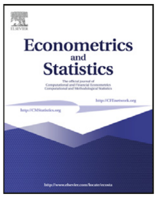
Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta



The Asymptotic Equivalence of Ridge and Principal Component Regression with Many Predictors

Christine De Mol[a,*], Domenico Giannone[b,d,e], Lucrezia Reichlin[c,e]

[a] *Department of Mathematics and ECARES, Université libre de Bruxelles, Campus Plaine CPI 217, Boulevard du Triomphe, Brussels, 1050, Belgium*
[b] *Department of Economics, University of Washington, Savery Hall Room 305, Seattle, 98105, WA, USA*
[c] *London Business School, The Regent's Park, London, NW1 4SA, United Kingdom*
[d] *International Monetary Fund, 700 19th St NW, Washington, 20431, DC, USA*
[e] *CEPR, 33 Great Sutton Street, London, EC1V 0DX, United Kingdom*

ARTICLE INFO

Article history:
Received 29 June 2023
Revised 26 February 2024
Accepted 26 February 2024
Available online xxx

Keywords:
High-dimensional time series
Factor models
Forecasting
Ridge regression
Asymptotic inference

ABSTRACT

The asymptotic properties of ridge regression in large dimension are studied. Two key results are established. First, consistency and rates of convergence for ridge regression are obtained under assumptions which impose different rates of increase in the dimension n between the first n_1 and the remaining $n - n_1$ eigenvalues of the population covariance of the predictors. Second, it is proved that under the special and more restrictive case of an approximate factor structure, principal component and ridge regression have the same rate of convergence and the rate is faster than the one previously established for ridge.

1. Introduction

This paper develops asymptotic analysis of ridge regression in large dimension and compares it with principal component (PC) regression. Ridge regression estimates coefficients by minimizing the residual sum of squares plus a penalty which has the effect of shrinking the estimates of the coefficients towards zero. Principal component regression, on the other hand, uses PC as predictors in a linear regression that is fit using least squares.

Ridge regression has a long tradition in applied mathematics (Tikhonov, 1963) and also in econometrics and statistics (Hoerl and Kennard, 1970; Leamer, 1973). The role of the penalty in the least squares regression is to reduce the variance of the regression coefficients at the cost of a bias. The L2-penalty 'shrinks' all coefficients towards zero. A shrinkage parameter governs the tradeoff between variance and bias. As all regressors remain relevant, ridge captures dense structures of the data. It differs from shrinkage methods such as LASSO in which the L1-penalty performs variable selection, thereby capturing sparse structures. When data are correlated, ridge performs better than sparsity-enforcing methods (for an intuitive explanation see James et al. (2013)). In high dimension, ridge was first studied by De Mol et al. (2008) who compare it with principal component (PC) regression. Recent related work is He (2023).

PC Regression is an alternative to shrinkage as it compresses the data into a few components which capture the bulk of the covariance of the data. As it is the case for shrinkage methods, compression provides a solution to the curse of

[*] Corresponding author.
E-mail address: christine.de.mol@ulb.be (C. De Mol).

https://doi.org/10.1016/j.ecosta.2024.02.005
2452-3062/© 2024 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

dimensionality problem. Seminal work by Forni et al. (2000) and Stock and Watson (2002) has established (n, T) consistency properties of principal component regression when the data have an approximate factor structure. Rates of convergence and general asymptotic properties, when both the number of predictors n and the sample size T go to infinity, have been studied by Bai and Ng (2002); Bai (2003) and Forni et al. (2009). Doz et al. (2012) and Barigozzi and Luciani (2019) establish consistency and rates for likelihood-based estimators.

Standard assumptions in this literature require that the ratio between the r-th largest and the $(r+1)$-th largest eigenvalues of the population covariance matrix of the data, where r is the number of factors, is rising proportionally to n so that the cumulative effects of the normalized factors on the cross-sectional units strongly dominate the idiosyncratic influences asymptotically.

De Mol et al. (2008) show that under similar assumptions, ridge yields (n, T) consistent forecast, as had been established for PC. The intuitive explanation of the result is that both estimators give high weight to the dominant principal components and either little weight (ridge) or zero weight (PC) to the remaining ones and therefore work well in forecasting when few principal components explain the bulk of the predictors' variation, as is the case with data that comove strongly.

Building on this analysis, we consider here more general assumptions on the data generating process, allowing the eigenvalues of the population covariance matrix to diverge at different rates. The first n_1 eigenvalues grow with n but not necessarily all linearly. The $n - n_1$ remaining eigenvalues increase with n at a slower rate, so that the gap between the first cluster of n_1 eigenvalues and the second cluster of the remaining $n - n_1$ eigenvalues also increases with n, but possibly at a sublinear (i.e. slower than linear) rate. Accordingly, the population prediction equation is split into two components corresponding to these two clusters. This means that we relax the assumptions on the approximate factor model in two directions. First, the eigenvalues of the covariance matrix of the idiosyncratic component can grow with the cross-sectional dimension, allowing for the possibility of pervasive idiosyncratic shocks (note that this case was already considered by De Mol et al. (2008)). Second, some of the eigenvalues of the dominating cluster can grow with the cross-sectional dimension at a slower-than-linear rate, allowing for weaker factors. Our model includes as a special case the strong-factor-structure traditional assumptions of the principal component literature, as well as the structure adopted by Bai and Ng (2023) under the name 'weaker loadings'. This setup is different from the weak factor models considered by Onatski (2012) but broadly aims at capturing similar empirical situations.

We prove consistency and rates of convergence of the ridge regression estimate to the component of the prediction driven by the dominating cluster, when the number of predictors n and the sample size T go to infinity. The bias between the forecast driven by this component and the one provided by the ridge estimate vanishes asymptotically provided that the ridge parameter is properly tuned as a function of n and T. We also compare these rates of convergence for ridge with those for PC under the same relaxed assumptions.

We then reconsider as a special case the factor structure analyzed in De Mol et al. (2008) and show that in this case the component recovered by ridge regression corresponds to the prediction based on the $r = n_1$ pervasive factors. Moreover, we can asymptotically capture not only the forecast driven by the dominating subspace or common component, but also the optimal forecast. Here the asymptotic analysis lets the number of predictors and the sample size go to infinity with no restriction on their relative growth rates. The rates we obtain are improved with respect to those derived by De Mol et al. (2008) and are the same as for PC.

Our results have two important implications. First, they establish that ridge is a valid alternative to PC and therefore provides asymptotic foundations for the use of L2-penalized regression in large dimension, including Bayesian regression with normal priors. The multivariate generalization of this approach is the Bayesian VAR with Minnesota priors which is a well established tool in time series econometrics. For an analysis of the Bayesian VAR in large dimension, see Bańbura et al. (2010) and the subsequent literature recently surveyed by Hauzenberger et al. (2024). Second, they show that compression and shrinkage are equivalent. Ridge via shrinkage and PC via compression both capture the component of the prediction that is associated with the dominating eigenvalues of the covariance matrix. Both ridge and PC regression are dense in the sense of Giannone et al. (2021) since the mass of regression coefficients is dispersed throughout all variables implying that all explanatory variables are included in the prediction, although the impact of each of them may be small.

