed from $\hat{\psi}_j = X\widehat{\phi}_j/\hat{\lambda}_j$. Then $\hat{\lambda}_1^2 \geq \hat{\lambda}_2^2 \geq ...\hat{\lambda}_N^2$ and $\hat{\phi}_j$, $j = 1, 2, ..., N$, are the eigenvalues and the corresponding orthonormalized eigenvectors of dimension $(N \times 1)$ of S_{xx}. Note that $\hat{\phi}_j'\hat{\phi}_j = 1$. See for instance Carrasco, Florens, and Renault (2007). Note also that $\hat{\lambda}_j^2$ are consistent estimators of the corresponding eigenvalues of Σ_{xx}, λ_j^2 as T goes to infinity. We allow for multiple eigenvalues. In the sequence $\lambda_1^2 \geq \lambda_2^2 \geq ...\lambda_N^2$, each eigenvalue is repeated in regard to its multiplicity order.

[6] Note that the eigenvalues are denoted in this paper by $\widehat{\lambda}_j^2$, so that $\widehat{\lambda}_j^2$ are the eigenvalues themselves and not the eigenvalues squared.

In this paper, we consider four regularizations which take the form:

$$(S_{xx}^{\alpha})^{-1} v = \sum_{j=1}^{\min(N,T)} \frac{\hat{q}_j}{\hat{\lambda}_j^2} \left\langle v, \widehat{\phi}_j \right\rangle \widehat{\phi}_j \tag{2}$$

and the corresponding estimators:

$$\hat{\delta}^{\alpha} = (S_{xx}^{\alpha})^{-1} S_{xy} \tag{3}$$

$$= \sum_{j=1}^{\min(N,T)} \frac{\hat{q}_j}{\hat{\lambda}_j} \left\langle y, \hat{\psi}_j \right\rangle_T \widehat{\phi}_j \tag{4}$$

where $\left\langle y, \hat{\psi}_j \right\rangle_T = y'\hat{\psi}_j/T$ and $\hat{q}_j$ controls the amount of shrinkage, which differs across the various methods, and depends on a tuning parameter (α for all the methods except PLS, and k for PLS) as well as $\hat{\lambda}_j$**.** The predictor for y takes the form

$$\hat{y} = X\hat{\delta}^{\alpha} = \sum_{j=1}^{\min(N,T)} \hat{q}_j \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\psi}_j \equiv M_T^{\alpha} y,$$

where $M_T^{\alpha} = \sum_{j=1}^{\min(N,T)} \hat{q}_j \hat{\psi}_j \hat{\psi}_j'/T$ and the division by T comes from the fact that $\hat{\psi}_j'\hat{\psi}_j/T = 1$.

We consider four regularization schemes. The first three are traditionally applied to invert integral equations (Kress, 1999) while the fourth was introduced by Wold in the seventies.

1) **Ridge (R)**

A first regularization scheme is closely related to the Ridge regression where

$$(S_{xx}^{\alpha})^{-1} = (S_{xx} + \alpha I)^{-1}, \tag{5}$$

$$M_T^{\alpha} y = X (S_{xx} + \alpha I)^{-1} S_{xy} = \sum_{j=1}^{\min(N,T)} \frac{\hat{\lambda}_j^2}{\hat{\lambda}_j^2 + \alpha} \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\psi}_j \tag{6}$$

where $\alpha > 0$ is the tuning parameter, I is the $(N \times N)$ identity operator and $\widehat{q}_j = \hat{\lambda}_j^2 / \left(\hat{\lambda}_j^2 + \alpha\right)$.

When implementing Ridge in practice, it is convenient to express it as follows:

$$M_T^{\alpha} y = X\hat{\delta}_R^{\alpha} \tag{7}$$

$$\hat{\delta}_R^{\alpha} = (S_{xx} + \alpha I)^{-1} S_{xy} \tag{8}$$

2) **Landweber Fridman (LF)**

LF is an iterative method. Let d be such that $0 < d < 1/\hat{\lambda}_1^2$ where $\hat{\lambda}_1^2$ is the largest eigenvalue of S_{xx}. $\hat{\delta} = (S_{xx}^{\alpha})^{-1} S_{xy}$ can be computed iteratively from

$$\begin{cases} \hat{\delta}_l = (1 - dS_{xx})\, \hat{\delta}_{l-1} + dS_{xy},\ l = 1, 2, ..., 1/\alpha - 1 \\ \hat{\delta}_0 = dS_{xy} \end{cases}$$

where $1/\alpha - 1$ is some positive integer corresponding to the number of iterations. We see that $\hat{\delta}_l$ is a polynomial in S_{xx}. This estimator can be alternatively written in terms of the singular value decomposition of X, $\hat{\delta}_{LF}^{\alpha} = \sum_{j=1}^{\min(N,T)} \frac{\widehat{q}_j}{\widehat{\lambda}_j} \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\phi}_j$, with $\widehat{q}_j = 1 - \left(1 - d\widehat{\lambda}_j^2\right)^{1/\alpha}$. Using $\hat{\phi}_j = \frac{X'\hat{\psi}_j}{T\widehat{\lambda}_j}$, we obtain

$$\hat{\delta}_{LF}^{\alpha} = \sum_{j=1}^{\min(N,T)} \frac{\left(1 - \left(1 - d\widehat{\lambda}_j^2\right)^{1/\alpha}\right)}{\widehat{\lambda}_j^2} \left\langle y, \hat{\psi}_j \right\rangle_T \frac{X'\hat{\psi}_j}{T} \tag{9}$$

and one has $M_T^{\alpha} y = X\hat{\delta}_{LF}^{\alpha}$.

3) **Spectral Cut-off (SC) and Principal Components (PC)**

The SC procedure selects the eigenvectors associated with the eigenvalues greater than some threshold $\alpha > 0$:

$$M_T^{\alpha} y = \sum_{\hat{\lambda}_j^2 \geq \alpha} \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\psi}_j. \tag{10}$$

In fact, note that $M_T^{\alpha} = \sum_{\hat{\lambda}_j^2 \geq \alpha} \hat{\psi}_j \hat{\psi}_j' / T$, so that $\widehat{q}_j = I\left(\widehat{\lambda}_j^2 \geq \alpha\right)$, where $I(.)$ is the indicator function. Alternatively, one can select the eigenvectors associated with the largest eigenvalues. If the regressors are centered, ϕ_j correspond to the principal components and hence are consistent estimators of the factors of X (up to a rotation, under the conditions of Bai and Ng (2006)). Using the results in Carrasco, Florens and Renault (2007, p. 5694), and letting $\widehat{\Psi} = \left[\widehat{\psi}_1 | \widehat{\psi}_2 | ... | \widehat{\psi}_{k(\alpha)}\right]$ denote the matrix of the $k(\alpha)$ eigenvectors associated with the largest eigenvalues ($\left\{\lambda_j^2\right\}_{j=1}^{k(\alpha)}$ s.t. $\lambda_j^2 \geq \alpha$), then eq. (10) can be rewritten as: $M_T^{\alpha} y = \widehat{\Psi}\left(\widehat{\Psi}'\widehat{\Psi}\right)^{-1} \widehat{\Psi}' y$. Here, M_T^{α} is the projection matrix on the space spanned by $\hat{\psi}_j$, $j = 1, ..., k(\alpha)$ associated with the $k(\alpha)$ largest (positive) eigenvalues.

Again, for practical purposes, one can define:

$$\hat{\delta}_{PC}^{\alpha} = \left(\widehat{\Psi}'\widehat{\Psi}\right)^{-1} \widehat{\Psi}' y$$
$$M_T^{\alpha} y = \widehat{\Psi}\hat{\delta}_{PC}^{\alpha}.$$

So we call "principal component" the method which consists in projecting on the first k principal components and we call "spectral cut-off" the method consisting in projecting on the eigenvectors corresponding to the eigenvalues greater than some threshold α. Both methods are equivalent, only the regularization parameters differ.

4) **Partial Least Squares (PLS)**

Let k be the number of steps of PLS, which is the tuning parameter in PLS. Assume X and y are centered. PLS is an iterative method which looks for components which simultaneously

explain X and y well. As it takes the prediction of y into account to select the components, we say that it is a supervised method. The previous methods (R, LF, SC) are not supervised. According to Helland (1988, p.596), see also Groen and Kapetanios (2008), the PLS estimator can be written as:

$$\hat{\delta}_{PLS}^{k} = V_k \left(V_k' X' X V_k\right)^{-1} V_k' y$$

where $V_k = \left(X'y,\ \ (X'X)\, X'y, ..., (X'X)^{k-1} X'y\right)$, and

$$M_T^{\alpha} y = X V_k \left(V_k' X' X V_k\right)^{-1} V_k' X' y. \tag{11}$$

Let us denote $P_l = X\,(X'X)^{l-1} X'y$ the $l-$th PLS factor.[7] Then M_T^{α} is the projection matrix on the first k PLS factors: $P_1, P_2, ..., P_k$. Note that $P_l = XX'P_{l-1},\ j = 2, 3, ...$This can be conveniently written as power functions of $S_{xx},$ see Delaigle and Hall (2012):

$$\hat{\delta}_{PLS}^{k} = \sum_{l=1}^{k} \hat{\gamma}_j S_{xx}^{l-1} S_{xy} \tag{12}$$

where $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_k)'$ satisfies $\hat{\gamma} = \hat{H}^{-1}\hat{\mu}$ where $\hat{H}$ is a $(k \times k)$ matrix with (i, l) element $\hat{H}_{i,l} = S_{xy}' S_{xx}^{i+l-1} S_{xy}$ and $\hat{\mu}_l = S_{xy}' S_{xx}^{l-1} S_{xy}$. As LF, PLS uses power functions of S_{xx} to approximate δ, however it is done in a supervised way.

Alternatively, PLS can be written in terms of the eigenvectors $\hat{\phi}_j$. By adapting the formulas in Blazère, Gamboa, and Loubes (2014a, b) to our notation, the PLS estimator is:

$$\hat{\delta}_{PLS}^{k} = \sum_{j=1}^{\min(N,T)} \frac{\hat{q}_{kj}}{\hat{\lambda}_j} \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\phi}_j$$

where

$$\hat{q}_{kj} = \sum_{(j_1,...,j_k) \in I_k^+} \hat{w}_{j_1 ... j_k} \left[1 - \prod_{l=1}^{k} \left(1 - \frac{\hat{\lambda}_j^2}{\hat{\lambda}_{j_l}^2}\right)\right], \tag{13}$$

$$\hat{w}_{j_1 ... j_k} = \frac{\hat{p}_{j_1}^2 ... \hat{p}_{j_k}^2 \hat{\lambda}_{j_1}^4 ... \hat{\lambda}_{j_k}^4 V\left(\hat{\lambda}_{j_1}^2 ... \hat{\lambda}_{j_k}^2\right)^2}{\sum_{(j_1,...,j_k) \in I_k^+} \hat{p}_{j_1}^2 ... \hat{p}_{j_k}^2 \hat{\lambda}_{j_1}^4 ... \hat{\lambda}_{j_k}^4 V\left(\hat{\lambda}_{j_1}^2 ... \hat{\lambda}_{j_k}^2\right)^2},$$

$$\hat{p}_{j_i} = \left\langle y, \hat{\psi}_{j_i} \right\rangle_T,$$

$$I_k^+ = \left\{(j_1, ..., j_k) : \min(N, T) \geq j_1 > ... > j_k \geq 1\right\},$$

[7] There are various algorithms for PLS. This one is the simplest to present.

and $V(x_1, ..., x_k)$ is the Vandermonde determinant defined as

$$V(x_1, ..., x_k) = \begin{vmatrix} 1 & x_1 & \cdots & x_1^{k-1} \\ 1 & x_2 & \cdots & x_2^{k-1} \\ \vdots & & & \vdots \\ 1 & x_k & \cdots & x_k^{k-1} \end{vmatrix} = \prod_{1 \leq i < j \leq k} (x_j - x_i).$$

Note that

$$0 \leq \hat{w}_{j_1 \ldots j_k} \leq 1$$

and

$$\sum_{(j_1,\ldots,j_k) \in I_k^+} \hat{w}_{j_1 \ldots j_k} = 1.$$

As a consequence,

$$1 - \hat{q}_{kj} = \sum_{(j_1,\ldots,j_k) \in I_k^+} \hat{w}_{j_1 \ldots j_k} \prod_{l=1}^{k} \left(1 - \frac{\hat{\lambda}_j^2}{\hat{\lambda}_{j_l}^2}\right).$$

The weight $\hat{w}_{j_1 \ldots j_k}$ can be seen as probabilities but it should be remembered that they are random, as they are functions of y.

2.1 Comparing the Degree of Regularization/Shrinkage Across Methods

To summarize, the four regularized inverses yield

$$\hat{y} = M_T^{\alpha} y = \sum_{j=1}^{\min(N,T)} \hat{q}_j \left\langle y, \hat{\psi}_j \right\rangle_T \hat{\psi}_j = \frac{1}{T} \sum_{j=1}^{\min(N,T)} \hat{q}_j \hat{\psi}_j \hat{\psi}_j' y \tag{14}$$

where $\hat{q}_j \equiv q\left(\alpha, \hat{\lambda}_j^2\right)$ with for R: $q\left(\alpha, \lambda_j^2\right) = \frac{\lambda_j^2}{\lambda_j^2 + \alpha}$, for SC: $q\left(\alpha, \lambda_j^2\right) = I\left(\lambda_j^2 \geq \alpha\right)$, and for LF: $q\left(\alpha, \lambda_j^2\right) = 1 - \left(1 - d\lambda_j^2\right)^{1/\alpha}$. For PC with k components: $q_j = \hat{q}_j = I(j \leq k)$. For PLS with k PLS factors, $\hat{q}_j = \hat{q}_{kj}$ given by (13). The matrix M_T^{α} is idempotent for PC and PLS but not for LF and R.

Note that, when using all regressors without regularization, the prediction is equivalent to that obtained by eq.(14) where $\hat{q}_j = 1$ for all j, as using all regressors is equivalent to using all the eigenvectors. For R, SC, and LF, clearly $0 \leq \hat{q}_j \leq 1$, so that the shrinking property of these regularization methods is obvious. For PLS, $\hat{q}_{kj}$ may be negative and even greater than one. A simple illustration of this property (see Blazère et al., 2014b) is given for $k < N$ and $j = 1$. In that

case, $1-\frac{\hat{\lambda}_1^2}{\hat{\lambda}_{j_l}^2}<0$ for all j_l and hence

$$\prod_{l=1}^{k}\left(1-\frac{\hat{\lambda}_1^2}{\hat{\lambda}_{jl}^2}\right)<0 \text{ if } k \text{ is odd},$$
$$\prod_{l=1}^{k}\left(1-\frac{\hat{\lambda}_1^2}{\hat{\lambda}_{jl}^2}\right)>0 \text{ if } k \text{ is even},$$

therefore $\hat{q}_{kj}>1$ for k odd and $\hat{q}_{kj}<1$ for k even. Moreover, $\hat{q}_{kj}$ is random which makes the analysis of the properties of PLS in the next section more difficult.

3 Rate of Convergence

A good estimator of δ has to rely on a good estimator of Σ_{xx}^{-1}. However, when N is large relative to T, the number of terms in Σ_{xx} is so large that there exists no consistent estimator unless one is willing to impose some structure. In this section, we study the rate of convergence of $x_t'\hat{\delta}^{\alpha}-x_t'\delta^2$ in two interesting cases. In the first case (called "ill-posed model"), the eigenvalues of Σ_{xx} are bounded and decline to zero gradually. In the second case (the factor model), a few eigenvalues dominate the others. We believe that these two cases can cover many relevant applications in economics.

3.1 Ill-posed Model

First, we define some notation and recall some results on norms, which will be very useful in the sequel. For an arbitrary vector $v=(v_1,v_2,...,v_L)'$, $\|v\|$ denotes the Euclidean norm, $\|v\|=\sqrt{\sum_{i=1}^{L}v_i^2}$. For an arbitrary $(N\times T)$ matrix A with elements a_{it}, $\|A\|$ denotes the matrix norm defined as $\max_x\|Ax\|/\|x\|=\sqrt{\lambda_{\max}(A'A)}$ where $\lambda_{\max}(M)$ denotes the largest eigenvalue of M. The Frobenius norm is defined as $\|A\|_F=\sqrt{\sum_{i=1}^{n}\sum_{t=1}^{T}a_{it}^2}=\sqrt{trace(A'A)}$. For an arbitrary matrix A and a conformable vector v, it holds that $\|Av\|\leq\|A\|\|v\|$. Moreover, as $\|A\|\leq\|A\|_F$, it also holds that $\|Av\|\leq\|A\|_F\|v\|$

Assumption 1. Assume that $y_t=x_t'\delta+\varepsilon_t$, $t=1,...,T$, where the regressor x_t is a $N-$dimensional vector, $\|\delta\|<D$ for some constant D, $E(y_t)=E(x_t)=0$. ε_t is a stationary martingale difference sequence with respect to $\{\varepsilon_{t-1},\varepsilon_{t-2},...,x_{it},x_{it-1},...,\ i=1,2,...,N\}$, such that $E(\varepsilon\varepsilon'|X)=\sigma_\varepsilon^2 I$ and either (a) the $N-$vectors (x_t), $t=1,...,T$ are i.i.d. with mean 0 and $E\left[\|x_t\|^4\right]<C$ for some constant C, or (b) $(x_{it}x_{jt})$, $t=1,...,T$, $i,j=1,...,N$ is a strong mixing stationary process for all N with α coefficient satisfying

$$\exists\varepsilon>0,\ \exists d\in 2\mathbb{N},\ d\geq 2:\sum_{l=1}^{\infty}(l+1)^{d-2}(\alpha_l)^{\varepsilon/(d+\varepsilon)}<\infty$$

and

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left[E\left(\left|x_{it}x_{jt}\right|^{2+\varepsilon}\right)\right]^{2/(2+\varepsilon)}<C$$

for some constant C.

Remarks. The assumption that the observations are centered is a simplifying assumption which could be relaxed at the cost of more complicated formulas. The homoskedasticity assumption is not crucial and could be replaced by some high-level assumption. Assumption 1 imposes that x_t is either independent or weakly dependent. Many common processes are mixing (see Doukhan, 1994). The conditions in Assumption 1(b) permits to guarantee that S_{xx} is a consistent estimator of Σ_{xx} (see Lemma 1 below).

The condition $E\left[\|x_t\|^4\right]<C$ implies that $E\|x_t\|^2=\sum_{j=1}^{N}E\left(x_{jt}^2\right)=trace\left(\Sigma_{xx}\right)=\sum_j\lambda_j^2<C$ where λ_j^2 are the eigenvalues of Σ_{xx}. Therefore, Σ_{xx} has a finite trace. This implies two properties. First the largest eigenvalue of Σ_{xx} is bounded. Second, the smallest eigenvalues of Σ_{xx} decline to zero fast as $N\to\infty$, which means that Σ_{xx} becomes ill-posed.

By Cauchy-Schwarz, the condition $E\|x_t\|^2<C$ implies that $\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\Sigma_{xx}\left(i,j\right)\right)^2<C$, where $\Sigma_{xx}\left(i,j\right)$ is the (i,j) element of the matrix Σ_{xx}. Consequently when N grows, the extra covariances become negligible. This assumption is a kind of sparsity assumption on the matrix Σ_{xx} and may seem very strong but it is actually a common assumption in papers dealing with functional regressions, namely regressions where the regressor is a curve in a Hilbert space. The covariance matrix is then a covariance operator which is usually assumed to have a finite trace, see for instance Hall and Horowitz (2007) and Delaigle and Hall (2012). Under this assumption, the covariance operator can be approximated by a finite dimensional operator. The assumption $\|\delta\|<D$ implies that the importance of additional regressors is limited. It is again similar to conditions found in nonparametric regressions.

Note that Assumption 1 does not impose any restriction on the growth rate of N. In particular, N can be much larger than T.

Assumption 1 is not satisfied in the case where the regressors X_1, X_2,..., X_N are orthonormal. In that case, the matrix $X'X/T$ is the identity matrix and its trace equals N. But we believe that in economics, extra regressors tend to be very correlated to included regressors so that the extra variance contributed by these regressors becomes negligible.

Our assumptions are such that $\|x_t\|=O_p\left(1\right)$ and $\|\delta\|=O\left(1\right)$. This may seem restrictive as in many applications, $\|x_t\|=O_p\left(N\right)$. This case can be handled by rescaling the variables. Let $\tilde{x}_t=x_t/\sqrt{N}$ and $\tilde{\delta}=\sqrt{N}\delta$, then our model can be rewritten as

$$y_t=\tilde{x}_t'\tilde{\delta}+\varepsilon_t.$$

Now $\|\tilde{x}_t\|=O_p\left(1\right)$ and provided $\left\|\tilde{\delta}\right\|=O_p\left(1\right)$, our theory goes through and the rates of convergence

given in this section are valid. The condition $\left\|\tilde{\delta}\right\| = O_p(1)$ is equivalent to $\|\delta\| = O_p\left(\frac{1}{\sqrt{N}}\right)$. This is satisfied for the factor model we will analyze in Section 3.2. and for the following model. Consider a functional linear regression (see for instance Hall and Horowitz, 2007):

$$y_t = \int_0^1 b(s)\, x_i(s)\, ds + \varepsilon_i$$

where $\int_0^1 b^2(s)\, ds < \infty$ and $\int_0^1 x_i^2(s)\, ds < \infty$ a.s. An example is given by a model with data sampled at different frequencies. Let y_t be the quarterly US output growth for a quarter t and $x_t(s)$ be the interest rate at time s during that quarter. The interest rate is observed at a higher frequency than the output growth. For instance, assume that it is observed at a daily frequency with dates denoted s_1, s_2, ..., s_N, so that the integral is approximated by an average and the estimated model becomes:

$$y_t = \frac{1}{N}\sum_{j=1}^{N} b(s_j)\, x_t(s_j) + \varepsilon_i.$$

Now denote $x_t = (x_t(s_1), x_t(s_2), ..., x_t(s_N))'$ and $\delta = (b(s_1), b(s_2), ..., b(s_N))'/N$, then we have $\|x_t\| = O_p\left(\sqrt{N}\right)$ and $\|\delta\| = O\left(1/\sqrt{N}\right)$. So, regularization provides an alternative to the MIDAS regression proposed by Ghysels, Santa-Clara, and Valkanov (2005).

Note that our model differs from the usual sparse models (see the survey by Fan, Lv and Qi, 2011), in particular we do not assume that the coefficients δ_i are zero for all i except in a small subset. Indeed, the elements of δ may be all different from zero in our model. Our theory does not cover the typical sparse model where $\|x_t\| = O_p\left(\sqrt{N}\right)$ and $\|\delta\| = O(1)$.

When no structure is imposed to Σ_{xx}, it is well-known that the sample covariance S_{xx} is not a consistent estimator of Σ_{xx} as N grows faster than T. More precisely, Ledoit and Wolf (2004, Lemma 3.1) show that $\|S_{xx} - \Sigma_{xx}\|^2 = O_p\left(N^2/T\right)$. Under Assumption 1, the rate of convergence of S_{xx} to Σ_{xx} does not depend on N:

Lemma 1 *Under Assumption 1,* $\|S_{xx} - \Sigma_{xx}\|_F^2 = O_p(1/T)$.

To study the rate of the MSE, we need to impose an additional assumption which will be useful to characterize the regularization bias.

Assumption 2. Σ_{xx} is non singular and there exists $\beta > 1$ such that

$$\sum_j \frac{\langle \delta, \phi_j \rangle^2}{\lambda_j^{2\beta-2}} < \infty. \tag{15}$$

where $\left\{\phi_j, \lambda_j^2\right\}$, $j = 1, 2, ..., N$ are the eigenvectors and eigenvalues of Σ_{xx}.

The fact that Σ_{xx} is non singular means that there are no redundant regressors. As a result, the population OLS parameter, $\delta = \Sigma_{xx}^{-1}\Sigma_{xy}$, is well defined. We could relax this condition and define δ as the population OLS with minimum norm, namely $\delta = \Sigma_{xx}^{-}\Sigma_{xy}$ where Σ_{xx}^{-} is the Moore Penrose generalized inverse of Σ_{xx}.

Condition (15) is a source condition and can be found in the papers on functional estimation using inverse problem techniques (see Carrasco, Florens, and Renault, 2007). The elements $\langle \delta, \phi_j \rangle$ can be thought of as Fourier coefficients in the decomposition of δ on the basis $\{\phi_j, j = 1, 2, ..., N\}$. (15) relates the rate at which $\langle \delta, \phi_j \rangle^2$ declines to zero relatively to that $\lambda_j^{2\beta-2}$. The larger β is, the easier it is to recover the signal δ from the basis $\{\phi_j, j = 1, 2, ..., N\}$. Importantly, we will show that the regularization bias is a function of β. The factor model is an example where the condition (15) is satisfied for β arbitrarily large (see Section 3.2).

Let δ^α be the regularized version of δ defined as $\delta^\alpha = \left(\Sigma_{xx}^\alpha\right)^{-1}\Sigma_{xy} = \left(\Sigma_{xx}^\alpha\right)^{-1}\Sigma_{xx}\delta$ using Assumption 1. Then, we have

$$\begin{aligned}\delta^\alpha &= \sum_{j=1}^{N} \frac{q_j}{\lambda_j^2} \langle \Sigma_{xx}\delta, \phi_j \rangle \phi_j \\ &= \sum_{j=1}^{N} q_j \langle \delta, \phi_j \rangle \phi_j\end{aligned}$$

where $q_j = q\left(\alpha, \lambda_j^2\right)$. The prediction error can be decomposed as the sum of two terms:

$$x_t'\hat{\delta}^\alpha - x_t'\delta = x_t'\hat{\delta}^\alpha - x_t'\delta^\alpha + x_t'\delta^\alpha - x_t'\delta.$$

The first term, $x_t'\hat{\delta}^\alpha - x_t'\delta^\alpha$, captures the estimation error, whereas the second term, $x_t'\delta^\alpha - x_t'\delta$, captures the regularization bias. To derive the rate of convergence, we analyze these two terms separately. The results are summarized in the following proposition.

Proposition 2 *Under Assumptions 1 and 2, we have as N and T go to infinity*

$$x_t'\hat{\delta}^\alpha - x_t'\delta = \begin{cases} O_p\left(\alpha^{\min\left(\frac{\beta}{2},1\right)}\right) + O_p\left(\frac{1}{\alpha\sqrt{T}}\right) & \textit{for R,} \\ O_p\left(\alpha^{\beta/2}\right) + O_p\left(\frac{1}{\alpha\sqrt{T}}\right) & \textit{for SC and LF,} \end{cases}.$$

where the first O_p term corresponds to the regularization bias, $x_t'\delta^\alpha - x_t'\delta$, and the second O_p term corresponds to the estimation error, $x_t'\hat{\delta}^\alpha - x_t'\delta^\alpha$.

Remarks:

1. The strict exogeneity is not required to establish Proposition 2.

2. The rate of convergence of Ridge is not as fast as that of SC and LF when $\beta > 2$.[8] It means that if the model is easy to estimate (if, for instance, $X\delta$ is explained by a few $\hat{\psi}_j$) then SC and

[8] The rate could be improved by using iterated Tikhonov (see Engl, Hanke, and Neubauer, 2000).

LF should be preferred to Ridge. This is a well-known property of Ridge called saturation. An illustration of this phenomenon is given by the factor model in the next section.

3. To derive the α which minimizes the sum of the two rates, we look for the value of α which equates the rate for the regularization bias and the rate for the estimation error. For SC and LF, we obtain $\alpha = T^{-1/(\beta+2)}$ and $x_t'\hat{\delta}^\alpha - x_t'\delta = O_p\left(T^{-\beta/(2(\beta+2))}\right)$. It follows that when β becomes large, the rate approaches the parametric rate, $T^{-1/2}$. This is not true for Ridge because of the saturation property.

4. PC and SC are the same estimation method. They differ only in the way the regularization parameter is defined. However, the rate of convergence of PC depends on the decay rate of the eigenvalues as we will see in the next proposition where we analyze the conditional mean-square error. The rate of convergence of PC (without conditioning) requires stronger assumptions on the eigenvalues, in particular the difference between subsequent eigenvalues $\lambda_j^2 - \lambda_{j+1}^2$ should not decrease to zero too fast when j goes to zero (see Cai and Hall (2006) and Hall and Horowitz (2007) among others). As pointed out by Hall and Horowitz (2007), one advantage of Ridge is that it does not require any restrictions on the spacing of eigenvalues. We see that, in our context, LF and SC do not require extra restrictions either and are consistent even when multiple eigenvalues are present. It should be noted that Hall and Horowitz (2007) derive an optimal rate for the estimation of δ, while here we focus on the estimation of the prediction $X\delta$. The presence of X induces a smoothing which improves the rate of convergence.

To establish the optimality of the selection criteria used to determine α, we need results on the conditional mean squared prediction error.

Proposition 3 *Under Assumptions 1-2, $E(\varepsilon|X) = 0$, we have as N and T go to infinity*

$$\left\{\frac{1}{T}E\left[\left\|X\hat{\delta}^\alpha - X\delta\right\|^2 |X\right]\right\}^{1/2} = \begin{cases} O_p\left(\alpha^{\min\left(\frac{\beta}{2},1\right)}\right) + O_p\left(\frac{1}{\sqrt{\alpha T}}\right) \text{ for R,} \\ O_p\left(\alpha^{\frac{\beta}{2}}\right) + O_p\left(\frac{1}{\sqrt{\alpha T}}\right) \text{ for SC and LF,} \\ O_p\left(\lambda_{k+1}^{\beta}\right) + \sqrt{\frac{k\sigma_\varepsilon^2}{T}} \text{ for PC with } k \text{ principal components} \end{cases}.$$

It is interesting to compare the results of Propositions 2 and 3. In Proposition 3, we assume strict exogeneity. The proof of Proposition 2 accounts for the estimation error $\Sigma_{xx} - S_{xx}$, while in Proposition 3, the conditioning on X makes this unnecessary. This explains why the estimation error is larger in Proposition 2 than in Proposition 3. The rate for PC will be used for comparison with PLS later on.

For PLS, the population regularized estimator (denoted by δ_{PLS}^k) has two expressions which are equal to each other. One expression is based on the spectral decomposition of Σ_{xx} :

$$\delta_{PLS}^k = \sum_{j=1}^{\min(N,T)} q_{kj} \langle \delta, \phi_j \rangle \phi_j$$

where q_{kj} is as in (13) with $\hat{\lambda}_{j_l}$ replaced by λ_{j_l} and $\hat{p}_{j_l}$ replaced by $p_{j_l} = \lambda_{j_l} \langle \delta, \phi_{j_l} \rangle$. The other expression is based on a non-orthogonal basis formed of Σ_{xx}^l

$$\delta_{PLS}^k = \sum_{l=1}^{k} \gamma_l \Sigma_{xx}^l \delta$$

where $\gamma = (\gamma_1, ..., \gamma_k)' = H^{-1} (\mu_1, ..., \mu_k)'$ and H is the $(k \times k)$ matrix with (i, l) element

$$H_{il} = \Sigma_{xy}' \Sigma_{xx}^{i+l-1} \Sigma_{xy} \tag{16}$$

and $\mu_l = \Sigma_{xy}' \Sigma_{xx}^{l-1} \Sigma_{xy}$. We will use the first expression to study the bias, whereas we will use the second expression to study the variance.

Proposition 4 *Under Assumptions 1 and 2, we have as N and T go to infinity*

$$x_t' \delta_{PLS}^k - x_t' \delta = O_p \left(\lambda_{k+1}^{\beta} \right).$$

Moreover,

$$E \left(\left(x_t' \delta_{PLS}^k - x_t' \delta \right)^2 \right) \leq E \left(\left(x_t' \delta_{PC}^k - x_t' \delta \right)^2 \right) \text{ for all } k.$$

In the special case where there is a constant $C > 0$ such that

$$\frac{1}{C} j^{-2} \leq \lambda_j^2 \leq C j^{-2} \tag{17}$$

then $x_t' \delta_{PLS}^k - x_t' \delta = O_p \left(k^{-3\beta} \right) = O_p \left(\lambda_k^{3\beta/2} \right)$ as k grows with $\min(T, N)$.

Assume moreover that $(x_t', y_t)'$, $t = 1, 2, ..., T$ is an iid sample and the largest eigenvalue of $\Sigma_{xx} < 1$. Let λ_H be the smallest eigenvalue of H defined in (16). Provided k is such that $1 \leq k \leq CT^{1/2}$ for an arbitrary constant C and diverges sufficiently slowly to ensure that $T^{-1/2} \lambda_H^{-1} \|\gamma\| + T^{-1} \lambda_H^{-3} \to 0$, then as N and T go to infinity,

$$x_t' \widehat{\delta}_{PLS}^k - x_t' \delta_{PLS}^k = O_p \left(\frac{1 + \|\gamma\|}{T^{1/2} \lambda_H} + \frac{1}{T \lambda_H^3} \right).$$

Remark 1. The rate we derived for the bias of PLS coincides with that of PC. It is only an upper bound which can be improved in the special case where $\frac{1}{C} j^{-2} \leq \lambda_j^2 \leq C j^{-2}$. We also show that, for the same number of components k, the squared bias of PLS is smaller than that of PC. It has been established earlier that PLS fits better than PC, in the sense that $\left\| y - X \delta_{PLS}^k \right\|^2 < \left\| y - X \delta_{PC}^k \right\|^2$, see De Jong (1993), Phatak and de Hoog (2002), and more recently Blazère et al. (2014a).

Remark 2. The assumption $\frac{1}{C} j^{-2} \leq \lambda_j^2 \leq C j^{-2}$ is often used in nonparametric estimation. For instance, it was considered by Hall and Horowitz (2007) in the context of a functional regression.

Remark 3. Blazère et al. (2014a) study the empirical risk of PLS, namely $E \left(\frac{1}{T} \left\| y - X \delta_{PLS}^k \right\|^2 \right)$ and show that it decreases at an exponential rate as k goes to infinity. Their results are not

applicable here because they rely on the assumption that N is fixed and that the condition number $(\lambda_1^2/\lambda_N^2)$ of the matrix $X'X$ is bounded. Here, on the contrary, the number of regressors, N, is growing with the number of observations, T, and the matrix $X'X$ is ill-posed so that the condition number goes to infinity. Note that the condition number is independent of the re-scaling of the matrix $X'X$.

Remark 4. The squared bias of PLS, $\left\|\delta_{PLS}^k - \delta\right\|^2 = \sum_j (1 - q_{kj})^2 \langle \delta, \phi_j \rangle^2$, does not decrease monotonically as it is the case for PC and other regularization methods. Indeed, q_{kj} tend to alternate between values greater and smaller than one, so that the bias of PLS may increase as k increases.

Remark 5. The results on the estimation error follow from Delaigle and Hall (2012, (5.11)). We imposed the same assumptions as they did. The rate of this term depends on the smallest eigenvalue λ_H of H. H being a determinate Hankel matrix, λ_H will converge to zero exponentially fast as k goes to infinity, see Delaigle and Hall (2012, Section 5.3.4.) and Berg and Szwarc (2011, Theorem 2.1). As a result, k has to grow sufficiently slowly for the estimation error to go to zero. It is interesting to compare it with the estimation error of PC. Let $\widehat{\delta}_{PC}^k$ denote the PC estimator with k components, and δ_{PC}^k denote the population estimator regularized by PC. The estimation error of PC increases linearly in k. So we see that, on the one hand, the bias of PLS is smaller than that of PC; on the other hand, the estimation error of PLS maybe larger than that of PC. It is not clear which one will result in the smallest MSE.

Summary of theoretical results for the ill-posed model. In summary, we found that in cases where the signal is difficult to recover (i.e. cases where $\beta < 2$), the convergence rates of the prediction errors of Ridge, SC, and LF are the same whereas, for easy to recover signals (i.e. cases with $\beta > 2$), Ridge is slower than the other two methods. When comparing PC with PLS, we found that the bias of PLS is smaller than the bias of PC for the same number of components, k, while its estimation error maybe larger for large k. As the MSE consists of both, it is not clear which one will have the smaller MSE.

3.2 Approximate Factor Model

We consider a factor model as in De Mol, Giannone and Reichlin (2008). Here, we do not postulate a linear relationship between y_t and x_t but assume that y_t and x_t depend on a small number of common factors F_t. The aim is to estimate the linear projection of y_t on x_t using a large N, large T asymptotics. The number of factors r is assumed to be fixed.