Notations

Throughout the paper, we will use the following notations. For a vector v in $\mathbb{R}^n$, we will denote its L2-norm by $\|v\|$, that is $\|v\| = \sqrt{\sum_{i=1}^{n} |v_i|^2}$. For a matrix A, we will use the spectral norm defined as $\|A\| = \max_{v;\, v'v=1} \sqrt{v'A'Av}$ (which is the maximal eigenvalue for a symmetric square matrix), where v' denotes the transpose of v and A' the transpose of A. Identity matrices will be denoted by I.

As concerns asymptotics, we will use the notations 'Big O' and 'Big Theta'. We recall that for two functions $f(n)$ and $g(n)$ depending on n, one says that $f(n) = O(g(n))$ asymptotically as $n \to \infty$ if $|f(n)| \le Mg(n)$, for all $n > n_0$, with $M > 0$ a constant independent of n. If, moreover, there is another constant $m > 0$ such that $mg(n) \le |f(n)| \le Mg(n)$, for all $n > n_0$, then one says that $f(n) = \Theta(g(n))$ asymptotically.

For stochastic variables, similar bounds are supposed to hold in probability, i.e. one says that $f(n) = O_p(g(n))$ if $\left|\frac{f(n)}{g(n)}\right|$ is bounded in probability, or else if for every $\eta > 0$ there is a constant $M(\eta)$ and an integer $n(\eta)$ such that if $n \ge n(\eta)$, then the probability $P\left(\left|\frac{f(n)}{g(n)}\right| \le M(\eta)\right)$ is greater or equal to $1 - \eta$.

Later on, we will also consider asymptotics in both variables n and T, the number of observed time samples. Besides, we will use repeatedly the following well-known result.

Lemma 1. *A zero-mean stochastic variable X_n is of the order of its standard deviation σ_n, i.e. $X_n = O_p(\sigma_n)$.*

Proof. Let X_n be a sequence of zero-mean stochastic variables with variance $Var(X_n) \equiv \sigma_n^2$. Then the Tchebycheff inequality implies that

$$P(|X_n| \geq \varepsilon) \leq \frac{Var(X_n)}{\varepsilon^2}, \quad \text{for any} \quad \varepsilon > 0, \tag{1}$$

or equivalently, setting $\eta = \varepsilon^{-2}$,

$$P\left(\left|\frac{X_n}{\sigma_n}\right| \leq \frac{1}{\sqrt{\eta}}\right) \geq 1 - \eta, \tag{2}$$

which means precisely that $X_n = O_p(\sigma_n)$. □

2. General Setting

Consider the linear regression model

$$y_{t+h} = X_t' \ \beta + u_{t+h} \tag{3}$$

where y_t is the one-dimensional target variable to be forecast at some horizon h, X_t is a high-dimensional time series of dimension n and u_t is a noise term. To allow for dependence of the forecast on a finite number of lags in the series, we assume that the corresponding lagged series are included in X_t.

Assumption 1. In model (3), i.e. $y_{t+h} = X_t' \ \beta + u_{t+h}$, we assume that

(i) the individual series $x_{it}, i = 1, \ldots, n,$ are normalized to have zero mean and unit variance;
(ii) u_t has mean zero and is orthogonal to each of the n individual series $x_{it}, i = 1, \ldots, n$, namely $E(x_{it}u_{t+h}) = 0$, for all i (by E we denote the expectation);
(iii) y_t and X_t are jointly stationary.

Then, the n-dimensional vector of the population regression coefficients, denoted by β, is given by

$$\beta = \Sigma_{XX}^{-1} \ \Sigma_{Xy} \tag{4}$$

where $\Sigma_{XX} = E(X_tX_t')$ is the $n \times n$ population covariance matrix, assumed to be invertible, and $\Sigma_{Xy} = E(X_ty_{t+h})$ is the $n \times 1$ population covariance with the dependent variable. Stationarity, implicitly assuming the existence of second-order moments, means that these covariance matrices do not depend on t and are componentwise bounded. Moreover, for a fixed n, all variances are bounded, i.e. $Var(y_t) < +\infty$, $Var(x_{it}) < +\infty$ for all i and $Var(u_t) < +\infty$. Notice that the covariance matrix Σ_{Xy}, and hence the regression coefficient β, depend on the forecast horizon h. We drop this dependence for notational convenience.

The aim is to forecast y_{t+h} based on the information contained in the observations of the high-dimensional time series X_t. Moreover, we want to investigate the following questions: (i) to what extent can we improve the forecast by increasing the number n of individual series? (ii) what can be recovered asymptotically when $n \to \infty$?

Let us first remark that

$$Var(y_{t+h}) = Var(X_t'\beta) + Var(u_{t+h}). \tag{5}$$

Since we want to investigate the asymptotic behavior of the forecast when the dimension n tends to infinity, we have to ensure that the variance of the forecast does not blow up and remains bounded not only for every fixed n but also for $n \to \infty$. Accordingly, we introduce the following assumption.

Assumption 2.

$$Var(X_t'\beta) = \beta'\Sigma_{XX}\beta = O(1) \quad \text{for} \quad n \to \infty. \tag{6}$$

Let us now denote by y the $T \times 1$ vector collecting the observations of the target variable $y_{t'}$ for $t' = 1+h, 2+h, \cdots, T' \equiv T+h$ available at time T' and range the corresponding observations of X_t for $t = 1, \ldots, T$ in the $T \times n$ matrix X.

To overcome the curse of dimensionality affecting the Ordinary Least Squares (OLS) estimator $\hat{\beta} = (X'X)^{-1}X'y$, we consider the following penalized least squares or ridge estimator

$$\hat{\beta}_\lambda = \arg\min_\beta \left\{ \frac{1}{T}\|y - X\beta\|^2 + \lambda\|\beta\|^2 \right\} \tag{7}$$

where λ is the shrinkage parameter and where the penalty involves the L2-norm of β. This ridge estimator is given by

$$\hat{\beta}_\lambda = \left(\frac{X'X}{T} + \lambda I\right)^{-1} \frac{X'y}{T} = \left(S_{XX} + \lambda I\right)^{-1} S_{Xy} \tag{8}$$

where $S_{XX} = \frac{X'X}{T}$ and $S_{Xy} = \frac{X'y}{T}$ are the sample covariance matrices and I is the identity matrix.

Another difficulty in high dimension is that the sample covariance matrices (the only ones we have access to from the data) are not necessarily close to the population covariance matrices (see Lemma 4 hereafter). This induces us to introduce the following assumption – in the spirit of factor models – that the eigenvalue spectrum of Σ_{XX} can be separated into two clusters, labelled by 1 and 2. The first one, of fixed dimension (independent of n), contains the largest eigenvalues, typically growing fast with n for correlated individual series, and on which the forecast will be based. The other one contains the smallest eigenvalues which remain bounded or grow with n at a slower rate. This is the part that we will not be able to capture asymptotically when $n \to \infty$, so that we consider it as irrelevant for the forecast. Precise assumptions on the behavior of the eigenvalues in the two clusters will be made later in the paper.

Assumption 3. The population covariance matrix admits the following eigendecomposition

$$\Sigma_{XX} = V_1 \; D_1 \; V_1' + V_2 \; D_2 \; V_2' \tag{9}$$

where D_i is a diagonal matrix containing the eigenvalues d_j's of Σ_{XX} belonging to the cluster i and V_i is the matrix containing as columns the corresponding eigenvectors. The eigenvalues are ordered by decreasing order of magnitude – and repeated according to their multiplicity in case of degeneracy (with, of course, no eigenvalue belonging to the two clusters). The first cluster is of fixed dimension n_1, independent of n, and contains the largest n_1 eigenvalues, whereas the second one has dimension $n_2 = n - n_1$ and contains the remaining ones. The eigenvectors are orthonormalized, so that we have $V_1 \; V_1' + V_2 \; V_2' = I_{n\times n}$, where $I_{n\times n}$ is the $n \times n$ identity matrix, and also $V_1' \; V_2 = 0 = V_2' \; V_1$, expressing the orthogonality of the eigensubspaces corresponding to the two clusters.

We will also need the eigendecompositions

$$\Sigma_{XX}^{1/2} = V_1 \; D_1^{1/2} \; V_1' + V_2 \; D_2^{1/2} \; V_2' \tag{10}$$

and, if all eigenvalues are bounded away from zero,

$$\Sigma_{XX}^{-1} = V_1 \; D_1^{-1} \; V_1' + V_2 \; D_2^{-1} \; V_2'. \tag{11}$$

For the empirical covariance $S_{XX} = \frac{X'X}{T}$, we will make use of the analogous spectral decomposition

$$S_{XX} = \hat{V}_1 \; \hat{D}_1 \; \hat{V}_1' + \hat{V}_2 \; \hat{D}_2 \; \hat{V}_2' \tag{12}$$

where the first cluster will contain the n_1 largest eigenvalues $\hat{d}_j$'s of S_{XX}, using similar assumptions as in Assumption 3. We will also need later the eigendecomposition

$$(S_{XX} + \lambda I)^{-1} = \hat{V}_1 \; (\hat{D}_1 + \lambda I)^{-1} \; \hat{V}_1' + \hat{V}_2 \; (\hat{D}_2 + \lambda I)^{-1} \; \hat{V}_2' \tag{13}$$

where I denote the identity matrices of appropriate dimension (n_1 or n_2).

Now let us decompose β as

$$\beta = V_1 \; V_1' \; \beta + V_2 \; V_2' \; \beta \equiv \beta_1 + \beta_2 \tag{14}$$

where β_1 (respectively β_2) is the projection of β on the first (respectively second) cluster. Then the variance $\beta' \Sigma_{XX} \beta$ of $X_t' \beta$ can also be decomposed into the two parts:

$$\begin{aligned} \beta' \Sigma_{XX} \beta &= \| \Sigma_{XX}^{1/2} \beta \|^2 = \beta_1' V_1 \; D_1 \; V_1' \beta_1 + \beta_2' V_2 \; D_2 \; V_2' \beta_2 \\ &= \| V_1 \; D_1^{1/2} V_1' \beta_1 \|^2 + \| V_2 \; D_2^{1/2} V_2' \beta_2 \|^2 = O(1), \end{aligned} \tag{15}$$

the last equality resulting from Assumption 2.

Lemma 2. *Under Assumption 2, we have the following bound on the norm of β_1 for $n \to \infty$*

$$\| \beta_1 \| = O\left(\frac{1}{\sqrt{d_1^{\min}}}\right) \tag{16}$$

where $d_1^{\min}$ denotes the smallest eigenvalue of the first cluster.

Proof. We have

$$\begin{aligned} \| \beta_1 \| &= \| V_1 \; D_1^{-1/2} \; V_1' \; V_1 \; D_1^{1/2} \; V_1' \beta_1 \| \\ &\le \| V_1 \; D_1^{-1/2} \; V_1' \| \; \| V_1 \; D_1^{1/2} \; V_1' \beta_1 \| \le \frac{1}{\sqrt{d_1^{\min}}} \; O(1) \end{aligned} \tag{17}$$

since $\| V_1 \; D_1^{1/2} \; V_1' \beta_1 \|$ is also an $O(1)$ by (15). □

3. Perturbation results for the covariance matrices

We will first establish some bounds on the difference – measured in the spectral norm – between the population and sample covariances. The basic assumption is that the sample covariances converge componentwise to their population counterparts. More specifically, we will introduce the following assumption.

Assumption 4. There exist constants $M_1 < \infty$ and $M_2 < \infty$ such that

$$T\, E\Big([S_{XX} - \Sigma_{XX}]_{i,j}^2 \Big) < M_1, \quad \text{for all} \quad i, j, \tag{18}$$

and

$$T\, E\Big(\big[S_{Xy} - \Sigma_{Xy}\big]_i^2 \Big) < M_2, \quad \text{for all} \quad i. \tag{19}$$

The bounds M_1 and M_2 are uniform with respect to the elements i, j and i, respectively.

Notice that the componentwise convergence can be obtained under mild conditions on the autocovariances and fourth cumulants of the predictors x_i and on the cross-covariances with the target variable y. We do not state primitive assumptions since the desired result can be obtained with a variety of detailed conditions. For a discussion see Forni et al. (2009) and Barigozzi (2022).

Lemma 3. *Under Assumption 4, we have that, asymptotically for large n and T,*

$$\| \Sigma_{XX} - S_{XX} \| = O_p\left(\frac{n}{\sqrt{T}} \right), \tag{20}$$

and

$$\| \Sigma_{Xy} - S_{Xy} \| = O_p\left(\frac{\sqrt{n}}{\sqrt{T}} \right). \tag{21}$$

Proof. Since $\|A\|^2 \leq Trace(A'A) = \sum_{i,j} A_{i,j}^2$, we have that

$$E\Big(\| \Sigma_{XX} - S_{XX} \|^2 \Big) \leq n^2 \max_{i,j} E[S_{XX} - \Sigma_{XX}]_{i,j}^2 < M_1 \, \frac{n^2}{T}, \tag{22}$$

and

$$E\Big(\| \Sigma_{Xy} - S_{Xy} \|^2 \Big) \leq n \max_i E\big[S_{Xy} - \Sigma_{Xy}\big]_i^2 < M_2 \, \frac{n}{T}. \tag{23}$$

The result then follows from Lemma 1. □

Weyl's inequality then yields the following classical perturbation result for the eigenvalues.

Lemma 4. *We have for $n, T \to \infty$ and for all indices j,*

$$\hat{d}_j = d_j + O_p\left(\frac{n}{\sqrt{T}} \right) \tag{24}$$

where d_j (respectively $\hat{d}_j$) is the j^{th} eigenvalue of Σ_{XX} (respectively S_{XX}).

Proof. The proof immediately follows from Weyl's inequality

$$|d_j - \hat{d}_j| \leq \| \Sigma_{XX} - S_{XX} \| = O_p\left(\frac{n}{\sqrt{T}} \right) \tag{25}$$

which holds for all j. □

This shows that in high dimension the two spectra can be quite far apart.

Even more crucial are results about the perturbation of the corresponding eigenvectors. Using the decompositions above into two clusters of eigenvalues, we want to establish asymptotic bounds on the cross terms $\|V_1' \, \hat{V}_2\|$ and $\|V_2' \, \hat{V}_1\|$. These terms characterize in some way the angles between the subspaces spanned by the respective eigenvectors of the two clusters.

We can derive the following result.