Assumption A. $y_t = \theta' F_t + v_t$, where v_t is orthogonal to x_t for all N and where the factors $F_t = (f_{1t}, ..., f_{rt})'$ are a r-dimensional stationary process with covariance $E(F_t F_t') = I_r$.

Assumption B. x_t is such that $x_t = \Lambda F_t + \xi_t$ where:

(i) the residuals ξ_t are a N-dimensional stationary process with covariance $E(\xi_t\xi_t') = \Psi$ of full rank for all N;

(ii) the $(N \times r)$ matrix Λ is a non-random matrix and full rank r for each N;

(iii) the residuals ξ_t are orthogonal to the factors F_t.

Assumption B allows for the idiosyncratic errors ξ_{it} to be cross-sectionally correlated as in the "approximate factor model" of Chamberlain and Rothschild (1983) and serially correlated. It allows also for the largest eigenvalue of Ψ to grow with N. By Assumptions A and B, $\Sigma_{xx} = E(x_t x_t') = \Lambda\Lambda' + \Psi$ and $\Sigma_{xy} = E(x_t y_t) = \Lambda\theta$. The population OLS regression coefficient is

$$\delta = \Sigma_{xx}^{-1}\Sigma_{xy} = \left(\Lambda\Lambda' + \Psi\right)^{-1}\Lambda\theta$$

and the linear projection of y_t on x_t is given by

$$x_t'\delta = x_t'\left(\Lambda\Lambda' + \Psi\right)^{-1}\Lambda\theta.$$

We need an extra condition to obtain the rates for the sample variance and covariance.

Assumption C. There exists a finite constant K such that for all $T \in \mathbb{N}$ and $i, j \in \mathbb{N}$

$$TE\left(e_{xx,ij}^2\right) < K \text{ and } TE\left(e_{xy,ij}^2\right) < K$$

where $e_{xx,ij}$ is the (i,j) element of the matrix $\Sigma_{xx} - S_{xx}$ and $e_{xy,ij}$ is the (i,j) element of the matrix $\Sigma_{xy} - S_{xy}$.

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be the largest and smallest eigenvalues of A.

Proposition 5 *Under Assumptions A, B, and C, we have as N and T go to infinity,*

$$x_t'\delta^\alpha - x_t'\delta = O_p\left(\alpha\sqrt{N}\frac{\|\delta\|}{\lambda_{\min}(\Lambda'\Lambda)}\right) \text{ for } R,$$

$$x_t'\delta^\alpha - x_t'\delta = O_p\left(\alpha^{\beta/2}\sqrt{N}\frac{\|\delta\| I(\alpha > \lambda_{\min}(\Lambda'\Lambda))}{\lambda_{\min}(\Lambda'\Lambda)^{\beta/2}}\right) \text{ for } SC,$$

$$x_t'\delta^\alpha - x_t'\delta = O_p\left(\left(\frac{\alpha}{d}\right)^{\beta/2}\sqrt{N}\frac{\|\delta\|}{\lambda_{\min}(\Lambda'\Lambda)^{\beta/2}}\right) \text{ for } LF,$$

for an arbitrary large $\beta \geq 2$ (independent of N and T). Moreover,

$$x_t'\hat{\delta}^\alpha - x_t'\delta^\alpha = O_p\left(\frac{N}{\alpha\sqrt{T}}\left[1 + \sqrt{N}\|\delta\|\right]\right) \text{ for } R \text{ and } SC,$$

and

$$x_t'\hat{\delta}^\alpha - x_t'\delta^\alpha = O_p\left(\frac{dN}{\alpha\sqrt{T}}\left[1 + \sqrt{N}\|\delta\|\right]\right) \text{ for } LF$$

where d refers to the term appearing in the Landweber-Fridman regularization (see Section 2). Moreover, $\|\delta\| = O\left(\frac{\sqrt{\lambda_{\max}(\Lambda'\Lambda)}}{\lambda_{\min}(\Lambda'\Lambda)}\right)$.

Remarks.

1. This result has been derived for Ridge by De Mol et al. (2008, Proposition 1) and is extended here to LF and SC.

2. Compared to Proposition 2, we see that the rate depends of the number of regressors N. It is due to the fact that $\|x_t\| = O_p\left(\sqrt{N}\right)$ and $\|\delta\| = O\left(\frac{\sqrt{\lambda_{\max}(\Lambda'\Lambda)}}{\lambda_{\min}(\Lambda'\Lambda)}\right)$. In the ill-posed model, we had $\|x_t\| = O_p(1)$ and $\|\delta\| = O(1)$.

3. Our rates take into account the fact that the factors are not observed through the difference between Σ_{xx} and S_{xx}.

Consider the following assumption which stipulates that all the eigenvalues of $\Lambda'\Lambda$ grow linearly with N.

Assumption D.

$$0 < \lim \inf_{N\to\infty} \frac{\lambda_{\min}(\Lambda'\Lambda)}{N} \leq \lim \sup_{N\to\infty} \frac{\lambda_{\max}(\Lambda'\Lambda)}{N} < \infty.$$

Under Assumption D, we have

$$x_t'\hat{\delta}^\alpha - x_t'\delta = \begin{cases} O_p\left(\frac{\alpha}{N}\right) + O_p\left(\frac{N}{\alpha\sqrt{T}}\right) \text{ for R,} \\ O_p\left(\left(\frac{\alpha}{N}\right)^{\beta/2} I\left(\alpha > \lambda_{\min}(\Lambda'\Lambda)\right)\right) + O_p\left(\frac{N}{\alpha\sqrt{T}}\right) \text{ for SC,} \\ O_p\left(\alpha^{\beta/2}\right) + O_p\left(\frac{1}{\alpha\sqrt{T}}\right) \text{ for LF} \end{cases}$$

because $d = O(1/N)$. The optimal α which minimizes the rate of the right-hand side is the value of α which equates the rates for the regularization bias and the estimation error. Choosing this α, we obtain the following results.

$$x_t'\hat{\delta}^\alpha - x_t'\delta = \begin{cases} O_p\left(T^{-1/4}\right) \text{ with } \alpha = NT^{-1/4} \text{ for R,} \\ O_p\left(\sqrt{T}^{-\beta/(\beta+2)}\right) \text{ with } \alpha = NT^{-\frac{1}{\beta+2}} \text{ for SC,} \\ O_p\left(\sqrt{T}^{-\beta/(\beta+2)}\right) \text{ with } \alpha = T^{-\frac{1}{\beta+2}} \text{ for LF} \end{cases}$$

for arbitrary large $\beta > 2$. We see that the rate for SC and LF is arbitrary close to $T^{-1/2}$, and is much faster than that obtained for Ridge. The good performance of SC makes sense because SC exploits the factor structure of the model and predicts y_t by projecting onto the first principal components. On the other hand, Ridge and LF are omnibus regularization methods. So the performance of LF, which is as good as SC, may seem surprising. However, it is consistent with the findings of Section 3.1. The factor model is an example of model where the source condition (15) holds for an arbitrary large β because the signal belongs to the span of a finite number, r, of principal components. Hence, Ridge is clearly at a disadvantage.

To address the convergence to the optimal forecast, $\theta' F_t$, we add the following assumption.

Assumption E. There exists $0 < \nu \leq 1$ such that

$$\lim \sup_{n\to\infty} \frac{1}{N^{1-\nu}} \lambda_{\max}(\Psi) < \infty.$$

When the errors are idiosyncratic, Assumption E holds with $\nu = 1$. De Mol et al. (2008) shows that, under Assumptions D and E,

$$x_t'\delta - \theta' F_t = O_p\left(N^{-\nu/2}\right).$$

We can deduct the rate of convergence of $x_t'\hat{\delta}^\alpha - \theta' F_t$ by using the decomposition

$$x_t'\hat{\delta}^\alpha - \theta' F_t = x_t'\hat{\delta}^\alpha - x_t'\delta + x_t'\delta - \theta' F_t$$

and Proposition 5.

Corollary 6 *Under Assumptions A-E, as N and T go to infinity, we have*

$$\Delta_{nT}\left(x_t'\hat{\delta}^\alpha - \theta' F_t\right) = O_p(1)$$

with

$$\Delta_{nT} = \begin{cases} \min\left(N^{\frac{\nu}{2}}, T^{\frac{1}{4}}\right) & \textit{for } R, \\ \min\left(N^{\frac{\nu}{2}}, \sqrt{T}\right) & \textit{for SC and LF.} \end{cases}$$

So, all three regularization methods converge to the optimal forecast but at different rates. There is no restriction on the relative rate of N and T for the consistency. We see that LF and SC reach the fastest possible rate. This result was previously established for SC by Bai (2003, Theorem 3) in the case of idiosyncratic noise.

Now, we discuss properties of PC and PLS. We are not able to say anything about the variance of PLS, so our comparison will focus on the bias. The bias of PC is zero for $k \geq r$. The bias of PLS is also equal to zero for $k = r$. To see this, observe that $\langle \delta, \phi_j \rangle = 0$ for $j > r$ because δ belongs to the range of Λ. Hence, $\left\|\delta_{PLS}^k - \delta\right\|^2 = \sum_{j=1}^r \left(q_{kj} - 1\right)^2 \langle \delta, \phi_j \rangle^2$ where

$$1 - q_{kj} = \sum_{(j_1,\ldots,j_k)\in I_k^*} w_{j_1\ldots j_k} \prod_{l=1}^k \left(1 - \frac{\lambda_j^2}{\lambda_{j_l}^2}\right),$$

$$w_{j_1\ldots j_k} = \frac{p_{j_1}^2 \ldots p_{j_k}^2 \lambda_{j_1}^4 \ldots \lambda_{j_k}^4 V\left(\lambda_{j_1}^2 \ldots \lambda_{j_k}^2\right)^2}{\sum_{(j_1,\ldots,j_k)\in I_k^+} p_{j_1}^2 \ldots p_{j_k}^2 \lambda_{j_1}^4 \ldots \lambda_{j_k}^4 V\left(\lambda_{j_1}^2 \ldots \lambda_{j_k}^2\right)^2},$$

$$p_{j_i} = \lambda_{jl} \langle \delta, \phi_{j_l} \rangle,$$

and I_k^+ is replaced by a restricted version denoted I_k^* :

$$I_k^* = \{(j_1, ..., j_k) : r \geq j_1 > ... > j_k \geq 1\}.$$

For $k = r$, we have

$$\prod_{l=1}^r \left(1 - \frac{\lambda_j^2}{\lambda_{j_l}^2}\right) = 0,$$

hence the bias is null.

Consider a special case where $\langle \delta, \phi_j \rangle = 0$ for all $j \neq r$. For $k = 1, 2.., r-1$, PC has a bias:

$$\delta_{PC}^k - \delta = -\sum_{j=k+1}^{r} \langle \delta, \phi_j \rangle \phi_j = -\langle \delta, \phi_r \rangle \phi_r = -\delta.$$

For PLS for $k = 1$, we have

$$q_{1j} = \sum_{l=1}^{r} w_l \left(1 - \prod_{l=1}^{r} \left(1 - \frac{\lambda_j^2}{\lambda_l^2}\right)\right)$$

with

$$\begin{aligned} w_l &= \frac{p_l^2 \lambda_l^2}{\sum_{l=1}^{r} p_l^2 \lambda_l^2} = 0 \text{ if } l \neq r, \\ &= 1 \text{ if } l = r. \end{aligned}$$

Hence, $q_{1j} = \frac{\lambda_j^2}{\lambda_r^2}$ and the bias of PLS is given by

$$\delta_{PLS}^k - \delta = (q_{kr} - 1) \langle \delta, \phi_r \rangle \phi_r = 0$$

for $k = 1$. Hence, the bias of PLS is zero for $k = 1$, while it is different from zero for PC. In this example, a single PLS factor permits to achieve a null bias while PC needs r factors.

4 Data-driven Selection of the Tuning Parameter via Cross-Validation

4.1 Criteria

The four regularization methods involve a regularization (or tuning) parameter, α. An important practical issue if how to choose α. Ideally, we would like to select α for which $\left\| X\hat{\delta}^{\alpha} - X\delta \right\|^2$ is as small as possible. However, this is not feasible because δ is unknown.

Following Li (1986, 1987), we investigate the following cross-validation techniques:[9]

(i) Generalized cross-validation (GCV):

$$\hat{\alpha} = \arg \min_{\alpha \in A_T} \frac{T^{-1} \left\| y - M_T^{\alpha} y \right\|^2}{\left(1 - T^{-1} tr\left(M_T^{\alpha}\right)\right)^2}, \tag{18}$$

(ii) Mallows' (1973) C_L:

$$\hat{\alpha} = \arg \min_{\alpha \in A_T} T^{-1} \left\| y - M_T^{\alpha} y \right\|^2 + 2\widehat{\sigma}_{\varepsilon}^2 T^{-1} tr\left(M_T^{\alpha}\right) \tag{19}$$

[9] Li (1987) also considers delete-one cross-validation, which is less suitable for time series analysis; therefore, we do not consider it.

where $\widehat{\sigma}_{\varepsilon}^2$ is a consistent estimator of the variance of ε_t, σ_{ε}^2, and A_T is the set within which α is selected.

Note that, for PC, $tr\left(M_T^{\alpha}\right) = k$, where k is the number of factors retained. The trace is also a measure of the effective degrees of freedom of the fit, which guides the penalty in the cross-validation procedure. In fact, usually, in a linear regression with p parameters, the penalty is a function the number of parameters (i.e. k for PC); in Ridge estimation, even though all the coefficients will be non-zero, their fit is restricted and controlled by the parameter α; when there is no regularization, $\alpha = 0$ and the degree of freedom equals the total number of parameters in the regression, while when the amount of regularization is very large α increases and the degree of freedom decreases.

The criteria we consider are different from those suggested by Bai and Ng (2002): it is therefore important to compare them with theirs to point out their differences and similarities. Bai and Ng (2002) and Stock and Watson (2002a) assume that x_t has r common latent factors:

$$x_t = \Lambda F_t + \xi_t.$$

For selecting the number of factors, Bai and Ng (2002) suggest the following criterion:

$$\min_k \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{it} - \lambda_i^{k\prime} \hat{F}_t^k \right)^2 + kg\left(N, T\right) \tag{20}$$

where $\hat{F}_t^k$ is the vector of the first k principal components, $\left(\hat{\psi}_1, ..., \hat{\psi}_k\right)'$, and $g\left(N, T\right)$ is a penalty which is function of N and T.

The differences are as follows. First, the criterion described by eq. (20) is valid when there is a factor model that explains the data, i.e. r is finite. The criteria described in eqs. (18) and (19) instead, do not assume a factor model and are valid regardless of whether the data are well-described by a factor model or not.

In addition, the criterion described by eq. (20) aims at finding the factors that best explain the variability in the predictors, regardless of the way the predictors relate to the variable to be predicted.

Furthermore, the penalty terms are also different: if the predictors, X, have a factor structure with k factors then $\hat{k}$ obtained from (20) is a consistent estimator of k (see Theorem 2 in Bai and Ng, 2002); however, the number of factors relevant for y may be smaller than k, in which case eqs. (18) and (19) may deliver a more accurate estimate of the number of factors. On the other hand, it is important to note that criteria (18) and (19) are AIC-type criteria and do not deliver a consistent estimate of the number of factors in y. This number may be over-estimated. However, this is not a major issue, as in general the model is never correctly specified. In their paper, Stock and Watson (2002a) perform Monte Carlo simulations to compare the forecasting performance of factor models where the number of factors is selected via AIC and BIC with that of factor models where the

number of factors is selected via Bai and Ng's (2002) criterion. For comparison, we therefore also include AIC and BIC in our Monte Carlo simulation exercises.

Since GCV and Mallows C_L are not justified for PLS, we use leave-one-out cross-validation to select the number of PLS components in the simulations.

4.2 Optimality of the Selection

In this section, we wish to establish the optimality of Mallows' C_L and generalized cross-validation (GCV) in the following sense

$$\frac{L_T(\hat{\alpha})}{\inf_{\alpha \in A_T} L_T(\alpha)} \to 1 \text{ in probability} \tag{21}$$

as N and T go to infinity where $L_T(\alpha) = \frac{1}{T}\left\|X\hat{\delta}^{\alpha} - X\delta\right\|^2$ and $\hat{\alpha}$ is defined in eqs. (18) and (19). A selection procedure which satisfies this condition is said to be asymptotically loss efficient. This property is weaker than the consistency of α to some true value. Instead, it establishes that using $\hat{\alpha}$ in the criterion $L_T(\hat{\alpha})$ delivers the same rate of convergence as if minimizing $L_T(\alpha)$ directly.

For this analysis, we focus on the ill-posed model and on the following three regularizations: SC, Ridge, and LF. For these three regularizations, the prediction of y is a linear function of y denoted $M_T^{\alpha} y$, where M_T^{α} depends on X and the regularization parameter α. For these estimators, Li's (1986, 1987) optimality results apply. Our task will be to check Li's (1986, 1987) conditions. For PLS, the matrix M_T^{α} depends on y also, therefore the prediction is not linear in y and Li's (1986, 1987) results do not apply.

To establish optimality, we need to replace the condition that ε_t is a martingale difference sequence by a stronger condition of independence and strict exogeneity. However, x_t can be serially correlated.

Assumption 3. (i) $\varepsilon_t|X$ i.i.d. $\left(0, \sigma_\varepsilon^2\right)$ and $E\left(\varepsilon_t^8\right) < \infty$. (i') $\varepsilon_t|X$ i.i.d. $\mathcal{N}\left(0, \sigma_\varepsilon^2\right)$.
(ii) $\hat{\sigma}_\varepsilon^2 \xrightarrow{P} \sigma_\varepsilon^2$.
(iii) The eigenvalues $\hat{\lambda}_j^2$ of XX'/T are such that for any m such that $m/T \to 0$,

$$\frac{\left(\frac{1}{T}\sum_{j=m+1}^{T} \hat{\lambda}_j^2\right)^2}{\frac{1}{T}\sum_{j=m+1}^{T} \hat{\lambda}_j^4} \to 0.$$

Proposition 7 *Under Assumptions 1, 2, and 3(i)(ii), Mallows' C_L and the generalized cross-validation criteria are asymptotically optimal in the sense of (21) for SC and LF with $A_T = \{\alpha : \alpha > C/T\}$ for some $C > 1$.*

Under Assumptions 1, 2, and 3(i')(ii), the Mallows C_L is asymptotically optimal in the sense of (21) for Ridge. Under Assumptions 1,2,3(i')(ii)(iii), the generalized cross-validation is asymptotically optimal in the sense of (21) for Ridge.