Lemma 5. *Asymptotically for large n and T, we have*

$$\|V_1' \, \hat{V}_2\| \leq \frac{1}{d_1^{\min} - d_2^{\max} + O_p\left(\frac{n}{\sqrt{T}} \right)} \, O_p\left(\frac{n}{\sqrt{T}} \right) \tag{26}$$

and the same bound holds for $\|V_2' \hat{V}_1\|$.

Proof. We have $S_{XX} \hat{V}_2 = (\hat{V}_1 \hat{D}_1 \hat{V}_1' + \hat{V}_2 \hat{D}_2 \hat{V}_2') \hat{V}_2 = \hat{V}_2 \hat{D}_2$ since $\hat{V}_1' \hat{V}_2 = 0$. Similarly, $V_1' \Sigma_{XX} = D_1 V_1'$. Hence

$$V_1' S_{XX} \hat{V}_2 = V_1' \hat{V}_2 \hat{D}_2 \tag{27}$$

and

$$D_1 V_1' \hat{V}_2 = V_1' \Sigma_{XX} \hat{V}_2 = V_1' (\Sigma_{XX} - S_{XX}) \hat{V}_2 + V_1' \hat{V}_2 \hat{D}_2. \tag{28}$$

This yields the bound

$$\begin{aligned} \|V_1' \hat{V}_2\| &\le \|D_1^{-1}\| \Big(\|\Sigma_{XX} - S_{XX}\| + \|\hat{D}_2\| \ \|V_1' \hat{V}_2\| \Big) \\ &= \frac{1}{d_1^{\min}} \Big(\|\Sigma_{XX} - S_{XX}\| + \hat{d}_2^{\max} \ \|V_1' \hat{V}_2\| \Big) \end{aligned} \tag{29}$$

where $d_1^{\min}$ denotes the smallest eigenvalue of D_1 and $\hat{d}_2^{\max}$ the largest of $\hat{D}_2$, and we used the fact that all matrices formed by orthogonal vectors have a spectral norm bounded by 1. Using Lemma 4, we get

$$\|V_1' \hat{V}_2\| \le \frac{1}{d_1^{\min}} \left(\|\Sigma_{XX} - S_{XX}\| + \left(d_2^{\max} + O_p\left(\frac{n}{\sqrt{T}} \right) \right) \|V_1' \hat{V}_2\| \right), \tag{30}$$

whence the bound (26) using Lemma 3. We can derive exactly the same bound for $\|V_2' \hat{V}_1\|$. □

Let us remark that (26) is non-trivial and yields a meaningful bound only if the gap between the eigenvalues of the two clusters of the population covariance matrix is strictly positive and, moreover, if this gap dominates the size of the perturbation $O_p(\frac{n}{\sqrt{T}})$ resulting from Lemma 4, namely if asymptotically

$$gap(n) \equiv d_1^{\min} - d_2^{\max} > O_p\left(\frac{n}{\sqrt{T}} \right). \tag{31}$$

Notice that if $gap(n)$ grows proportionally to n, then it is sufficient to assume that T is large enough, independently of n. However, if the gap grows with n at a slower rate, then we have to assume that T grows sufficiently fast with respect to n.

Lemma 5 is somehow related to the Davis-Kahan theorem ((Davis and Kahan, 1970) – see also Yu et al. (2014)) but the comparison with these results is not completely straightforward because of different contexts. We provide here a simple proof of the result under the form we need for our purpose.

As we will see, the perturbation results of the present section are the key for deriving the asymptotic consistency results of the next section.

4. Consistency rates for large *n* and *T*

In this section, we will derive the key result of the paper, namely that, under appropriate spectral assumptions on the population covariance matrix, the ridge regression estimator (8) yields asymptotically for large n and T the same forecast as the forecast driven by the dominant spectral cluster, namely that $X_t' \hat{\beta}_\lambda$ converges in probability to $X_t' \beta_1$, provided that the parameter λ is tuned appropriately ($\lambda \sim n/\sqrt{T}$).

We first establish the following lemma.

Lemma 6. *Asymptotically for large n and T, we have the bound*

$$\|X_t' \hat{V}_2\| \le O_p\left(\sqrt{d_1^{\max}} \right) \frac{1}{d_1^{\min} - d_2^{\max} + O_p(\frac{n}{\sqrt{T}})} O_p\left(\frac{n}{\sqrt{T}} \right) + O_p\left(\sqrt{d_2^{\max}} \right). \tag{32}$$

Proof. Let us remark that

$$\begin{aligned} \|X_t' \hat{V}_2\| &= \|X_t' V_1 V_1' \hat{V}_2 + X_t' V_2 V_2' \hat{V}_2\| \\ &\le \|X_t' V_1\| \ \|V_1' \hat{V}_2\| + \|X_t' V_2\| \ \|V_2' \hat{V}_2\|. \end{aligned} \tag{33}$$

We have $\|X_t' V_1\| \le O_p(\sqrt{d_1^{\max}})$ and $\|X_t' V_2\| = O_p(\sqrt{d_2^{\max}})$ since

$$E\|X_t' V_1\|^2 = E\|V_1' X_t X_t' V_1\| = \|V_1' \Sigma_{XX} V_1\| \le d_1^{\max} \tag{34}$$

and the same holds when substituting the indices 1 by 2. Using Lemmas 1 and 5 as well as the bound $\|V_2' \hat{V}_2\| \le \|V_2'\| \ \|\hat{V}_2\| \le 1$ for the orthogonal matrices $\|\hat{V}_2\|$ and $\|V_2'\|$, the result follows. □

The next proposition yields an asymptotic upper bound for the error we commit when forecasting $X_t' \beta_1$ by $X_t' \hat{\beta}_\lambda$, or else for the bias between the forecast driven by the dominating cluster and the one provided by the ridge regression estimator.

Proposition 1. *Asymptotically for large n and T, the point forecast error is bounded as follows*

$$\begin{aligned}|X_t'(\hat{\beta}_\lambda - \beta_1)| \leq_p\ & \frac{n}{\sqrt{T}\left(d_1^{\min} + O_p\left(\frac{n}{\sqrt{T}}\right)\right)}\left(1 + \frac{\sqrt{n}}{\sqrt{d_1^{\min}}}\right)\\ &+ \frac{1}{\sqrt{d_1^{\min}}}\left(\frac{n\sqrt{d_1^{\max}}}{\sqrt{T}(d_1^{\min} - d_2^{\max} + O_p\left(\frac{n}{\sqrt{T}}\right))} + \sqrt{d_2^{\max}}\right)\\ &+ \frac{1}{\lambda}\,\frac{\sqrt{n}}{\sqrt{T}}\left(1 + \frac{\sqrt{n}}{\sqrt{d_1^{\min}}}\right)\left(\frac{n\sqrt{d_1^{\max}}}{\sqrt{T}(d_1^{\min} - d_2^{\max} + O_p\left(\frac{n}{\sqrt{T}}\right))} + \sqrt{d_2^{\max}}\right)\\ &+ \lambda\,\frac{1}{\sqrt{d_1^{\min}}}\,\frac{\sqrt{n}}{d_1^{\min} + O_p\left(\frac{n}{\sqrt{T}}\right)},\end{aligned} \tag{35}$$

where $\leq_p \cdot$ means $\leq O_p(\cdot)$.

Proof. Let us consider the difference

$$\begin{aligned}\hat{\beta}_\lambda - \beta_1 &= \left(S_{XX} + \lambda I\right)^{-1} S_{Xy} - V_1\ V_1'\ \Sigma_{XX}^{-1}\ \Sigma_{Xy}\\ &= \left(S_{XX} + \lambda I\right)^{-1}(S_{Xy} - \Sigma_{Xy})\\ &\quad + \left[\left(S_{XX} + \lambda I\right)^{-1} - V_1\ V_1'\ \Sigma_{XX}^{-1}\right]\Sigma_{Xy}.\end{aligned} \tag{36}$$

Noticing that $V_1\ V_1'\ \Sigma_{XX}^{-1} = V_1\ V_1'\ V_1\ D_1^{-1}\ V_1' = (V_1\ D_1\ V_1')^{-1}$ and using the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we get

$$\begin{aligned}\hat{\beta}_\lambda - \beta_1 &= \left(S_{XX} + \lambda I\right)^{-1}(S_{Xy} - \Sigma_{Xy})\\ &\quad + \left(S_{XX} + \lambda I\right)^{-1}(V_1\ D_1\ V_1' - S_{XX} - \lambda I)(V_1\ D_1\ V_1')^{-1}\Sigma_{Xy}\\ &= \left(S_{XX} + \lambda I\right)^{-1}\left[(S_{Xy} - \Sigma_{Xy}) + (\Sigma_{XX} - S_{XX})\beta_1 - \lambda\beta_1\right].\end{aligned} \tag{37}$$

This yields the following bound for the point forecast error

$$\begin{aligned}|X_t'(\hat{\beta}_\lambda - \beta_1)| \leq\ & \|X_t'\left(S_{XX} + \lambda I\right)^{-1}\|\\ & \times\left[\|S_{Xy} - \Sigma_{Xy}\| + \|\Sigma_{XX} - S_{XX}\|\ \|\beta_1\| + \lambda\|\beta_1\|\right].\end{aligned} \tag{38}$$

To bound the term between brackets, we can use Lemmas 2 and 3 which yield

$$\begin{aligned}&\|S_{Xy} - \Sigma_{Xy}\| + \|\Sigma_{XX} - S_{XX}\|\ \|\beta_1\| + \lambda\|\beta_1\|\\ &\leq O_p\left(\frac{\sqrt{n}}{\sqrt{T}}\right) + O_p\left(\frac{n}{\sqrt{T}}\right)O\left(\frac{1}{\sqrt{d_1^{\min}}}\right) + \lambda\ O\left(\frac{1}{\sqrt{d_1^{\min}}}\right).\end{aligned} \tag{39}$$

To bound the first term, we use the spectral decomposition of S_{XX}

$$\begin{aligned}\|X_t'\left(S_{XX} + \lambda I\right)^{-1}\| &= \|X_t'\ \hat{V}_1\ (\hat{D}_1 + \lambda)^{-1}\ \hat{V}_1' + X_t'\ \hat{V}_2\ (\hat{D}_2 + \lambda)^{-1}\ \hat{V}_2'\|\\ &\leq \frac{O_p(\sqrt{n})}{\hat{d}_1^{\min}} + \frac{1}{\lambda}\ \|X_t'\ \hat{V}_2\|,\end{aligned} \tag{40}$$

where we used the bounds

$$\|X_t'\| \leq O_p(\sqrt{n}), \tag{41}$$

which follows from Lemma 1 and Assumption 1,

$$\|\hat{V}_1\ (\hat{D}_1 + \lambda)^{-1}\ \hat{V}_1'\| \leq \frac{1}{\hat{d}_1^{\min}}, \tag{42}$$

and

$$\|(\hat{D}_2 + \lambda)^{-1}\ \hat{V}_2'\| \leq \frac{1}{\lambda}. \tag{43}$$

Let us remark that the latter two spectral bounds are always valid but will be bad when λ does not dominate the eigenvalues of the second cluster while being smaller than those of the first. But this is what we expect of an appropriate choice of λ since we want to capture the forecast driven by the dominating cluster. Finally, replacing $\hat{d}_1^{\min}$ by $d_1^{\min} + O_p(n/\sqrt{T})$ thanks to Lemma 4, using Lemma 6 to bound $\|X_t' \hat{V}_2\|$, and injecting all previous bounds into (38), we obtain the asymptotic upper bound (35) for the point forecast error. □

To derive consistency rates, we introduce further assumptions about the asymptotic growth rates of the extreme population eigenvalues of the two clusters. More specifically, as a simple model, we will assume that $d_1^{\max}$ grows proportionally to the dimension n whereas $d_1^{\min}$ grows at the possibly slower rate $n^{1-\delta}$ with $\delta \geq 0$. The maximal eigenvalue of the second cluster $d_2^{\max}$ is supposed to grow proportionally to $n^{1-\alpha}$ with $\alpha \leq 1$ and $\alpha > \delta$, so that, asymptotically, the gap (31), $gap(n) \equiv d_1^{\min} - d_2^{\max} = n^{1-\delta} - n^{1-\alpha}$, is positive and grows with n. This is summarized into the following assumption, using the notation 'Big Theta', $\Theta(\cdot)$, introduced in the notation subsection at the end of Section 1.

Assumption 5. The extreme eigenvalues of the two clusters of the population covariance matrix behave asymptotically as follows:

$$d_1^{\max} = \Theta(n); \qquad d_1^{\min} = \Theta(n^{1-\delta}); \qquad d_2^{\max} = \Theta(n^{1-\alpha}), \tag{44}$$

with $0 \leq \delta < \alpha \leq 1$.

Under this simple model, we can establish the following bound for the forecast error (35).

Proposition 2. *Under Assumption 5, we get the consistency rate*

$$|X_t'(\hat{\beta}_\lambda - \beta_1)| \leq O_p\left(\frac{n^{\frac{3}{2}\delta}}{\sqrt{T}} + n^{\frac{\delta-\alpha}{2}}\right), \tag{45}$$

asymptotically for $n, T \to \infty$, if the ridge parameter is set to $\lambda = n/\sqrt{T}$.

Proof. Using the approximation $\frac{n}{n^{1-\delta}-n^{1-\alpha}} = \frac{n^\delta}{1-n^{\delta-\alpha}} \sim n^\delta$ and plugging the growth rates (44) into (35), we get

$$\begin{aligned} |X_t'(\hat{\beta}_\lambda - \beta_1)| \leq_p \; & \frac{n^\delta}{\sqrt{T}}\left(1 + n^{\frac{1}{2}\delta}\right) + n^{\frac{\delta-1}{2}}\left(\frac{\sqrt{n}}{\sqrt{T}}\, n^\delta + n^{\frac{1-\alpha}{2}}\right) \\ & + \frac{1}{\lambda}\left(\frac{\sqrt{n}}{\sqrt{T}}\, n^\delta + n^{\frac{1-\alpha}{2}}\right)\frac{\sqrt{n}}{\sqrt{T}}\left(1 + n^{\frac{\delta}{2}}\right) \\ & + \lambda\; n^{\frac{3}{2}\delta - 1}, \end{aligned} \tag{46}$$

where we dropped all terms $O_p(n/\sqrt{T})$ in the denominators in (35), assuming that $\sqrt{T}$ is growing faster than n^δ. We see that asymptotic consistency requires $n^{\frac{3}{2}\delta}/\sqrt{T} \to 0$ for $n, T \to \infty$, which implies indeed $n^\delta/\sqrt{T} \to 0$. Keeping only the asymptotically dominating terms, we finally obtain

$$\begin{aligned} |X_t'(\hat{\beta}_\lambda - \beta_1)| \leq_p \; & \frac{n^{\frac{3}{2}\delta}}{\sqrt{T}} + n^{\frac{\delta-\alpha}{2}} \\ & + \frac{1}{\lambda}\left(\frac{n^{1+\frac{3}{2}\delta}}{T} + \frac{n^{1+\frac{\delta-\alpha}{2}}}{\sqrt{T}}\right) \\ & + \lambda\; n^{\frac{3}{2}\delta - 1}. \end{aligned} \tag{47}$$