Remarks. Many papers applying Li's (1987) optimality results impose a high-level assumption of the type:

$$\inf_{\alpha \in A_T} TR_T(\alpha) \to \infty$$

where $R_T(\alpha) = E[L_T(\alpha)|X]$. Given our Proposition 3, we know the rate of $TR_T(\alpha)$ and we are able to check this condition.

For the optimality of GCV for SC and LF, a restriction needs to be imposed on the set A_T, namely $A_T = \{\alpha : \alpha > C/T\}$ for some $C > 1$. This is not very restrictive since we need $T\alpha$ to go to infinity for the mean squared error to go to zero.

Note that Assumption 3(iii) is trivially satisfied if $N/T \to 0$. Indeed, in that case $\hat{\lambda}_j^2 = 0$ for $j > N$. According to Lemma C.5 of Carrasco, Hu and Ploberger (2014), we have

$$\left(\sum_{j=m+1}^{N} \hat{\lambda}_j^2 \right)^2 \leq (N-m) \sum_{j=m+1}^{N} \hat{\lambda}_j^4.$$

Hence,

$$\frac{\left(\frac{1}{T} \sum_{j=m+1}^{T} \hat{\lambda}_j^2 \right)^2}{\frac{1}{T} \sum_{j=m+1}^{T} \hat{\lambda}_j^4} \leq \frac{(N-m)}{T} \to 0.$$

The proof of Proposition 7 is relegated to the Not-for-Publication Appendix (Carrasco and Rossi, 2016). In the proof, we consider separately two cases: the case where the index set of the regularization parameter is discrete and the case where it is continuous. SC and LF have both a discrete index set. For SC, it comes from the fact that q_j is a step function that jumps only at λ_j. For LF, $1/\alpha$ is the number of iterations which is countable. On the other hand, Ridge has a continuous index set. We use the results of Li (1987) for SC and LF and the results of Li (1986) for Ridge.

5 Monte Carlo Analysis

We analyze the performance of the data reduction approaches where the tuning parameter is selected according to criteria (18) and (19) in both large and small samples via Monte Carlo simulations.

5.1 Description of the Data Generating Process

Unless otherwise noted, we consider the following Data Generating Process (DGP):

$$\underset{(N \times 1)}{x_t} = \underset{(N \times r)}{\Lambda} \underset{(r \times 1)}{F_t} + \underset{(N \times 1)}{\xi_t},$$

where F_t is an $(r \times 1)$ vector of $iidN(0, I)$, ξ_t is an $(N \times 1)$ vector of $iidN(0, I)$ random variables, and Λ is an $(N \times r)$ matrix of $iidN(0, I)$; all these variables are uncorrelated with each other. In addition,

$$y_t = \theta' F_t + v_t, \quad t = 1, 2, ..., T, \tag{22}$$

where θ is $(r \times 1)$, $v_t \sim iidN(0, \sigma_v^2)$, and $\sigma_v^2 = 1$. We consider two cases: the large sample case, where $N = 200$ and $T = 500$, and the small sample case, where $N = 100$ and $T = 50$.

In matrix notation, let $X \equiv [x_1, ..., x_T]'$, $F \equiv [F_1, ..., F_T]'$, $\xi \equiv [\xi_1, ..., \xi_T]'$, $y = [y_1, ..., y_T]'$, $v = [v_1, ..., v_T]'$. Also, let $r_{\max}$ be the maximum number of factors in our criteria. Then,

$$\underset{(T \times 1)}{y} = \underset{(T \times r)}{F} \underset{(r \times 1)}{\theta} + \underset{(T \times 1)}{v} \tag{23}$$

$$\underset{(T \times N)}{X} = \underset{(T \times r)}{F} \underset{(r \times N)}{\Lambda'} + \underset{(T \times N)}{\xi}. \tag{24}$$

Equations (23), (24) encompass several interesting cases, ranging from typical factor models, where the variability in X is explained by a few factors (i.e. r is small), to models where a large number of predictors have information at the same time (e.g. r is large). We consider the following cases:

(*DGP 1: Few Factors Structure*) θ is the $(r \times 1)$ vector of ones with $r = 4$; $r_{\max} = r + 10$.[10]

(*DGP 2: Many Factors Structure*) θ is the $(r \times 1)$ vector of ones with $r = 50$; $r_{\max} = \min(N, T/2)$.

(*DGP 3: Five Factors but only One Relevant*) $r = 5$; $r_{\max} = \min(r + 10, \min(N, T/2))$; $\theta = (1, 0_{1\times 4})'$ so that y_t depends only on the first factor while x_t depends on five factors, $F = [F_1, F_2']'$, where F_1 is the relevant factor while F_2 is the vector of four irrelevant factors. The factors are uncorrelated; the first factor has unit variance while the four factors in F_2 have a diagonal covariance matrix with coefficients 2, 3, 3, 4, which makes them dominant relative to the relevant factor (in the sense of having a larger variance). The remaining factors have an identity covariance matrix and are independent of the other factors. In this case, to achieve identification of the factors, we set $y = \widehat{F}\,\theta + v$, where $\widehat{F}$ are the estimated factors from the X generated in the same Monte Carlo simulation, and we set $\sigma_v = 0.1$.

(DGP 4: x_t Has a Factor Structure but Unrelated to y_t) θ is a vector of zeros of dimension $(r \times 1)$ with $r = 5$; $r_{\max} = r + 10$. The factors are uncorrelated and have a diagonal covariance matrix with coefficients 1, 2, 3, 3, 4, similarly to DGP 3.

[10] The minimum number of factors is zero for all methods.

(DGP 5: Eigenvalues Declining Slowly) $r = N$, $r_{\max} = \min(N, T/2)$, θ is an $(N \times 1)$ vector of ones, $\Lambda = M \odot \xi$, where ξ is an $(N \times N)$ matrix of $iidN(0,1)$ random numbers and M is an $(N \times N)$ matrix s.t. $M = \begin{bmatrix} 1, & 1, ..., & 1 \\ 1/2, & 1/2, ..., & 1/2 \\ ... & & ... \\ 1/N & 1/N & 1/N \end{bmatrix}$ and $\odot$ denotes the element-wise product.

(DGP 6: Near Factor Model) $\theta = 1$**,** $r = 1$, $r_{\max} = r + 10$, $\Lambda' = N^{-1/2} 1_{r \times N}$.

DGP 1 describes a situation where the observed data for the $i-th$ cross section unit at time t, $x_{i,t}$, is described by common factors, F_t, and an idiosyncratic component, ξ_t. The number of common factor is four; this is the same set-up as in Bai and Ng (2002), where factor analysis allows substantial dimension reduction. DGP 2 is similar to DGP 1 except that the number of common factors is fifty. DGP 3 describes a situation where the number of common factors of $x_{i,t}$ is five, but only one factor is useful for forecasting y_t.[11] Note that this design is very different from Stock and Watson (2002a), as the latter focus on a situation where the common factors of $x_{i,t}$ are all relevant for forecasting y_t. DGP 4 is similar to DGP 3 except that, although $x_{i,t}$ has five common factors, none of them are relevant for forecasting y_t. DGP 5 describes a situation where many factors are relevant, although their strength is heterogeneous: their importance declines as the number of factors increases. DGP 6 describes a near factor model situation, where the strength of the factors is governed by the rate of decay in N, which we set to 1/2 (the faster the rate of decay, the smaller the factor loading).

5.2 Description of the Methods

In the Monte Carlo experiments, we compare the methods that we propose with traditional factor models, where the number of factors is selected using Bai and Ng's (2002) criterion, eq. (20), which we refer to as "traditional PC". In particular, we use Bai and Ng's (2002) criterion PC_{p2} implemented with a maximum number of factors equal to $r_{\max}$. All factors have been standardized to have unit variance. We use traditional PC as the benchmark, that is, we report results on in-sample mean square error (MSE) and out-of-sample root mean square forecast error (RMSFE) of all the other methods relative to it.

Among the data reduction methods, we consider PC where the number of factors is chosen either by generalized cross-validation (eq. 18), labeled "PC with Cross-Validation", or via Mallows' criterion (eq. 19), labeled "PC with Mallows", where $tr(M_T^{\alpha}) = k$. We also consider Ridge and LF, where the tuning parameter α is chosen by either criteria. When implementing Mallow's criterion in Ridge and LF, we estimate σ_v^2 (that is, the error variance for the "largest" model[12]) using the

[11] The design of the DGP follows Kelly and Pruitt (2015, p. 301).

[12] That is, σ_{ε}^2 in eq. (19).

in-sample errors obtained estimating k using cross-validation; this is because we need a consistent estimate of σ_v^2; in PC we estimate σ_v^2 from the model with all regressors. PLS is implemented by choosing k via leave-one-out cross-validation (see the discussion at the end of Section 4.1).

Forecasts based on principal components have been obtained as follows (Stock and Watson, 2002a): at each time t, we standardize the vector of predictors including observations from time 1 to t (for the recursive estimation method) or from time $(t-R+1)$ to t (for the rolling estimation method); we then determine the number of factors using the various criteria and extract the vector of estimated factors over time, $\{F_j\}_{j=1}^{t}$ for the recursive and $\{F_j\}_{j=t-R+1}^{t}$ for the rolling estimation methods, respectively. The lagged value of the factors are then used to estimate the parameter θ in eq. (22), including a constant (for the recursive estimation method, for example, we used the factors dated at time 1 to $(t-1)$ and the corresponding y's); then, the forecasts for the target variable at time $(t+1)$ are obtained by multiplying the estimated parameter value by F_t.

Forecasts based on Ridge and PLS are obtained similarly by estimating $\hat{\delta}^{\alpha}$ using in-sample data, then forecasting by multiplying it by the regressors at time t. The range of values for α for Ridge and LF was determined by plotting the shape of the GCV and Mallows' objective functions in one Monte-Carlo simulation (note that researchers could choose the range of values in the same way, in practice).[13] When implementing LF, we choose $d = 0.018/\max\left(\lambda^2\right)$.

When forecasting, the GCV criteria, the Mallows criteria, the information criteria and the number of factors are estimated only once, in the first estimation window, to limit computations. In all cases, we allow the number of factors to possibly equal zero.[14]

We also compare our methods with the case where all the variables X are used (labeled "All Regressors") and the case where the number of factors is chosen by AIC or BIC (labeled "PC with AIC" and "PC with BIC", respectively). Forecasts based on all the potential predictors are based on parameter estimates obtained by OLS.[15]

We study both the in-sample fit as well as the forecasting performance of the methods. The out-of-sample forecasts are obtained by using either a recursive window estimation method, starting at time $T/2$, or a rolling estimation method, where the size of the window in the rolling estimation scheme is $T/2$. In all experiments, the number of Monte Carlo replications is 1,000.

[13] The range is as follows: DGP 1: for Ridge, α/N=0:0.001:0.1; and for LF: α/N=0.000001:0.00002:0.0003; DGP 2: for Ridge, α/N=0:0.0001:0.01; and for LF: α/N=0.00001:0.00001:0.0002; DGP 3: for Ridge, α/N=0:0.0005:0.1; and for LF: α/N=0.000001:0.0000025:0.00005; DGP 4: for Ridge, α/N=0:0.001:0.1; and for LF: α/N=0.000001:0.00001:0.0004; DGP 5: for Ridge, α/N=0:0.001:0.15; and for LF: α/N=0.000001:0.00002:0.0004; DGP 6: for Ridge, α/N=0:0.001:0.1; and for LF: α/N=0.000001:0.001:0.016.

[14] That is, for example, in PC we choose the value of k that minimizes the criterion function: $\frac{T^{-1}\left\|y-M_T^{\alpha}y\right\|^2}{\left(1-kT^{-1}\right)^2}$, allowing for the case $k=0$, in which case the value of the criterion function is $T^{-1}y'y$.

[15] Note that we do not report results based on all predictors for cases where $N > T$, as the OLS estimation would not make sense.

5.3 Discussion of the Results

Table 1 reports results for large samples, where $N = 200$ and $T = 500$. The first column in panel A reports the DGP; the second reports the true number of factors, r; the third column reports the number of factors estimated via traditional PC, labeled "Bai-Ng k". Then, for each of the dimension reduction methods, the table reports: the average value (across Monte Carlo simulations) of the tuning parameter, which differs depending on the method; the in-sample MSE of each method relative to traditional PC, labeled "MSER"; and the RMSFE relative to that of traditional PC, labeled "MSFER", either implemented with a recursive ("rec") or a rolling scheme ("roll"). The fourth to seventh columns refer to principal component estimation where the tuning parameter is the number of factors (labeled "k") estimated via GCV. The next three columns report results for PLS. Panel B has a similar structure, but reports results for Ridge and LF; the only difference is that, instead of reporting the number of components, it reports the degrees of freedom (labeled "DoF"), calculated as the trace of M_T^{α} (Hastie, Tibshirani and Friedman, 2009, p. 68). Panel C reports results for the number of factors selected by AIC and BIC (labeled "k") as well as the MSE and the RMSFE of the method relative to traditional PC; the last three columns of panel C report results for the model that includes all regressors. Finally, Panel D reports results for Lasso (labeled "Lasso"), implemented using 10-fold cross-validation (see e.g. Hastie et al., 2009, p. 69).

Table 1 shows that the GCV criterion correctly estimates the number of factors in both situations where the true number of factors is either large or small (DGPs 1 and 2). When the number of factors is small (DGP 1), Bai and Ng's (2002) method also correctly selects the number of factors and the resulting MSE and MSFE are very similar to the ones obtained using GCV. LF improves a lot relative to factor models, while Ridge and PLS perform only slightly worse. AIC and BIC criteria perform similarly to the other information criteria, as it would be expected in the large sample setup of this experiment.

When the number of factors is large but finite (DGP 2), traditional PC selects too many factors. This is due to the fact that the DGP does not satisfy Bai and Ng's (2002) assumptions. While the over-estimated number of factors results in an in-sample MSE that is lower than the one obtained using GCV, it penalizes the forecasts and results in an out-of-sample MSFE that is larger than the one obtained using GCV (or information criteria such as AIC and BIC). Among the three information criteria, cross-validation performs the best, followed very closely by the BIC, then AIC. LF performs even better than principal components with GCV, while PLS and Ridge perform similarly to the latter. Clearly, using all regressors in DGP 2 results in forecasts that perform similarly to traditional PC, as the latter selects the maximum number of factors – see Panel C.

When the number of factors is small but not all factors are relevant predictors of y_t (DGP 3), Bai and Ng (2002) still correctly selects the number of factors in the data, although they include

not only the relevant factors but also the irrelevant ones. Cross-validation slightly over-estimates the number of relevant factors. Among the other data reduction techniques, LF performs quite well in forecasting, while PLS and Ridge perform similarly to traditional PC. We conjecture that the reason why principal components with GCV does not improve much relative to traditional PC might be due to the fact that it is difficult to distinguish the performance of models based on 1 or 5 factors in small samples. The difference should become more visible either by considering smaller sample sizes or by increasing the true number of factors or by decreasing the relevance of the factors. The former will be considered in Table 2; the latter corresponds to DGP 4.

In DGP 4, none of the factors are relevant for explaining y_t, and GCV selects fewer factors than traditional PC. Clearly, five factors again improve in-sample fit but worsen of the out-of-sample forecasting performance. Forecasts based on all regressors are always clearly dominated by data reduction methods in terms of out-of-sample forecast performance. In this design, Ridge and PLS perform similarly in terms of forecasting ability relative to traditional principal components, while LF still performs better than the latter.

DGP 5 describes a situation where the eigenvalues "decline gradually". In this case, traditional PC selects the maximum number of factors, and thus performs the same as a regression that includes all the regressors; GCV, instead, selects on average between eleven and twelve factors, and significantly improves the out-of-sample forecasting performance relative to traditional PC. Similar forecast improvements are obtained by Ridge, LF, PLS and the BIC, while AIC performs worse (although better than traditional PC).

In DGP 6, the eigenvalues are small in magnitude; thus, the DGP too does not satisfy Bai and Ng's (2002) assumptions. This results in cross-validation performing better than Bai and Ng (2002) in forecasting, and similarly to Ridge, AIC and BIC, while again LF performs the best and PLS performs the worst.

Panel D in Table 1 shows that Lasso typically performs no better, and oftentimes worse, than the other data reduction techniques across the simulation designs. Panels E and F instead show that results using Mallows' criterion are similar to using GCV except for DGPs 2 and 5, where it performs worse.

For each DGP, the second line in Table 1 reports the standard deviations; that is, square root of the variance (across Monte Carlo replications) of the estimated number of factors, or the standard deviation of the in-sample Mean Squared Error normalized by 100, $\sqrt{var(MSE_{in})}/100$, or the square root of the out-of-sample RMSFE relative to traditional PC. Note that Bai and Ng's (2002) method seems to show little variability in selecting the number of factors in some cases; this is due to the large sample size and the relatively small variance of the error term.

Table 2 reports the finite sample performance of the methods we consider ($N = 100$, $T = 50$). It is very interesting to see that, in small samples, forecast improvements provided by PC with

GCV are much bigger relative to traditional PC selection methods for DGPs 3 and 4. LF again performs very well. AIC performs poorly, while BIC performs worse than PC with GCV in DGPs 2 and 5, but similarly otherwise. Lasso is competitive when the true number of factors is large (DGP 2 and 5), but performs worse than PC with GCV in several other settings.

INSERT TABLES 1-2 HERE

5.4 Summary

Overall, the main conclusions we draw from the large sample Monte Carlo analysis are that principal components with cross-validation has the potential to improve over traditional PC in many relevant settings, and that further improvements can be obtained by LF, which is a robust method that performs very well across simulation designs. Traditional information criteria (AIC, BIC) perform similarly to principal components with GCV in large samples, but are outperformed by the latter in small samples.

6 Empirical Analysis: Forecasting US Output Growth and Inflation Using Data-Reduction Methods

What are the advantages of forecasting using large datasets of predictors in practice? Clearly, one advantage is that several predictors may simultaneously contain useful information, and using only an ad-hoc subset of them might not extract all the available information. Another compelling reason why forecasts based on large dimensional datasets might perform well is because they might be more robust to instabilities. In fact, it is well-known that the forecasting ability of predictors changes over time (see e.g. Stock and Watson, 2003, and in particular Rossi, 2013): some predictors might be useful in some periods of time but may lose their predictive ability in other periods. Thus, using models where, at each point in time, the forecasters can choose among a large dataset of potential predictors might turn out to guard against instabilities in the predictors' forecasting ability.

We evaluate the empirical performance of data reduction methods for forecasting US real output growth and inflation h-period-ahead using a large dataset of predictors similar to those used in Stock and Watson (2003). We collect quarterly data from January 1959. We use a fixed rolling window estimation scheme with a window size of 40 observations (that is, ten years of data). The target variable is either real GDP growth:

$$Y_{t+h}^h = (400/h)\ln(RGDP_{t+h}/RGDP_t),$$

where $RGDP_t$ is real GDP, or inflation:

$$Y_{t+h}^h = (400/h)\ln(P_{t+h}/P_t) - 400\ln\left(P_t/P_{t-1}\right),$$

where P_t is the price level at time t.

6.1 The Forecasting Models

We consider several competing forecasting models. The benchmark is an autoregressive model:

$$Y_{t+h}^h = \varphi_0 + \varphi(L) Y_t + u_{t+h}^h,\ t = 1, ..., T,$$

where $\varphi(L) = \sum_{j=0}^{p} \varphi_j L^j$, and where the lag length, p, is estimated recursively by the BIC.

The alternative forecasting models based on regularization utilize additional macroeconomic variables Z_t, a (31x1) vector; the models include the following:

- Factor models, where the number of factors is chosen either by the Bai and Ng (2002) information criterion, labeled "Bai and Ng", or by generalized cross-validation, eq. (18), labeled "Cross V.", or by Mallows, eq. (19), labeled "Mallows";

- Ridge, eq. (6), where the tuning parameter is chosen via eqs. (18), (19) (labeled "Ridge");

- PLS, where the tuning parameter is chosen by leave-one-out cross-validation (labeled "PLS");

- LF, where the tuning parameter is chosen via eqs. (18), (19) (labeled "LF");

- We also include equal-weight forecast combinations based on the dimension reduction methods implemented with either GCV or Mallows' criteria, that is Ridge, factor models and LF (labeled "CV Comb." and "Mall. Comb.", respectively).[16]

For these alternative models, we consider two specifications that produce direct forecasts. The first is a specification where the pool of regressors contains only h-period lagged predictors ($x_{t-h} = Z_{t-h}$); the second is a specification where the pool of regressors contains additional two lags of the predictors (x_{t-h} includes Z_{t-h},Z_{t-h-1}, and Z_{t-h-2}).[17] The gains from regularization techniques are stronger when the predictors are correlated; thus, we expect the regularization techniques to perform better (relative to traditional techniques) in the latter case.

The tuning parameters for all the methods are chosen recursively.

We also consider other techniques that have been proposed to deal with large dimensional datasets, such as Bayesian Model Averaging (labeled "BMA") and equal weight forecast combinations (labeled "Comb."). The performance of BMA has been investigated by Wright (2009) for forecasting inflation, while equal weight forecast combinations have been recently discussed by Timmermann (2006) and Rossi (2013), who find that they are two of the toughest benchmarks to beat when forecasting output growth and inflation in the U.S. In implementing equal weight forecast combinations, we estimate autoregressive distributed lag models using several economic explanatory

[16] The grid for the tuning parameter is chosen following the same criteria as in the Monte Carlo section; in particular, the grid for LF is 0.00001:0.0001:0.005 and the grid for Ridge is α/N =0.0001:0.01:0.08. When implementing LF, we choose $d = 0.018/\max(\lambda^2)$. We do not consider Lasso given its relatively poor performance in the simulations.

[17] We only include two additional lags of the predictors as our sample is small.

variables one-at-a-time (where the additional lags of the additional explanatory economic variables are chosen by the BIC) – see Rossi and Sekhposyan (2014) for details on the estimation procedure and the data.

6.2 Empirical Results

The empirical results for the techniques implemented with generalized cross-validation are reported in Panel I in Tables 3 and 4, while those for Mallows' criterion are reported in Panel II in the same tables. Table 3 reports results for the specification that contains only h-period lagged predictors, while table 4 reports results for the specification that includes two additional lags. Panel A in the tables reports results for one-quarter ahead forecasts, while panel B reports results for one-year-ahead forecasts. The tables show the ratio of the RMSFE of the various regularization methods relative to that of the autoregressive benchmark model; values less than unity favor the regularization method relative to the benchmark. The p-value of the Diebold and Mariano (1995) and West (1996) test of equal predictive ability is reported in parenthesis.

The results in Table 3, Panel A, show that BMA and equal weight forecast combinations are the best models for one-month ahead forecasts; in particular, BMA is the best for inflation and forecast combination is the best for output growth. Note that two other models outperform the AR benchmark: Ridge when forecasting output, and LF when forecasting inflation. Thus, the empirical evidence points to the fact that using only a few factors may not provide competitive forecasts, and that potentially useful information is spread out among several predictors; this is also consistent with the finding that forecast combinations and BMA perform very well. When forecasting one-year ahead, Ridge is the best model for forecasting output growth while BMA remains the best model for forecasting inflation.

When comparing factor models, selection based on generalized cross-validation typically performs better than traditional factor models for output growth at both one and four quarters horizons and for inflation at the one-quarter horizon, although the BIC performs even better.

Interestingly, Table 3 also shows that equal weight forecast combinations among the dimension reduction techniques ("CV Comb." and "Mall. Comb.") provide further gains. In particular, forecast combinations implemented among dimension reduction methods where the parameter is chosen via Mallows' criterion perform even better than BMA when forecasting inflation at the four quarter horizon. Such combinations across data reduction methods achieved via GCV further improve upon Ridge (the best forecasting model) when forecasting GDP growth four quarters ahead.

Table 4 shows that further gains are obtained by adding additional lags of the predictors, in the sense that the RMSFE of dimension reduction techniques tends to decrease, sometimes substantially, especially when forecasting output growth. When forecasting output growth one-quarter

ahead, the forecasting performance of the regularization methods improves substantially: the best regularization method is Ridge, which improves significantly (at the 10% level) on the autoregressive model, while still performing worse than forecast combinations. When forecasting output growth one-year ahead, instead, augmenting the set of predictors with lags substantially improves the performance of factor models selected via GCV, which performs better than forecast combinations, and improves even further the performance of Ridge implemented with Mallows selection criteria, which is the best forecasting model. There are no substantial improvements from adding additional lags when forecasting inflation at short horizons, however. Again, combining forecasts based on regularization techniques generate further gains. In fact, when including additional lags, forecasts based on combinations of regularization methods is the best for forecasting GDP growth at short horizons and are quite competitive for forecasting inflation at longer horizons as well.

Overall, regularization techniques help in forecasting output growth one-year-ahead, even relative to traditional forecast combinations, and Ridge provides significant forecast improvements that are close to those obtained via forecast combination. On the other hand, forecasting inflation is very difficult; in a way, this result is not surprising, given the well-known result that inflation, during this period, was essentially unforecastable. While regularization techniques do not help too much in forecasting inflation when considered individually, they become very competitive at longer horizons when combined.

INSERT TABLES 3-4 HERE

6.3 Can Large Datasets of Predictors Effectively Guard Against Instabilities?

As previously mentioned, one of the advantages of using a large dataset of predictors is to guard against instabilities in specific predictors' forecasting ability. That is, data-reduction methods produce forecasts that are robust to the fact that specific predictors might lose their forecasting ability over time since they extract information on all potential predictors at each point in time. We investigate the robustness of the forecasting performance of the data-reduction methods in two ways. The first is by analyzing the stability of their forecasting performance relative to the benchmark. In fact, if data-reduction methods forecast better because they guard against time variation in predictors' forecasting ability, then they should perform better consistently over time. We check whether this is the case by testing whether forecasts based on data reduction methods perform better than the benchmark over time using Giacomini and Rossi's (2010) Fluctuation test. The second is by analyzing the stability of the properties of the forecasts themselves, such as rationality. Again, the advantage of data reduction methods should result in forecasts that are highly correlated with the actual realizations consistently over time. We can test whether this is the case by using tests for forecast rationality robust to instabilities (Rossi and Sekhposyan, 2015).

For brevity, we consider only the case of h-period lagged predictors ($X_{t-h} = Z_{t-h}$).[18]

Figures 1-4 display the analysis of models' relative predictive ability over time. The figures plot Giacomini and Rossi's (2010) Fluctuation test statistic, a test of equal predictive ability repeated in rolling windows over the out-of-sample portion of the data. The size of the rolling window is 60 observations, that is fifteen years of data. The test statistic is denoted with a continuous line, while the critical values are denoted by dotted lines. Negative values of the test statistic denote situations in which the model forecasts better than the autoregressive benchmark, in the sense that its RMSFEs are lower than those of an autoregressive model in the previous 60 observations. On the other hand, positive values above the critical value line indicate that the model performs significantly worse than the autoregressive benchmark. The figures indeed reveal that factor model forecasts based on GCV are consistently better than those based on traditional factor models over time. In fact, comparing the upper two panels in Figure 1, it is clear that forecasts based on factor models selected via traditional criteria perform significantly worse than the autoregressive benchmark, while those selected via GCV do not; furthermore, the relative RMSFE of the latter over time is constantly below that of the former. The figures also show that BMA, LF, Ridge and combinations can effectively produce forecasts that often perform well consistently over time.

INSERT FIGURES 1-4 HERE

To summarize, the best performing models are forecast combinations, Ridge and BMA. However, while these models perform better than an autoregressive benchmark, nothing guarantees that their forecasts are unbiased, nor that they satisfy some minimal requirements that typically characterize rational forecasts. Tables 5 and 6 display results based on the traditional Mincer and Zarnowitz (1969) test. We regress the forecast error onto a constant and the forecast; if both coefficients are zero, forecasts are rational.[19] The results are encouraging: forecasts based on Ridge, LF, combinations based on data-reduction methods, traditional forecast combinations and BMA are typically rational for output growth at all horizons. Only BMA and GCV-based forecast combinations are rational for inflation at all horizons, although several data-reduction methods produce rational forecast at longer horizons.[20]

INSERT TABLE 5 AND 6 HERE

However, it is well-known that forecasts of output and inflation are unstable over time (Stock and Watson, 1996), and instabilities invalidate the Mincer and Zarnowitz (1969) test (Rossi, 2013).

[18] Results for the case with additional lags are reported in the Not-for-Publication Appendix.

[19] The test is implemented using a HAC robust covariance estimator based on Newey and West (1997) with four lags.

[20] LF and forecast combinations are rational for inflation only for four-quarter-ahead forecasts.

In addition, we are particularly interested in evaluating whether forecasts based on data-reduction techniques are rational *systematically* over time. This can be achieved by using Rossi and Sekhposyan's (2015) Fluctuation Rationality test, which is a forecast rationality test repeated in rolling windows over the out-of-sample portion of the data. Figures 5 to 8 display the Fluctuation Rationality test (solid line) together with its critical value (dotted line). When the test statistic is above the critical value line, we conclude that the forecast is not rational. It is reassuring to see that forecasts that are rational according to Tables 5 and 6 are also rational over time; this again supports our conjecture that data reduction methods may effectively guard against instabilities in predictors' forecasting ability over time.

INSERT FIGURES 5-8 HERE

7 Conclusions

This paper investigates whether dimension reduction techniques such as principal components, Ridge, Landweber-Fridman and partial least squares implemented with generalized cross-validation and Mallows's criteria have potential in forecasting. Theoretical results show that the three main methods (R, LF and SC) have the same rate of convergence when the signal is difficult to recover, whereas LF and SC have a faster rate than R when the signal is easy to recover. Moreover, LF and SC reach the fastest possible rate in the case of the factor model. Monte Carlo simulation results show that these alternative data reduction techniques can be potentially useful in environments where either the data do not have a factor structure, or where there is a factor structure but the factors are not strong predictors for the target variable.

Empirical results show that, when forecasting output growth and inflation in the US, forecasts of output growth based on Ridge and LF are rational; furthermore, Ridge has the potential to improve relative to an autoregressive benchmark and also, in some cases, relative to tougher benchmarks such as forecast combination and BMA. It is reassuring to see that, when Ridge performs well, it does so systematically over time. For inflation, the best forecasting model remains BMA. However, substantial forecast improvements can be obtained in some cases by combining data reduction methods.

It might be possible that using other techniques to extract factors (such as Forni et al., 2015) might guard against instabilities and perform well when forecasting in practice, although we leave this issue for future research.[21]

[21] Onatski (2015) studied the quality of approximation of the principal component estimate when the number of factors might be misspecified. We are not interested in the quality of approximation of the principal component; rather, we focus on the quality of approximation of the predicted target variable. In other words, Onatski (2015) studies $\text{tr}\left[(\widehat{\Lambda}\widehat{F}_t' - \Lambda F')(\widehat{\Lambda}\widehat{F}_t' - \Lambda F')'\right]/NT$, while we study the mean square prediction error of y, $\left\|X\hat{\delta}^{\alpha} - X\delta\right\|^2$.

8 References

Amemiya, T. (1966) "On the Use of Principal Components of Independent Variables in Two-stage Least-Squares Estimation", *International Economic Review* 7, 283-303.

Bai, J. (2003) "Inferential Theory for Factor Models of Large Dimensions", *Econometrica* 71(1), 135-171.

Bai, J. and S. Ng (2002) "Determining the Number of Factors in Approximate Factor Models", *Econometrica* 70(1), 191-221.

Bai, J. and S. Ng (2006) "Evaluating Latent and Observed Factors in Macroeconomics and Finance," *Journal of Econometrics* 131, 507-537.

Bai, J. and S. Ng (2008) "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics* 146, 304-317.

Bai, J. and S. Ng (2009) "Boosting Diffusion Indices," *Journal of Applied Econometrics* 4, 607-629.

Barbarino, A. (2015) "Sufficient Reductions in Dynamic Factor Models," *mimeo*, Federal Reserve Board.

Berg, C. and R. Szwarc (2011) "The smallest eigenvalue of Hankel matrices," *Constructive Approximation* 34, 107-133.

Blazère, M., J.-M. Loubes and F. Gamboa (2014a) "Partial Least Square: A New Statistical Insight through the Prism of Orthogonal Polynomials," arXiv:1405.5900v1.

Blazère, M., J.-M. Loubes and F. Gamboa (2014b) "A Unified Framework for the Study of PLS Estimator's Properties," arXiv:1411.0229v1.

Brown, S. (1989) "The Number of Factors in Security Returns," *The Journal of Finance*, Vol. XLIV, 1247-1262.

Carrasco, M. and B. Rossi (2016), "Not-for-Publication Appendix to: In-sample Inference and Forecasting in Misspecified Factor Models", *mimeo*, available at: www.econ.upf.edu/~brossi/ CarrascoRossi_Appendix.pdf.

Carrasco, M., L. Hu and W. Ploberger (2014) Supplement to "Optimal Test for Markov Switching Parameters," *Econometrica* 82(2), 765–784, *Econometrica Supplementary Material*.

Carrasco, M., J. P. Florens, and E. Renault (2007) "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization", *Handbook of Econometrics*, Vol. 6B, edited by J.J. Heckman and E.E. Leamer.

Cai, T. and P. Hall (2006) "Prediction in functional linear regression," *Annals of Statistics*, 34, 2159-2179.

Chen, X, and H. White (1998) "Central limit and functional central limit theorems for Hilbert-valued dependent heterogenous arrays with applications," *Econometric Theory*, 14, 260-284.

Cheng, X. and B.E. Hansen (2015) "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics* 186, 280-293.

Dauxois, Pousse, and Romain (1982) "Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference," *Journal of Multivariate Analysis*, 12, 136-154.

De Jong, S. (1993) "PLS fits closer that PCR," *Journal of Chemometrics* 7(6) 551-557.

De Mol, C., D. Giannone and L. Reichlin (2008) "Forecasting Using a Large Number of Predictions: Is Bayesian Shrinkage a Valid Alternative to Principal Components?", *Journal of Econometrics* 146, 318-328.

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

Djogbenou, A. (2015) "Model Selection in Factor-Augmented Regressions with Estimated Factors," *mimeo*, University of Montreal.

Doukhan, P. (1994) *Mixing*, Springer-Verlag, New York.

Engl, H., M. Hanke, and A. Neubauer (2000) *Regularization of Inverse Problem*s, Kluwer Academic Publishers, Dordrecht.

Fan, J., J. Lv, and L. Qi (2011) "Sparse High Dimensional Models in Economics," *Annual Review in Economics* 3, 291-317.

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005), "The Generalized Dynamic Factor Model: One-sided Estimation and Forecasting", *Journal of the American Statistical Association* 100, 830-840.

Forni, M., A. Giovannelli, M. Lippi and S. Soccorsi (2015) "Dynamic Factor Model with Infinite-Dimensional Factor Space: Forecasting", *mimeo*, Einaudi Institute.

Fuentes, J., P. Poncela and J. Rodriguez (2015), "Sparse Partial Least Squares in Time Series for Macroeconomic Forecasting", *Journal of Applied Econometrics* 30(4), 576-595.

Ghysels, E., P. Santa-Clara, and R. Volkanov (2005) "There is a risk-return trade-off after all", *Journal of Financial Econometrics*, 76, 509-548.

Giacomini, R. and B. Rossi (2010) "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics* 25(4), 595-620.

Giovannelli, A. and T. Proietti (2015) "On the Selection of Common Factors for Macroeconomic Forecasting", *mimeo*, University Tor Vergata.

Gonçalves, S., M. McCracken, and B. Perron (2015) "Tests of Equal Accuracy for Nested Models with Estimated Factors", *mimeo*, Université de Montréal.

Groen, J. and G. Kapetanios (2008) "Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting", *mimeo*, Queen Mary University.

Groen, J. and G. Kapetanios (2013) "Model Selection Criteria for Factor-Augmented Regres-

sions", *Oxford Bulletin of Economics and Statistics* 75(1), 37-63.

Hall, P. and J. Horowitz (2007) "Methodology and convergence rates for functional linear regression," *Annals of Statistics*, 35, 70-91.

Hastie, T., R. Tibshirani, and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag.

Helland, I. (1988) "On the Structure of Partial Least-Squares Regression", *Communications in Statistics Simulation and Computation* 17(2), 581-607.

Hillebrand, E. and T. Lee (2012) "Stein-Rule Estimation and Generalized Shrinkage Methods for Forecasting Using Many Predictors", in: Millimet, D. and Terrell, D. (eds.), *Advances in Econometrics,* Volume 30, Chapter 6, 171-196, Emerald Publishers.