We see that with the choice of the ridge parameter $\lambda = \frac{n}{\sqrt{T}}$, the last two terms are of the same order as the first two ones, which then govern the asymptotic rate. □

Hence, the bias between the forecast driven by the dominating cluster and the one provided by the ridge regression estimator vanishes asymptotically for $n, T \to \infty$ provided that the ridge parameter is properly tuned. The consistency rate for the variance is just the square of the rate in (45). Let us stress the fact that for $\alpha = \delta$, i.e. in the absence of a spectral gap in the population covariance matrix, there is one term which does not vanish asymptotically for $n \to \infty$, meaning that we cannot show consistency for the capture of the forecast driven by the dominating subspace. Notice also that, in the case of a fixed dimension n, it is known that the ridge parameter λ should be asymptotically of the order of $1/\sqrt{T}$ to achieve the best bias-variance tradeoff. In a way, our results generalize such a tuning to the high-dimensional case.

5. Consistency rates under a factor model

In this section, we consider the case where the high-dimensional time series is driven by a smaller dimensional one, i.e. where it obeys to a so-called factor model, implying that the forecast of y_t depends only on a finite (and small) number r of unobserved factors. As in De Mol et al. (2008), we assume the following:

Assumption 6. We have

$$y_{t+h} = \gamma F_t + v_{t+h} \tag{48}$$

where v_{t+h} 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 matrix $E(F_t F_t') = I_{r\times r}$.

The forecast based on the projection on the unobserved factors F_t, namely

$$y^*_{t+h|t} = \gamma F_t, \tag{49}$$

is optimal since its forecast accuracy cannot be improved in view of the assumption of orthogonality between the residuals v_{t+h} and the observed predictors X_t. However, for fixed n, the optimal forecast is unfeasible, even with an infinite sample size T, since the factors are unobserved.

We further assume that the observed predictors X_t are related to the common factors as follows:

Assumption 7. We have

$$X_t = \Lambda F_t + \xi_t, \tag{50}$$

where the residuals ξ_t are a n-dimensional stationary process with covariance matrix $E(\xi_t \xi_t') = \Psi$ of full rank for all n and are orthogonal to the factors F_t. The $n \times r$ matrix Λ of the loadings is a non-random matrix of full rank r for each n.

Notice that, in this model, only the common part ΛF_t driven by the factors is informative about the future of the target variable whereas the residuals – the so-called 'idiosyncratic component' ξ_t – are not. This assumption is justified by the finding in Luciani (2014) who studies the role of 'non-pervasive' shocks when forecasting with factor models with a sparse idiosyncratic component. He shows, both in simulations and on a large panel of US quarterly data, that the idiosyncratic component does not help forecasting macroeconomic variables. Moreover, the assumption that Ψ is of full rank entails that there are no redundant predictors, in the sense that, when we increase the number of predictors, information is not duplicated.

Under Assumptions 6 and 7, we have:

$$\Sigma_{XX} = E(X_t X_t') = \Lambda\Lambda' + \Psi \quad \text{and} \quad \Sigma_{Xy} = E(X_t y_{t+h}) = \Lambda\gamma' \tag{51}$$

where Σ_{XX} is invertible for all n. Consequently, for a given number n of predictors, the population OLS regression coefficient (4) is unique and the forecast is given by:

$$y_{t+h|t} = X_t'\beta = X_t'(\Lambda\Lambda' + \Psi)^{-1}\Lambda\gamma'. \tag{52}$$

Such a factor model induces the presence of two clusters in the spectrum of the covariance matrix Σ_{XX}, the first one being dominated by $\Lambda\Lambda'$ while the second one is driven by Ψ. As done in De Mol et al. (2008), the following assumptions are usually made on the corresponding eigenvalues.

Let us first consider the case of an exact factor model where we have uncorrelated components ξ_i, i.e. where $\Psi = I$. Then $\Lambda\Lambda'$ and Σ_{XX} share the same eigenvectors. For the first cluster, corresponding to the first $n_1 = r$ largest eigenvalues, it is assumed that for $n \to \infty$, the eigenvalues of $\Lambda\Lambda'$ all grow as $\Theta(n)$, so that for the extreme eigenvalues of the first cluster, we have

$$d_1^{\max} = \Theta(n) \quad \text{and} \quad d_1^{\min} = \Theta(n). \tag{53}$$

This means that in (44) we have $\delta = 0$. As for the second cluster, we have only the eigenvalues of $\Psi = I$, so that $d_2^{\max} = 1$. In this case, we also see from the expression $\beta = (\Lambda\Lambda' + I)^{-1}\Lambda\gamma'$, that β belongs to the r-dimensional range of the matrix Λ, or in other words that $\beta = \beta_1$ and $\beta_2 = 0$.

However, we would like to allow the residuals ξ_t to be weakly correlated across predictors, though less pervasive than the common component ΛF_t. To treat the case where $\Psi \neq I$, we need to be more careful, since in general $\Lambda\Lambda'$ cannot be diagonalized on the same basis. Now, assuming as above that the first cluster is spanned by the eigenvectors of $\Lambda\Lambda'$ and that $n_1 = r$, and assuming that the maximal eigenvalue of Ψ, $d^{\max}(\Psi) = \Theta(n^{1-\alpha})$ with $0 < \alpha \leq 1$, we remark that adding to $\Lambda\Lambda'$ the positive-definite matrix Ψ will not modify the leading growth rate (53), which remains valid. On the eigenspace corresponding to the second cluster, $\Lambda\Lambda'$ vanishes, so that the maximal eigenvalue of Σ_{XX} coincides with that of Ψ. Hence

$$d_2^{\max} = d^{\max}(\Psi) = \Theta(n^{1-\alpha}) \quad \text{with} \quad 0 < \alpha \leq 1. \tag{54}$$

Also in the case $\Psi \neq I$, we see from the expression $\beta = (\Lambda\Lambda' + \Psi)^{-1}\Lambda\gamma'$ that β is a linear combination of the columns of Λ, hence belongs to the first cluster, or else that $\beta = \beta_1$ and $\beta_2 = 0$. Hence we have

$$\|\beta\| = O\left(\frac{1}{\sqrt{n}}\right), \tag{55}$$

which was an essential ingredient for the consistency proofs by De Mol et al. (2008). This is similar to the bound (16) which here is entailed by the factor model instead of relying on condition (6) as assumed in Lemma 2.