Huang, H. and T.-H. Lee (2010) "To Combine Forecasts or To Combine Information?", *Econometric Reviews* 29, 534-570.

Inoue, A. and L. Kilian (2008) "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation", *Journal of the American Statistical Association* 103(482), 511-522.

Kelly, B. and S. Pruitt (2015) "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors", *Journal of Econometrics* 186(2), 277-476.

Kim, H.H. and N.R. Swanson (2014a) "Forecasting Financial and Macroeconomic Variables using Data Reduction Methods: New Empirical Evidence", *Journal of Econometrics*, 178, 352-367.

Kim, H.H. and N.R. Swanson (2014b) "Mining Big Data Using Parsimonious Factor and Shrinkage Methods", *mimeo*, Rutgers University.

Kim, H.H. and N.R. Swanson (2015) "Methods for Pastcasting, Nowcasting and Forecasting Using Factor-MIDAS with an Application to Real-Time Korean GDP", *mimeo*, Rutgers University.

Kress, R. (1999) *Linear Integral Equations*, Springer.

Ledoit, O. and M. Wolf (2004) "A Well-conditioned Estimator for Large-Dimensional Covariance Matrices", *Journal of Multivariate Analysis* 88, 365-411.

Li, K-C.(1986) "Asymptotic Optimality of C_L and Generalized Cross-validation in Ridge Regression with Application to Spline Smoothing", *The Annals of Statistics* 14, 1101-1112.

Li, K-C.(1987) "Asymptotic optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set", *The Annals of Statistics* 15, 958-975.

Mallows, C.L. (1973) "Some Comments on C_p", *Technometrics* 15, 661-675.

Mao Takongmo, C. and D. Stevanovic (2014) "Selection of the Number of Factors in Presence of Structural Instability: a Monte Carlo study", *L'Actualité économique* 91, 177-233.

Mincer, J. and V. Zarnowitz (1969) "The Evaluation of Economic Forecasts," in J. Mincer (ed.), *Economic Forecasts and Expectations*, New York: National Bureau of Economic Research.

Newey, W.K. and K.O. West (1987) "A Simple, Positive Semi-Definite, Heteroskedasticity and

Autocorrelation Consistent Covariance Matrix", *Econometrica* 55(3), 703-708.

Ng, S. (2013) "Variable Selection in Predictive Regressions". In: Elliott G, Timmermann A. (eds.), *Handbook of Economic Forecasting*, Vol. 2B, Elsevier-North-Holland: Amsterdam.

Onatski, A. (2015) "Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models", *Journal of Econometrics* 186, 388-406.

Phatak, A. and F. de Hoog (2002) "Exploiting the Connection between PLS, Lanczos Methods and Conjugate Gradients: Alternative Proofs of Some Properties of PLS", *Journal of Chemometrics* 16(7), 361-367.

Rossi, B. (2013) "Advances in Forecasting Under Instabilities". In: Elliott G, Timmermann A. (eds.), *Handbook of Economic Forecasting*, Vol. 2B, Elsevier-North-Holland: Amsterdam.

Rossi, B. and T. Sekhposyan (2014) "Evaluating Predictive Densities of US Output Growth and Inflation in a Large Macroeconomic Data Set", *International Journal of Forecasting* 30(3), 662-682.

Rossi B. and Sekhposyan T. (2015) "Forecast Rationality Tests in the Presence of Instabilities, with Applications to Federal Reserve and Survey Forecasts", *Journal of Applied Econometrics*, forthcoming.

Stock, J.H. and M.W. Watson (1996) "Evidence on Structural Instability in Macroeconomic Time Series Relations", *Journal of Business and Economic Statistics* 14(1), 11-30.

Stock, J. and M. Watson (2002a) "Macroeconomic Forecasting Using Diffusion Indexes", *Journal of Business and Economic Statistics* 20(2), 147-162.

Stock, J. and M. Watson (2002b) "Forecasting Using Principal Components From a Large Number of Predictors", *Journal of the American Statistical Association* 97(460), 1167-79.

Stock, J. and M. Watson (2003) "Forecasting Output and Inflation: The Role of Asset Prices", *Journal of Economic Literature* 41(3), 788-829.

Stock, J. and M. Watson (2006) "Forecasting With Many Predictors". In: Elliott G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 1, Amsterdam: Elsevier-North Holland, 135–196.

Stock, J. and M. Watson (2012) "Generalized Shrinkage Methods for Forecasting Using Many Predictors", *Journal of Business and Economic Statistics* 30(4), 481-493.

Tikhonov, A.N. and V.Y. Arsenin (1977) *Solutions of Ill-Posed Problems*, Winston, New York.

Timmermann, A. (2006) "Forecast Combinations". In: Elliott G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 1, Amsterdam: Elsevier-North Holland, 135–196.

Tu, Y. and T. Lee (2013) "Forecasting Using Supervised Factor Models", *mimeo*, University of California, Riverside.

West, K.D. (1996) "Asymptotic Inference about Predictive Ability," *Econometrica* 64(5), 1067-

1084.

Wright, J.H. (2009) "Forecasting US Inflation by Bayesian Model Averaging", *Journal of Forecasting* 28(2), 131-144.

Tables and Figures

Table 1 Panel A (GCV, N=200, T=500)

		Bai-Ng	PC with Cross-Validation				PLS			
	r	k	k	MSER	RMSFER		k	MSER	RMSFER	
					rec	roll			rec	roll
DGP 1	4.00	4.00	4.83	1.00	1.01	1.02	3.16	0.93	1.05	1.06
(s.e.)	–	0.00	1.90	1.00	1.33	1.74	0.40	1.45	1.28	1.35
DGP 2	50.00	200.00	50.96	1.50	0.49	0.25	8.99	1.44	0.48	0.25
(s.e.)	–	0.00	2.59	1.35	0.68	0.32	0.61	1.26	0.48	0.19
DGP 3	5.00	5.61	1.93	1.00	1.00	0.99	2.12	0.99	1.01	1.01
(s.e.)	–	0.71	2.11	1.02	0.99	1.00	0.47	1.08	1.06	1.10
DGP 4	5.00	5.61	0.86	1.01	0.99	0.98	1.36	0.99	1.01	1.01
(s.e.)	–	0.75	1.90	1.02	1.00	0.99	0.89	1.20	1.12	1.18
DGP 5	200.00	200.00	11.94	1.70	0.45	0.21	2.38	1.57	0.46	0.22
(s.e.)	–	0.00	8.27	1.54	0.40	0.16	0.80	2.12	0.44	0.20
DGP 6	1.00	0.00	6.51	0.81	0.86	0.89	2.00	0.50	1.04	1.09
(s.e.)	–	0.00	3.38	0.84	0.87	0.91	0.00	0.60	1.23	1.37

Table 1 Panel B (GCV, N=200, T=500)

		Bai-Ng	Ridge					LF				
	r	k	α	DoF	MSER	RMSFER		α	DoF	MSER	RMSFER	
						rec	roll				rec	roll
DGP 1	4.00	4	1.56	30.99	0.92	1.03	1.02	0.00	17.48	0.96	0.92	0.90
(s.e.)	–	0.00	0.31	4.98	1.04	1.06	1.06	0.00	5.06	1.03	0.94	0.92
DGP 2	50.00	200	0.06	84.38	1.45	0.54	0.26	0.00	61.84	1.56	0.33	0.14
(s.e.)	–	0.00	0.01	5.83	2.66	0.63	0.23	0.00	4.88	2.56	0.34	0.12
DGP 3	5.00	5.61	4.27	11.96	0.97	1.01	1.00	0.00	8.55	0.99	0.96	0.94
(s.e.)	–	0.71	2.17	7.75	1.13	1.02	1.00	0.00	6.73	1.09	1.00	1.02
DGP 4	5.00	5.61	17.94	6.75	0.99	1.00	0.99	0.00	22.58	0.94	0.91	0.88
(s.e.)	–	0.75	5.85	8.19	1.13	1.01	0.99	0.00	3.10	0.97	0.92	0.89
DGP 5	200.00	200	3.25	35.50	1.54	0.44	0.21	0.00	68.11	1.36	0.33	0.14
(s.e.)	–	0.00	3.13	12.19	1.67	0.40	0.16	0.00	4.24	1.14	0.31	0.12
DGP 6	1.00	0	1.52	73.12	0.62	0.89	0.89	0.01	70.99	0.62	0.59	0.54
(s.e.)	–	0.00	0.32	8.71	0.87	0.90	0.90	0.00	5.30	0.76	0.61	0.58

Table 1 Panel C (N=200, T=500)

		Bai-Ng	PC with AIC				PC with BIC				All Regressors		
	r	k	k	MSER	RMSFER		k	MSER	RMSFER		MSER	RMSFER	
					rec	roll			rec	roll		rec	roll
DGP 1	4.00	4.00	4.88	0.99	1.01	1.02	3.94	1.00	1.02	1.04	0.60	2.44	5.07
(s.e.)	–	0.00	1.98	1.00	1.33	1.74	0.30	1.00	1.58	2.08	0.76	2.53	6.59
DGP 2	50.00	200.00	51.73	1.50	0.71	0.56	49.88	1.51	0.49	0.26	1.00	1.00	1.00
(s.e.)	–	0.00	4.29	1.37	2.75	2.85	0.50	1.34	0.91	0.45	1.00	1.00	1.00
DGP 3	5.00	5.61	1.95	1.00	1.00	0.99	1.01	1.01	0.99	0.99	0.60	2.44	5.05
(s.e.)	–	0.71	2.13	1.02	0.99	1.00	0.16	1.00	0.99	1.01	0.80	2.70	6.97
DGP 4	5.00	5.61	1.88	1.00	0.99	0.99	1.02	1.01	0.99	0.98	0.61	2.43	5.04
(s.e.)	–	0.75	1.97	1.01	1.00	0.99	0.15	1.00	0.99	0.98	0.77	2.55	6.79
DGP 5	200.00	200.00	13.45	1.69	0.85	0.76	1.62	1.82	0.45	0.21	1.00	1.00	1.00
(s.e.)	–	0.00	10.12	1.57	2.65	2.97	1.26	1.41	0.39	0.16	1.00	1.00	1.00
DGP 6	1.00	0.00	6.58	0.81	0.86	0.89	1.84	0.83	0.88	0.91	0.45	1.86	3.87
(s.e.)	–	0.00	3.38	0.84	0.87	0.90	1.45	0.87	0.90	0.94	0.58	1.97	5.18

Table 1 Panel D (N=200, T=500)

		Bai-Ng	LASSO			
	r	k	λ_V	MSER	RMSFER	
					rec	roll
DGP 1	4.00	4.00	38.63	0.93	1.09	1.09
(s.e.)	–	0.00	11.30	1.45	1.16	1.15
DGP 2	50.00	200.00	102.10	1.44	0.61	0.29
(s.e.)	–	0.00	13.33	1.26	0.71	0.26
DGP 3	5.00	5.61	12.29	0.99	1.01	1.01
(s.e.)	–	0.71	6.46	1.08	1.03	1.01
DGP 4	5.00	5.61	2.56	0.99	0.99	0.98
(s.e.)	–	0.75	5.77	1.20	1.01	1.00
DGP 5	200.00	200.00	20.73	1.57	0.45	0.21
(s.e.)	–	0.00	16.04	2.12	0.40	0.16
DGP 6	1.00	0.00	79.05	0.50	0.97	0.97
(s.e.)	–	0.00	21.70	0.60	1.00	1.00

Table 1 Panel E (Mallows, N=200, T=500)

		Bai-Ng	PC with Mallows			
	r	k	k	MSER	RMSFER	
					rec	roll
DGP 1	4.00	4.00	14.00	0.98	1.03	1.04
(s.e.)	–	0.00	0.00	0.98	1.03	1.06
DGP 2	50.00	200.00	200.00	1.00	1.00	1.00
(s.e.)	–	0.00	0.00	1.00	1.00	1.00
DGP 3	5.00	5.61	15.00	0.98	1.03	1.04
(s.e.)	–	0.71	0.00	0.99	1.03	1.04
DGP 4	5.00	5.61	0.14	1.01	0.99	0.98
(s.e.)	–	0.75	1.42	1.01	0.99	0.98
DGP 5	200.00	200.00	130.20	1.27	1.00	1.00
(s.e.)	–	0.00	95.38	5.09	1.00	1.00
DGP 6	1.00	0.00	11.00	0.80	0.86	0.89
(s.e.)	–	0.00	0.00	0.82	0.87	0.90

Table 1 Panel F (Mallows, N=200, T=500)

		Bai-Ng	Ridge					LF				
	r	k	α	DoF	MSER	RMSFER		α	DoF	MSER	RMSFER	
						rec	roll				rec	roll
DGP 1	4.00	4.00	1.62	29.84	0.92	1.03	1.02	0.00	16.92	0.96	0.93	0.90
(s.e.)	–	0.00	0.26	3.60	1.00	1.06	1.06	0.00	3.80	1.00	0.95	0.94
DGP 2	50.00	200.00	0.05	89.86	1.42	0.54	0.26	0.00	65.51	1.53	0.32	0.14
(s.e.)	–	0.00	0.01	5.19	2.43	0.60	0.22	0.00	9.44	2.55	0.39	0.14
DGP 3	5.00	5.61	4.35	10.97	0.98	1.01	1.00	0.00	8.03	0.99	0.96	0.94
(s.e.)	–	0.71	2.04	5.63	1.07	1.00	0.99	0.00	5.97	1.08	1.03	1.05
DGP 4	5.00	5.61	18.34	5.77	0.99	1.00	0.99	0.00	22.10	0.94	0.92	0.88
(s.e.)	–	0.75	5.24	5.67	1.07	0.98	0.97	0.00	1.91	0.95	0.92	0.89
DGP 5	200.00	200.00	3.63	32.29	1.56	0.44	0.21	0.00	68.12	1.36	0.33	0.14
(s.e.)	–	0.00	3.30	10.65	1.60	0.39	0.16	0.00	4.29	1.14	0.31	0.12
DGP 6	1.00	0.00	1.83	65.10	0.65	0.89	0.89	0.01	66.73	0.63	0.60	0.55
(s.e.)	–	0.00	0.29	6.22	0.80	0.93	0.93	0.00	2.24	0.68	0.59	0.54

Table 2 Panel A (GCV, N=100, T=50)

		Bai-Ng	PC with Cross-Validation				PLS			
	r	k	k	MSER	RMSFER		k	MSER	RMSFER	
					rec	roll			rec	roll
DGP 1	4.00	6.18	4.57	1.01	0.75	0.58	4.29	0.62	0.77	0.55
(s.e.)	–	0.74	1.81	1.07	0.92	0.65	3.58	1.52	0.81	0.50
DGP 2	50.00	24.00	17.47	1.49	0.05	0.01	21.95	0.04	0.01	0.00
(s.e.)	–	0.07	6.49	2.92	0.03	0.01	12.48	0.27	0.00	0.00
DGP 3	5.00	8.54	1.97	1.12	0.61	0.45	3.22	0.93	0.67	0.46
(s.e.)	–	0.85	2.21	1.17	0.63	0.46	3.67	1.77	0.67	0.38
DGP 4	5.00	8.61	1.02	1.15	0.58	0.40	1.99	0.99	0.63	0.44
(s.e.)	–	0.88	2.32	1.19	0.65	0.46	2.96	1.68	0.68	0.38
DGP 5	100.00	17.64	2.49	1.49	0.34	0.03	2.59	0.95	0.00	0.00
(s.e.)	–	0.82	3.83	1.57	0.69	0.02	3.58	2.08	0.00	0.00
DGP 6	1.00	0.01	3.23	0.79	0.96	0.89	2.46	0.18	0.98	0.85
(s.e.)	–	0.10	3.14	1.02	1.01	0.95	3.38	0.64	1.05	0.83

Table 2 Panel B (GCV, N=100, T=50)

		Bai-Ng	Ridge					LF				
	r	k	α	DoF	MSER	RMSFER		α	DoF	MSER	RMSFER	
						rec	roll				rec	roll
DGP 1	4.00	6.18	1.74	16.89	0.74	0.79	0.56	0.00	11.21	0.90	0.42	0.28
(s.e.)	–	0.74	0.97	11.03	1.76	0.86	0.52	0.00	9.72	1.62	0.58	0.35
DGP 2	50.00	24.00	0.35	34.05	0.39	0.01	0.00	0.00	29.87	0.58	0.00	0.00
(s.e.)	–	0.07	0.33	7.72	1.21	0.00	0.00	0.00	8.02	1.73	0.00	0.00
DGP 3	5.00	8.54	2.14	15.41	0.78	0.69	0.46	0.01	9.28	0.97	0.41	0.26
(s.e.)	–	0.85	1.49	14.96	2.04	0.74	0.42	0.00	12.82	1.79	0.57	0.31
DGP 4	5.00	8.61	74.63	2.86	1.11	0.58	0.39	0.00	20.27	0.59	0.23	0.12
(s.e.)	–	0.88	36.88	6.30	1.46	0.63	0.36	0.00	12.40	1.39	0.45	0.22
DGP 5	100.00	17.64	9.47	10.80	1.12	0.00	0.00	0.01	12.35	1.11	0.00	0.00
(s.e.)	–	0.82	6.05	10.04	1.92	0.00	0.00	0.01	14.49	2.24	0.00	0.00
DGP 6	1.00	0.01	6.02	18.24	0.45	0.93	0.82	0.02	17.79	0.45	0.35	0.29
(s.e.)	–	0.10	5.37	12.84	1.44	0.96	0.81	0.01	12.70	1.30	0.86	0.70

Table 2 Panel C (N=100, T=50)

		Bai-Ng	PC with AIC				PC with BIC			
	r	k	k	MSER	RMSFER		k	MSER	RMSFER	
					rec	roll			rec	roll
DGP 1	4.00	6.18	5.17	0.98	0.79	0.64	3.83	1.05	0.78	0.60
(s.e.)	–	0.74	2.59	1.09	0.98	0.80	0.75	1.05	1.04	0.72
DGP 2	50.00	24.00	21.27	1.12	0.88	0.05	9.07	2.93	0.81	0.01
(s.e.)	–	0.07	4.09	1.66	0.99	0.04	8.39	5.32	0.99	0.01
DGP 3	5.00	8.54	2.51	1.10	0.69	0.54	1.11	1.17	0.60	0.45
(s.e.)	–	0.85	3.15	1.22	0.87	0.74	0.47	1.10	0.63	0.46
DGP 4	5.00	8.61	2.63	1.10	0.69	0.54	1.09	1.18	0.57	0.40
(s.e.)	–	0.88	3.20	1.22	0.90	0.80	0.45	1.10	0.61	0.45
DGP 5	100.00	17.64	6.78	1.29	0.93	0.06	1.37	1.60	0.82	0.05
(s.e.)	–	0.82	7.66	1.85	1.00	0.03	1.00	1.35	0.97	0.02
DGP 6	1.00	0.01	4.30	0.76	1.01	0.96	1.58	0.87	0.94	0.86
(s.e.)	–	0.10	3.32	0.98	1.09	1.10	1.22	0.97	0.97	0.88

Table 2 Panel D (N=100, T=50)

		Bai-Ng	LASSO			
	r	k	λ_V	MSER	RMSFER	
					rec	roll
DGP 1	4.00	6.18	17.14	0.62	0.91	0.64
(s.e.)	–	0.74	8.39	1.52	1.03	0.63
DGP 2	50.00	24.00	38.66	0.04	0.02	0.00
(s.e.)	–	0.07	8.59	0.27	0.00	0.00
DGP 3	5.00	8.54	11.52	0.93	0.73	0.48
(s.e.)	–	0.85	7.74	1.77	0.79	0.45
DGP 4	5.00	8.61	3.60	0.99	0.58	0.39
(s.e.)	–	0.88	8.00	1.68	0.73	0.41
DGP 5	100.00	17.64	6.24	0.95	0.00	0.00
(s.e.)	–	0.82	9.15	2.08	0.00	0.00
DGP 6	1.00	0.01	9.96	0.18	0.96	0.84
(s.e.)	–	0.10	12.17	0.64	1.05	0.88

Table 2 Panel E (Mallows, N=100, T=50)

		Bai-Ng	PC with Mallows			
	r	k	k	MSER	RMSFER	
					rec	roll
DGP 1	4.00	6.18	14.00	0.81	1.03	1.04
(s.e.)	–	0.74	0.00	0.89	1.12	1.07
DGP 2	50.00	24.00	24.00	1.00	1.00	1.00
(s.e.)	–	0.07	0.00	1.00	1.00	1.00
DGP 3	5.00	8.54	15.00	0.84	1.00	1.00
(s.e.)	–	0.85	0.00	0.93	1.01	1.00
DGP 4	5.00	8.61	2.50	1.12	0.85	0.76
(s.e.)	–	0.88	5.60	1.38	1.15	1.15
DGP 5	100.00	17.64	14.45	1.08	1.00	1.00
(s.e.)	–	0.82	11.75	2.16	1.00	1.00
DGP 6	1.00	0.01	5.92	0.77	1.08	1.12
(s.e.)	–	0.10	5.49	1.23	1.24	1.45

Table 2 Panel F (Mallows, N=100, T=50)

		Bai-Ng	Ridge					LF				
	r	k	α	DoF	MSER	RMSFER		α	DoF	MSER	RMSFER	
						rec	roll				rec	roll
DGP 1	4.00	6.18	1.69	13.65	0.80	0.75	0.53	0.00	9.55	0.92	0.43	0.28
(s.e.)	–	0.74	0.43	2.20	1.12	0.80	0.49	0.00	2.58	1.19	0.49	0.29
DGP 2	50.00	24.00	0.54	28.73	0.55	0.01	0.00	0.00	26.85	0.77	0.00	0.00
(s.e.)	–	0.07	0.29	3.95	0.99	0.00	0.00	0.00	7.01	2.15	0.00	0.00
DGP 3	5.00	8.54	2.49	8.24	0.95	0.63	0.42	0.01	5.32	1.06	0.45	0.29
(s.e.)	–	0.85	0.96	2.80	1.16	0.56	0.32	0.00	2.15	1.12	0.43	0.24
DGP 4	5.00	8.61	74.91	2.16	1.14	0.55	0.37	0.00	14.57	0.72	0.31	0.17
(s.e.)	–	0.88	36.61	3.86	1.26	0.50	0.28	0.00	1.47	0.76	0.29	0.14
DGP 5	100.00	17.64	11.00	7.56	1.25	0.00	0.00	0.01	6.79	1.30	0.00	0.00
(s.e.)	–	0.82	5.22	5.37	1.54	0.00	0.00	0.01	5.49	1.61	0.00	0.00
DGP 6	1.00	0.01	8.62	10.58	0.60	0.88	0.78	0.03	10.98	0.58	0.52	0.43
(s.e.)	–	0.10	4.81	4.88	0.99	0.89	0.75	0.01	4.01	0.89	0.68	0.56

Note to Tables 1 to 2. The tables report results for our proposed methods, traditional PC and a regression including all regressors. For each DGP, the first line in the table reports the following: r is the true number of factors; for PC (i.e. Bai-Ng, PC with Cross-Validation, PC with AIC or BIC and PC with Mallows) k is the estimated number of factors on average across Monte Carlo replications, "MSER" is the estimated in-sample Mean Squared Error of each method relative to traditional PC, "RMSFER" is the square root of the estimated Mean Squared Forecast Error of each method relative to traditional PC,

obtained either via rolling ("Rol") or recursive ("Rec") estimation schemes. For Ridge and LF, the tables report α(from eqs. (5) and (9) respectively) and the estimated number of degrees of freedom, "DoF", equal to $tr\left(M_T^{\alpha}\right)$. For Lasso, the tables report the number $k$of non-zero components in the regression coefficient $\widehat{\delta}_L = \arg\min_{\delta}\left\{\frac{1}{2T}\sum_{t=1}^{T}\left(y_t - x_t'\delta\right)^2 + \lambda\sum_{j=1}^{N}\left|\delta_j\right|\right\}$, where δ_j is the j-th component of δ and the value of λ is associated with the minimum MSE for 10-fold cross-validation. The second line reports the standard deviation (across Monte Carlo replications); for example, the standard deviation of the MSER is $\sqrt{var(MSE)}/100$ relative to the respective standard deviation obtained using traditional PC. Table 1 reports results for GCV and Mallows' criterion in large samples (T=500, N=200). Table 2 reports similar results for small samples (T=50, N=100).

Table 3 Panel A. Empirical Analysis (Forecast Horizon = 1 quarter)

	RMSFE (relative to AR)							
I. Cross-validation								
	Factor Models		Ridge	PLS	LF	CV Comb.	BMA	Comb.
	Bai-Ng	Cross V.						
GDP Growth	1.1548	1.0599	0.9663	1.0909	1.0361	0.9711	0.9517	0.9196
	(0.0935)	(0.2521)	(0.2502)	(0.1100)	(0.2148)	(0.3105)	(0.1743)	(0.0025)
Inflation	1.1343	1.0888	1.0447	1.1419	0.9975	0.9919	0.9780	1.0001
	(0.0122)	(0.0474)	(0.4850)	(0.1045)	(0.9432)	(0.8066)	(0.4968)	(0.9909)
II. Mallows								
	Factor Models		Ridge	LF	Mall. Comb.			
	BIC	Mallows						
GDP Growth	0.9795	1.1548	1.0259	1.0361	0.9858			
	(0.6656)	(0.0935)	(0.5388)	(0.2148)	(0.7016)			
Inflation	1.0706	1.1028	1.0320	0.9975	0.9995			
	(0.0407)	(0.0151)	(0.4031)	(0.9432)	(0.9866)			

Table 3 Panel B. Empirical Analysis (Forecast Horizon = 4 quarters)

	RMSFE (relative to AR)							
A. Cross-validation								
	Factor Models		Ridge	PLS	LF	CV Comb.	BMA	Comb.
	Bai and Ng	Cross V.						
GDP Growth	0.9320	0.9033	0.8370	0.9056	1.0295	0.8091	0.8847	0.8402
	(0.4803)	(0.3235)	(0.0017)	(0.3060)	(0.2386)	(0.0003)	(0.0650)	(0.0002)
Inflation	0.9705	0.9794	1.0394	1.4083	1.0231	0.9077	0.9068	0.9487
	(0.7637)	(0.8216)	(0.8085)	(0.2953)	(0.6744)	(0.3177)	(0.0503)	(0.0710)
B. Mallows								
	Factor Models		Ridge	LF	Mall. Comb			
	BIC	Mallows						
GDP Growth	0.8512	0.9320	0.8069	1.0294	0.7854			
	(0.0423)	(0.4803)	(0.0096)	(0.2394)	(0.0006)			
Inflation	0.9724	0.9705	1.0464	1.0231	0.9034			
	(0.7636)	(0.7637)	(0.7726)	(0.6744)	(0.2941)			

Notes to Table 3. The table reports the RMSFE of the model listed in the columns relative to that of the autoregressive model (i.e. RMSFE equals RMSFE of Model / RMSFE of Autoregression). In parentheses we report p-values of the Diebold and Mariano (1995) test statistic for testing the null of equal predictive ability against the alternative of unequal predictive ability using Newey and West's (1987) HAC estimate of the variance with 2 lags. The pool of regressors contains only h-period lagged predictors ($x_{t-h}= Z_{t-h}$).

Table 4 Panel A. Empirical Analysis (Forecast Horizon = 1 quarter)

	RMSFE (relative to AR)							
I. Cross-validation								
	Factor Models		Ridge	PLS	LF	CV Comb	BMA	Comb.
	Bai and Ng	Cross V.						
GDP Growth	1.1088	1.0225	0.9365	1.0884	1.0361	0.9312	0.9517	0.9321
	(0.0886)	(0.7128)	(0.0614)	(0.1769)	(0.2144)	(0.0490)	(0.1743)	(0.0505)
Inflation	1.3894	1.1570	1.2090	1.0924	0.9975	1.0486	0.9780	1.0001
	(0.0018)	(0.0174)	(0.0072)	(0.0616)	(0.9433)	(0.0842)	(0.4968)	(0.9909)
II. Mallows								
	Factor Models		Ridge	LF	Mall. Comb			
	BIC	Mallows						
GDP Growth	0.9781	1.1101	0.9317	1.0361	0.9294			
	(0.6886)	(0.0847)	(0.1071)	(0.2144)	(0.0715)			
Inflation	1.0458	1.2203	1.0549	0.9975	1.0271			
	(0.0640)	(0.0511)	(0.0732)	(0.9433)	(0.3400)			

Table 4 Panel B. Empirical Analysis (Forecast Horizon = 4 quarters)

	RMSFE (relative to AR)							
I. Cross-validation								
	Factor Models		Ridge	PLS	LF	CV Comb	BMA	Comb.
	Bai and Ng	Cross V.						
GDP Growth	0.9090	0.8296	0.8392	1.0144	1.0295	0.7811	0.8847	0.8402
	(0.2637)	(0.0470)	(0.0014)	(0.8659)	(0.2373)	(0.0000)	(0.0650)	(0.0002)
Inflation	1.0818	1.0697	0.9909	1.2307	1.0231	0.9070	0.9068	0.9487
	(0.4684)	(0.6060)	(0.9028)	(0.0006)	(0.6741)	(0.2233)	(0.0503)	(0.0710)
II. Mallows								
	Factor Models		Ridge	LF	Mall. Comb			
	BIC	Mallows						
GDP Growth	0.8728	0.9092	0.7596	1.0295	0.7391			
	(0.1101)	(0.2648)	(0.0022)	(0.2374)	(0.0001)			
Inflation	1.0252	1.0824	1.0388	1.0231	0.9160			
	(0.8384)	(0.4641)	(0.5850)	(0.6741)	(0.2188)			

Notes to Table 4. The table reports the RMSFE of the model listed in the columns relative to that of the autoregressive model (i.e. RMSFE equals RMSFE of Model / RMSFE of Autoregression). In parentheses we report p-values of the Diebold and Mariano (1995) test statistic for testing the null of equal predictive ability against the alternative of unequal predictive ability using Newey and West's (1987) HAC estimate of the variance with 2 lags. The pool of regressors contains two additional lags of the predictors (x_{t-h} includes Z_{t-h}, Z_{t-h-1}, and Z_{t-h-2}).

Table 5. Empirical Analysis (Forecast Horizon = 1 quarter)

Forecast Rationality Regressions

Panel A. GDP Growth				
	Coefficients		Wald Test	
	Constant	Slope	Statistic	P-value
Factor Models:				
Bai and Ng	1.77	-0.67	58.5	0.00
Cross V.	1.36	-0.58	28.4	0.00
Ridge	-0.35	0.01	1.03	0.60
PLS	1.72	-0.64	31.1	0.00
LF	3.85	-1.35	4.51	0.10
Comb. CV	-0.06	-0.11	2.25	0.32
Comb. Mall.	0.87	-0.39	7.89	0.02
BMA	0.34	-0.26	6.00	0.05
Comb.	0.04	-0.07	0.61	0.74

Panel B. Inflation				
	Coefficients		Wald Test	
	Constant	Slope	Statistic	P-value
Factor Models:				
Bai and Ng	-0.03	-0.76	73.33	0.00
Cross V.	-0.03	-0.79	32.77	0.00
Ridge	-0.02	-0.81	25.25	0.00
PLS	-0.03	-0.93	54.33	0.00
LF	-0.02	-3.41	9.01	0.01
Comb. CV	-0.02	-0.59	5.13	0.08
Comb. Mall.	-0.03	-0.63	8.82	0.01
BMA	-0.01	-0.46	3.23	0.20
Comb.	-0.04	-0.58	8.60	0.01

Table 6. Empirical Analysis (Forecast Horizon = 4 quarter)
Forecast Rationality Regressions

Panel A. GDP Growth				
	Coefficients		Wald Test	
	Constant	Slope	Statistic	P-value
Factor Models:				
Bai and Ng	1.14	-0.44	24.74	0.00
Cross V.	0.95	-0.41	21.34	0.00
Ridge	-0.97	0.22	1.99	0.37
PLS	0.93	-0.41	23.44	0.00
LF	2.87	-1.03	3.55	0.17
Comb. CV	-1.18	0.27	3.23	0.20
Comb. Mall.	-0.21	-0.02	1.47	0.48
BMA	-0.08	-0.07	1.31	0.52
Comb.	-0.55	0.14	0.77	0.68

Panel B. Inflation				
	Coefficients		Wald Test	
	Constant	Slope	Statistic	P-value
Factor Models:				
Bai and Ng	-0.11	-0.45	18.08	0.00
Cross V.	-0.08	-0.47	12.71	0.00
Ridge	-0.13	-0.54	6.76	0.03
PLS	-0.11	-0.80	19.47	0.00
LF	-0.07	-1.13	2.08	0.35
Comb. CV	-0.11	-0.17	1.20	0.55
Comb. Mall.	-0.13	-0.18	1.90	0.39
BMA	-0.01	-0.12	0.52	0.77
Comb.	-0.10	-0.31	4.68	0.10

Notes to Tables 5 and 6. The pool of regressors contains only the lagged predictors (x_{t-h}=Z_{t-h}).

Figure 1. Forecasting GDP, h=1. Forecast Performance Relative to the AR Model.

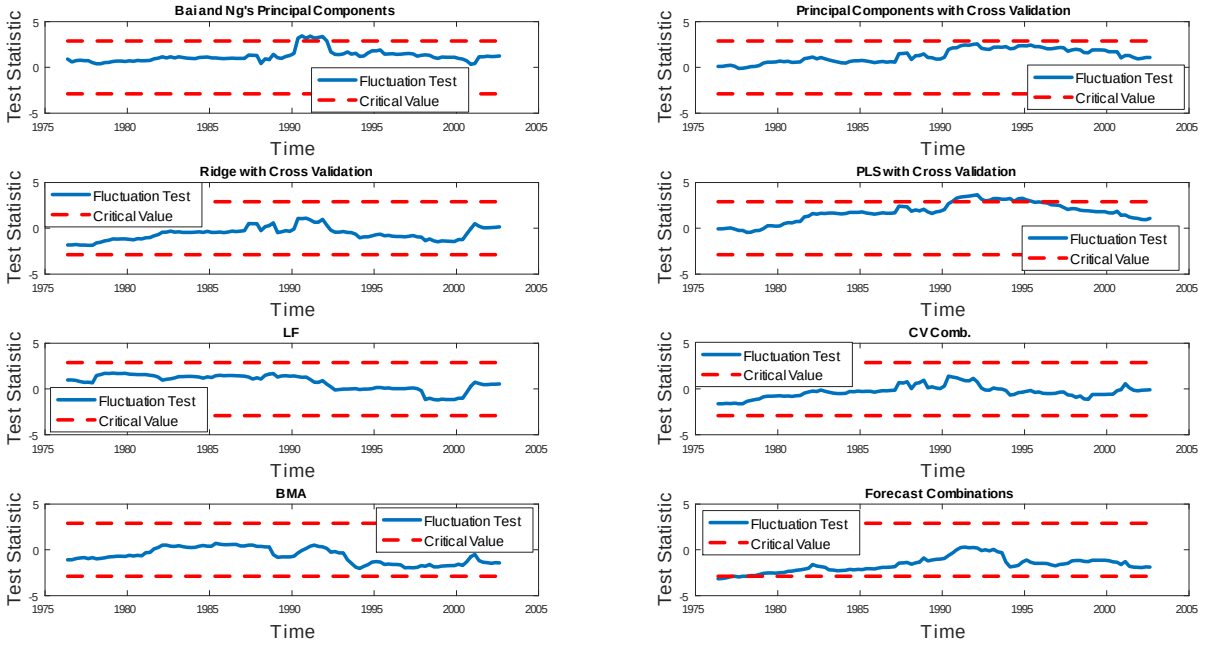


Figure 2. Forecasting Inflation, h=1. Forecast Performance Relative to the AR Model.