The following result replicates Proposition 2 of Section 4 in the case of a factor model, setting $\beta_1 = \beta$ and $\delta = 0$.

Proposition 3. *Under the factor-model Assumptions 6 and 7 with $d_1^{\max} = d_1^{\min} = \Theta(n)$ and $d_2^{\max} = \Theta(n^{1-\alpha})$, $\alpha > 0$, we get the consistency rate for the ridge estimator $\hat{\beta}_\lambda$*

$$|X_t'(\hat{\beta}_\lambda - \beta)| \leq O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{(\sqrt{n})^\alpha}\right), \tag{56}$$

asymptotically for $n, T \to \infty$, provided that the ridge parameter is set to $\lambda = n/\sqrt{T}$.

Proof. Redo the proof of Proposition 1, suppressing all subscripts 1 in the derivation from equation (36) to equation (39), using the bound (55) instead of (16), and keeping the rest of the derivations unchanged. Then use Proposition 2 setting $\delta = 0$. □

As shown in De Mol et al. (2008), in the case of a factor model, we can asymptotically capture not only the forecast $y_{t+h|t} = X_t'\beta$ driven by the dominating subspace or common component, but also the optimal forecast $y_{t+h|t}^* = \gamma F_t$. This result is reproduced in the following proposition.

Proposition 4. *Under the factor-model Assumptions 6 and 7 and if $d_2^{\max} = \Theta(n^{1-\alpha})$, we have*

$$y_{t+h|t} - y_{t+h|t}^* = O_p\left(\frac{1}{(\sqrt{n})^\alpha}\right), \tag{57}$$

asymptotically for $n \to \infty$.

Proof. By a straightforward matrix identity, we get

$$\beta = \Sigma_{XX}^{-1}\Sigma_{Xy} = (\Lambda\Lambda' + \Psi)^{-1}\Lambda\gamma' = \Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma'. \tag{58}$$

According to (52), (49) and (50), we have

$$\begin{aligned} y_{t+h|t} - y_{t+h|t}^* &= X_t'\beta - F_t'\gamma' = (F_t'\Lambda' + \xi_t')\beta - F_t'\gamma' \\ &= F_t'\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma' - F_t'\gamma' + \xi_t'\beta. \end{aligned} \tag{59}$$

To bound the first two terms, we notice that

$$\|\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1} - I\| = \|(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\| \leq \frac{1}{d^{\min}(\Lambda'\Psi^{-1}\Lambda)} \tag{60}$$

where $d^{\min}(A)$ denotes the minimum eigenvalue of the matrix A (and $d^{\max}(A)$ its maximum one). To bound the third term, we use the fact that

$$\begin{aligned} E\left[\left(\xi_t'\beta\right)^2\right] &= \beta'\Psi\beta = \gamma(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\Lambda'\Psi^{-1}\Psi\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma' \\ &\leq \|\gamma\|^2\|(\Lambda'\Psi^{-1}\Lambda)^{-1}\|\,\|(\Lambda'\Psi^{-1}\Lambda)(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\| \\ &\leq \|\gamma\|^2\,\frac{1}{d^{\min}(\Lambda'\Psi^{-1}\Lambda)}, \end{aligned} \tag{61}$$

which using $\|\gamma\| = O(1)$ and Lemma 1 implies

$$\xi_t'\beta = O_p\left(d^{\min}(\Lambda'\Psi^{-1}\Lambda)\right)^{-1/2}. \tag{62}$$

Since by the assumptions (53) and (54) made on the factor model,

$$d^{\min}(\Lambda'\Psi^{-1}\Lambda) \geq d^{\min}(\Psi^{-1})\;d^{\min}(\Lambda'\Lambda) = \left(d^{\max}(\Psi)\right)^{-1}d^{\min}(\Lambda'\Lambda) \geq 1, \tag{63}$$

and since the term (62) dominates (60), we get the announced rate (57). □

Combining the results of the two last propositions, we get the following asymptotically consistency rate for the ridge regression estimator to the optimal forecast under a factor model.

Theorem 1. *Under the factor-model Assumptions 6 and 7 with $d_2^{\max} = \Theta(n^{1-\alpha})$, we get the following consistency rate for the bias of the ridge estimator $\hat{\beta}_\lambda$ with $\lambda = n/\sqrt{T}$:*

$$|X_t'\hat{\beta}_\lambda - y_{t+h|t}^*| \leq O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{(\sqrt{n})^\alpha}\right), \tag{64}$$

asymptotically for $n, T \to \infty$, with no restriction on the path in the (n, T)-plane. For a strong factor model, this holds with $\alpha = 1$.

Proof. This property results immediately from Propositions 3 and 4. □

This asymptotic rate constitutes an improvement with respect to the rate derived in De Mol et al. (2008) where the term depending on T is only $T^{-1/4}$.

6. Comparison with Principal Component Regression

The ridge estimator (7)-(8), via the addition of the penalty $\lambda\|\beta\|^2$ to the least squares term, modifies the spectrum of the empirical covariance matrix S_{XX} by shifting away from zero its smallest eigenvalues, which are problematic for inverting the matrix in high dimension. On the other hand, estimators based on principal components sharply truncates this spectrum, keeping only the largest eigenvalues and eliminating the others. Hence, assuming that we know the right truncation point, i.e. the number n_1 of the eigenvalues in the dominating cluster, the corresponding estimator of the regression coefficient based on PC is given by

$$\hat{\beta}_{PC} = \hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1'\ S_{Xy}. \tag{65}$$

Indeed, using the diagonal form (12) of the empirical covariance matrix S_{XX}, it is easy to see that this estimator coincides with the solution of the least squares regression when keeping only the first n_1 PC as predictors.

Then, under the model (44) of Assumption 5 for the eigenvalues, we get the following asymptotic rate for the capture of the forecast $X_t'\beta_1$ driven by the dominating cluster.

Proposition 5. *If $d_1^{\min} = \Theta(n^{1-\delta})$, we have the consistency rate*

$$|X_t'(\hat{\beta}_{PC} - \beta_1)| \le O_p\left(\frac{n^{\frac{3}{2}\delta}}{\sqrt{T}}\right), \tag{66}$$

asymptotically for $n, T \to \infty$.

Proof. We have

$$\hat{\beta}_{PC} - \beta_1 = \hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1'\ S_{Xy} - V_1\ D_1^{-1}\ V_1'\ \Sigma_{Xy} \tag{67}$$

or else

$$\hat{\beta}_{PC} - \beta_1 = \hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1'\ (S_{Xy} - \Sigma_{Xy}) + (\hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1' - V_1\ D_1^{-1}\ V_1')\ \Sigma_{Xy}. \tag{68}$$

Using the identity

$$\hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1' - V_1\ D_1^{-1}\ V_1' = \hat{V}_1\ \hat{D}_1^{-1}\ \hat{V}_1'\left[V_1\ D_1\ V_1' - \hat{V}_1\ \hat{D}_1\ \hat{V}_1'\right] V_1\ D_1^{-1}\ V_1' \tag{69}$$

we get

$$\hat{\beta}_{PC} - \beta_1 = \hat{V}_1\,\hat{D}_1^{-1}\,\hat{V}_1'\Big[(S_{Xy} - \Sigma_{Xy}) + [V_1\,D_1\,V_1' - \hat{V}_1\,\hat{D}_1\,\hat{V}_1']\,V_1\,D_1^{-1}\,V_1'\,\Sigma_{Xy}\Big]. \tag{70}$$