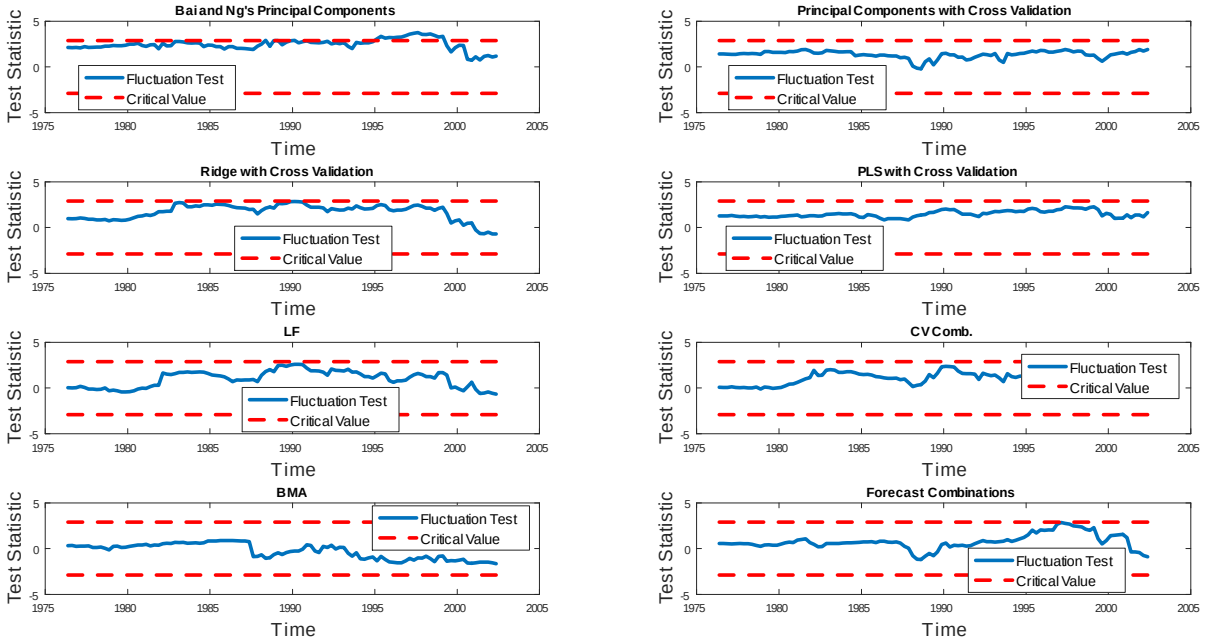


Figure 3. Forecasting GDP, h=4. Forecast Performance Relative to the AR Model.

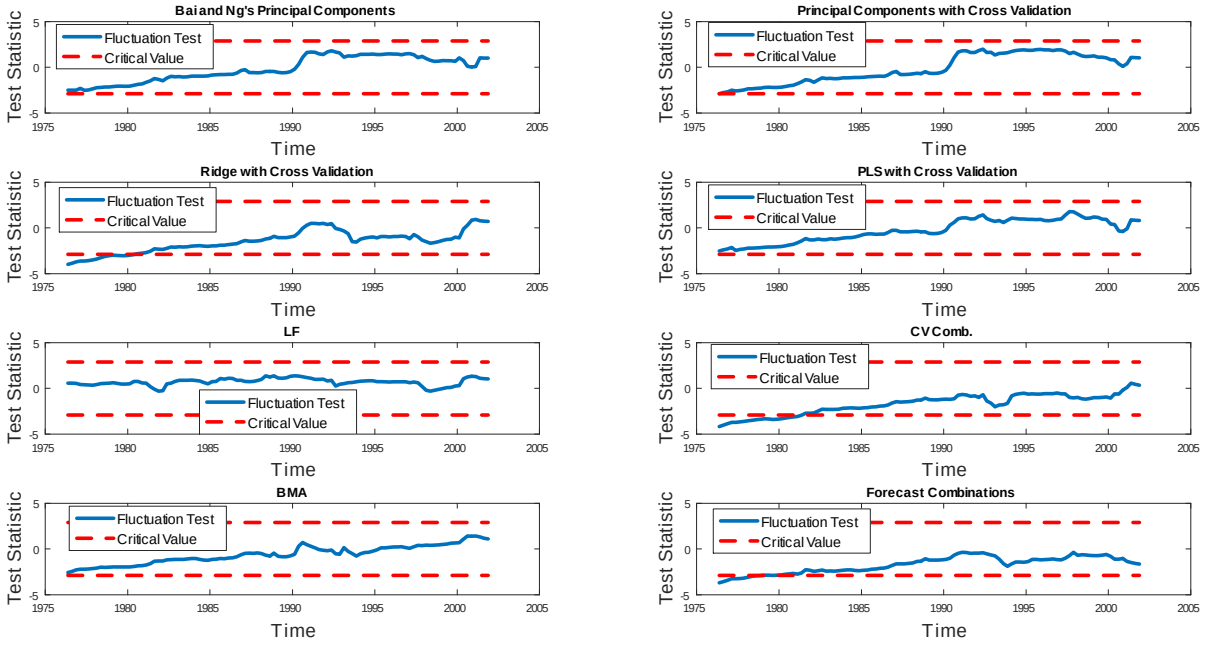


Figure 4. Forecasting Inflation, h=4. Forecast Performance Relative to the AR Model.

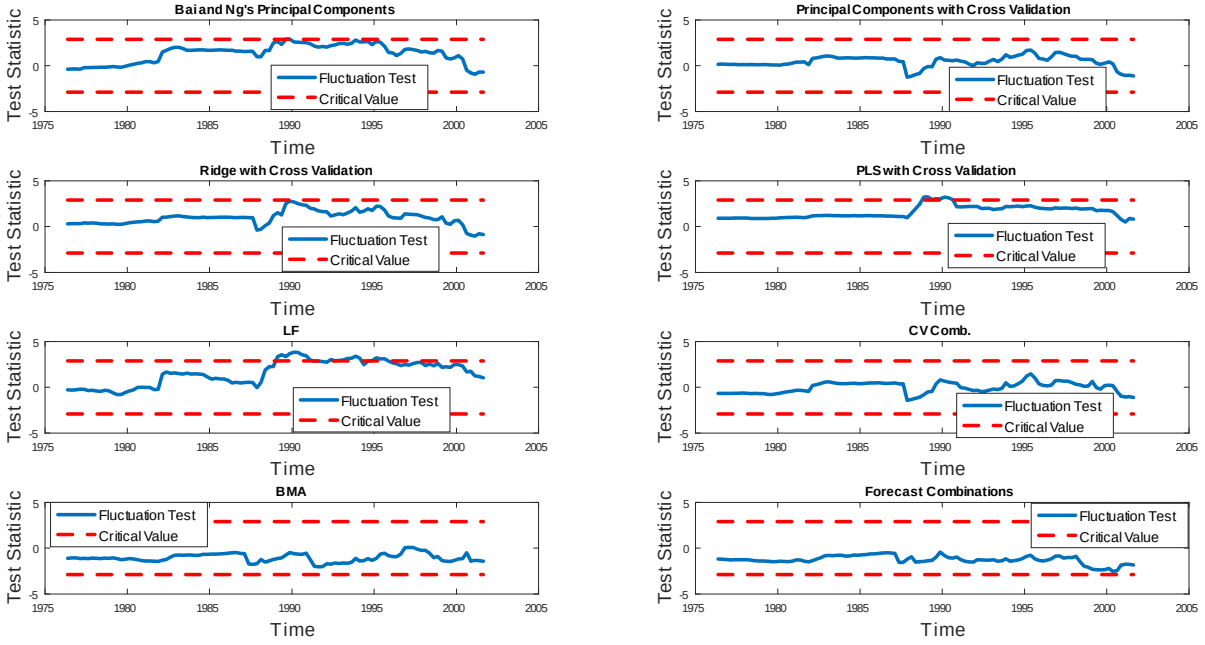


Figure 5. Forecasting GDP, h=1. Forecast Rationality.

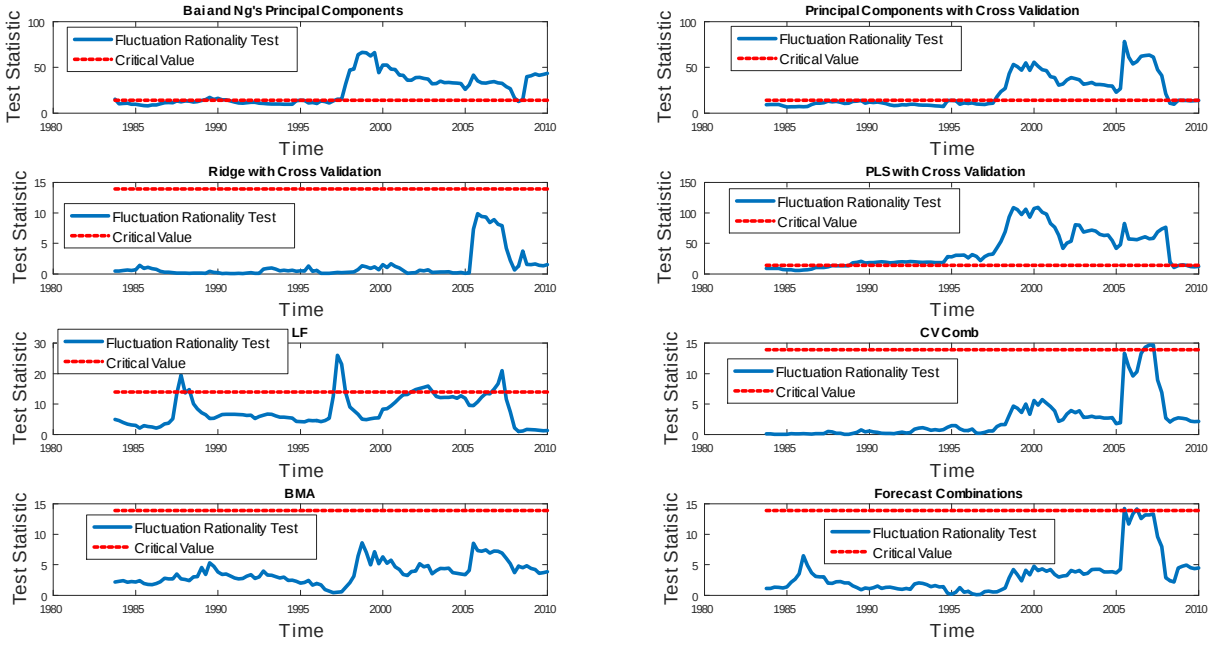


Figure 6. Forecasting Inflation, h=1. Forecast Rationality.

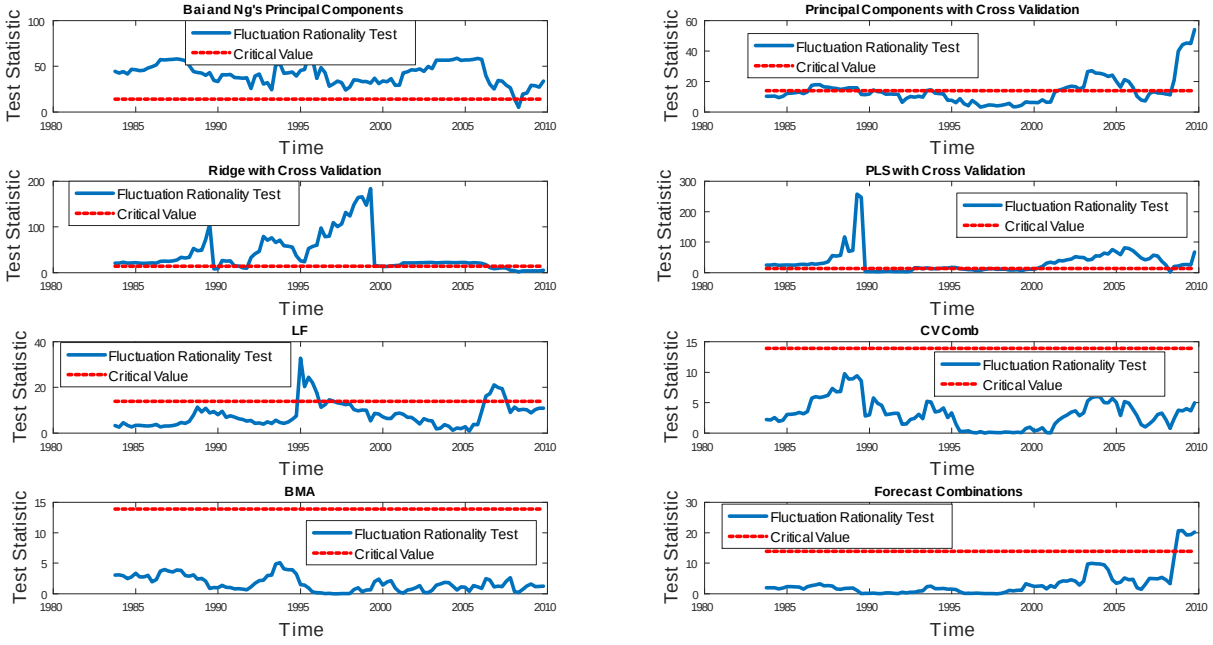


Figure 7. Forecasting GDP, h=4. Forecast Rationality.

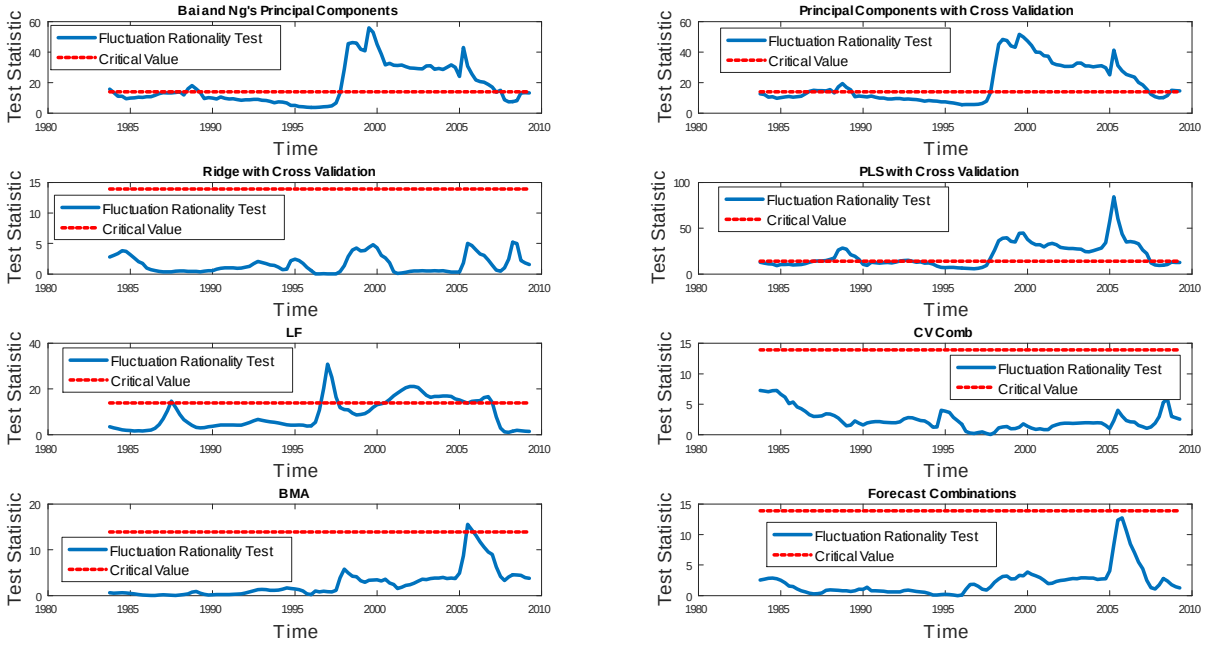
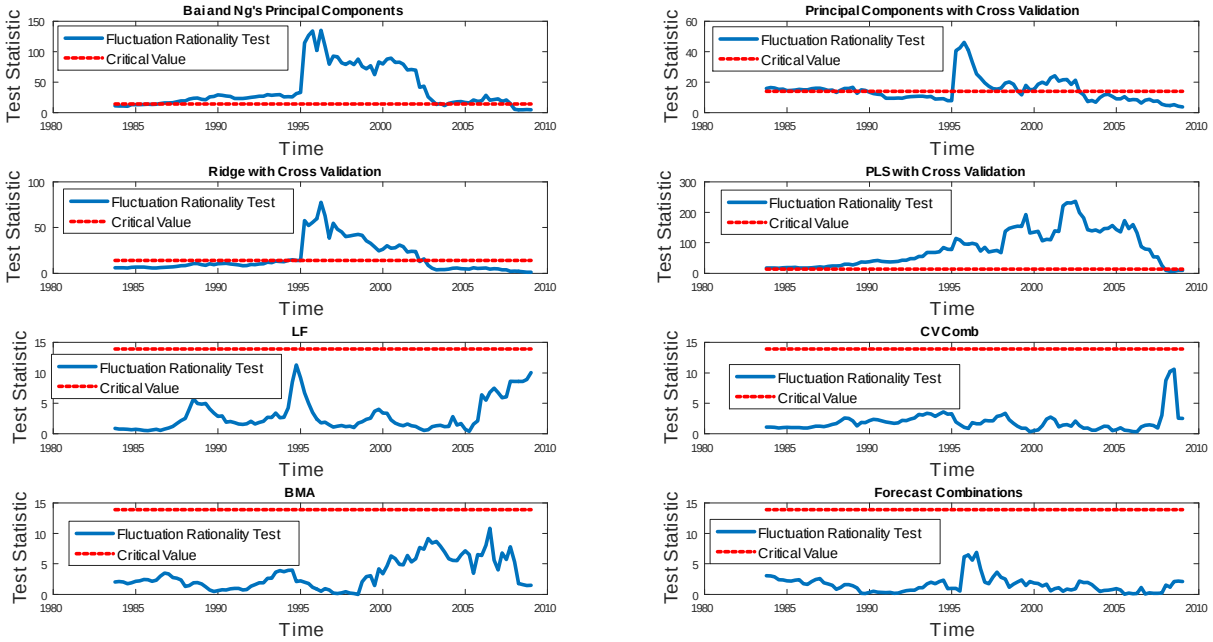


Figure 8. Forecasting Inflation, h=4. Forecast Rationality



Notes to Figures 1-4. The figures report Giacomini and Rossi's (2010) Fluctuation test (solid line) and critical values (dotted lines) for the forecasting models listed in the title. Figures 1-2 focus on forecasting

output growth and inflation in the short-run, while figures 3-4 focus on the long-run. The pool of regressors contains one lag of the predictors (x_{t-h} includes Z_{t-h}).

Notes to Figures 5-8. The figures report Rossi and Sekhposyan's (2015) Fluctuation Rationality test (solid line) and critical values (dotted lines) for the forecasting models listed in the title. Figures 5-6 focus on forecasting output growth and inflation in the short-run, while figures 7-8 focus on the long-run. The pool of regressors contains one lag of the predictors (x_{t-h} includes Z_{t-h}).