Using Lemma 3 on the behavior of the covariance matrices as well as Lemma 2 and the bound (41), we get the following bound for the point forecast error

$$|X_t'(\hat{\beta}_{PC} - \beta_1)| \le_p \frac{\sqrt{n}}{\hat{d}_1^{\min}}\left(\frac{\sqrt{n}}{\sqrt{T}} + \frac{n}{\sqrt{T}}\frac{1}{\sqrt{d_1^{\min}}}\right). \tag{71}$$

Replacing $\hat{d}_1^{\min}$ by $d_1^{\min} + O_p(\frac{n}{\sqrt{T}})$ according to Lemma 4, assuming the behavior $d_1^{\min} = \Theta(n^{1-\delta})$, and keeping only the dominating term, we get the rate (66). □

As in the case of the ridge estimator, we see that asymptotic consistency requires $n^{\frac{3}{2}\delta}/\sqrt{T} \to 0$ for $n, T \to \infty$. The rate for the variance is just the square of this rate for the bias.

Under the factor model described in Section 5, we have $\beta = \beta_1$ and $\delta = 0$, so that we get a rate $1/\sqrt{T}$ for any path, independently of the growth of n. This result could also be proved anew by replacing everywhere in the derivation here above β_1 by β and $V_1\ D_1^{-1}\ V_1'$ by $VD^{-1}V'$, the complete eigendecomposition of Σ_{XX}. Moreover, we can then pretend to capture also the optimal forecast $y^*_{t+h|t}$ driven by the factors and use Proposition 4, which combined with the previous Proposition 5, yields the following theorem.

Theorem 2. *Under the factor-model Assumptions 6 and 7, we get the following consistency rate for the bias of the principal component estimator $\hat{\beta}_{PC}$ with respect to the optimal forecast :*

$$|X_t'\hat{\beta}_{PC} - y^*_{t+h|t}| \le O_p\left(\frac{1}{\sqrt{T}} + \frac{1}{(\sqrt{n})^{\alpha}}\right), \tag{72}$$

asymptotically for $n, T \to \infty$, with no restriction on the path in the (n, T)-plane. For a strong factor model, this holds with $\alpha = 1$.

We notice that this is exactly the same rate as for the ridge estimator. However, let us stress the fact that, to get consistency, knowing the exact number of factors $r = n_1$ is crucial here to set the truncation point right on the number of principal components used. On the other hand, in the ridge estimator, setting the parameter λ to $n/\sqrt{T}$, i.e. to the right order of growth, would be sufficient to this purpose. Such flexibility might perhaps constitute an advantage for ridge regression compared to principal component regression, at least in such a high-dimensional time series context. However, in practice, it is not easy to pick the right tuning since this only gives an asymptotic rate. In regularization theory in a deterministic setting (with a fixed design), there is a vast literature dedicated to the choice of the regularization parameter in ridge or of the truncation point for principal component regression (and to the relation between the two choices). In the present stochastic setting and for factor models, there are also many papers dealing with the determination of the number of factors, including by Bai and Ng (2002) who use information criteria. However, to the best of our knowledge, nothing similar is available for the ridge parameter and our results could constitute a starting point to investigate this question. In finite samples, as done in the empirics reported by De Mol et al. (2008), cross-validation remains a route of choice to select these tuning parameters.

7. Conclusions

The paper establishes (n, T) asymptotic results for ridge regression and PC regression. It contains two novel results. First, it proves that under less restrictive assumptions than strong factors (or ‘approximate factor model’), the ridge regression estimator yields asymptotically for large n and T the same forecast as the forecast driven by the common component of the data, i.e. the component associated to the dominant eigenvalues of the population covariance matrix. Second, in the special case of a strong factor structure, it establishes improved (n, T) consistency rates for ridge regression, which turn out to be the same as those available for PC in the literature. Our results imply that compression via PC and shrinkage via ridge are equivalent. Both methods capture the bulk of the information contained in a large number of strongly correlated predictors and are therefore suitable for predictors that strongly comove, as it is the case in several economic situations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements and Disclaimer

The paper is dedicated to Marco Lippi for his 80th birthday. We thank the two anonymous referees as well as the Editors for their useful comments and suggestions, which led to an improved version of the paper.

The views expressed in this paper are those of the authors and do not necessarily reflect the position of the International Monetary Fund or its management or its executive directors.

References

Bai, J., 2003. Inferential Theory for Factor Models of Large Dimensions. Econometrica 71 (1), 135–171.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.
Bai, J., Ng, S., 2023. Approximate factor models with weaker loadings. Journal of Econometrics 235 (2), 1893–1916.
Barigozzi, M., 2022. On Estimation and Inference of Large Approximate Dynamic Factor Models via the Principal Component Analysis. arXiv:2211.01921.
Barigozzi, M., Luciani, M., 2019. Quasi Maximum Likelihood Estimation and Inference of Large Approximate Dynamic Factor Models via the EM algorithm. arXiv:1910.03821.
Bańbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian vector auto regressions. Journal of Applied Econometrics 25 (1), 71–92.
Davis, C., Kahan, W.M., 1970. The Rotation of Eigenvectors by a Perturbation. III. SIAM Journal on Numerical Analysis 7 (1), 1–46.
De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146 (2), 318–328.
Doz, C., Giannone, D., Reichlin, L., 2012. A Quasi-Maximum Likelihood Approach for Large Approximate Dynamic Factor Models. The Review of Economics and Statistics 94 (4), 1014–1024.
Forni, M., Giannone, D., Lippi, M., Reichlin, L., 2009. Opening the Black Box: Structural Factor Models with Large Cross Sections. Econometric Theory 25, 1319–1347.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The Generalized Dynamic-Factor Model: Identification and Estimation. The Review of Economics and Statistics 82 (4), 540–554.
Giannone, D., Lenza, M., Primiceri, G.E., 2021. Economic predictions with big data: The illusion of sparsity. Econometrica 89 (5), 2409–2437.
Hauzenberger, N., Huber, F., Koop, G., 2024. Macroeconomic forecasting using BVARs (forthcoming in: *Research Methods and Applications on Macroeconomic Forecasting*, M. Clements and A. Gavao, eds.). Edward Elgar.
He, Y., 2023. Ridge Regression Under Dense Factor Augmented Models. Journal of the American Statistical Association (forthcoming) 0 (0), 1–13.
Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12 (1), 55–67.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer.
Leamer, E.E., 1973. Multicollinearity: A Bayesian Interpretation. The Review of Economics and Statistics 55 (3), 371–380.
Luciani, M., 2014. Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. International Journal of Forecasting 30 (1), 20–29.
Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics 168 (2), 244–258.
Stock, J.H., Watson, M.W., 2002. Forecasting Using Principal Components from a Large Number of Predictors. Journal of the American Statistical Association 97, 1167–1179.
Tikhonov, A.N., 1963. Solution of Incorrectly Formulated Problems and the Regularization Method. Soviet Mathematics Doklady 4, 1035–1038.
Yu, Y., Wang, T., Samworth, R.J., 2014. A useful variant of the Davis-Kahan theorem for statisticians. Biometrika 102 (2), 315–323.