



Ridge Regression Under Dense Factor Augmented Models

Yi He

To cite this article: Yi He (2023): Ridge Regression Under Dense Factor Augmented Models, Journal of the American Statistical Association, DOI: [10.1080/01621459.2023.2206082](https://doi.org/10.1080/01621459.2023.2206082)

To link to this article: <https://doi.org/10.1080/01621459.2023.2206082>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.



View supplementary material [↗](#)



Published online: 05 Jun 2023.



Submit your article to this journal [↗](#)



Article views: 2413



View related articles [↗](#)



View Crossmark data [↗](#)

Ridge Regression Under Dense Factor Augmented Models

Yi He^a 

Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

This article establishes a comprehensive theory of the optimality, robustness, and cross-validation selection consistency for the ridge regression under factor-augmented models with possibly dense idiosyncratic information. Using spectral analysis for random matrices, we show that the ridge regression is asymptotically efficient in capturing both factor and idiosyncratic information by minimizing the limiting predictive loss among the entire class of spectral regularized estimators under large-dimensional factor models and mixed-effects hypothesis. We derive an asymptotically optimal ridge penalty in closed form and prove that a bias-corrected k -fold cross-validation procedure can adaptively select the best ridge penalty in large samples. We extend the theory to the autoregressive models with many exogenous variables and establish a consistent cross-validation procedure using the what-we-called double ridge regression method. Our results allow for nonparametric distributions for, possibly heavy-tailed, martingale difference errors and idiosyncratic random coefficients and adapt to the cross-sectional and temporal dependence structures of the large-dimensional predictors. We demonstrate the performance of our ridge estimators in simulated examples as well as an economic dataset. All the proofs are available in the supplementary materials, which also includes more technical discussions and remarks, extra simulation results, and useful lemmas that may be of independent interest.

ARTICLE HISTORY

Received September 2021
Accepted April 2023

KEYWORDS

Cross-validation;
High-dimensional linear
model; Mixed-effects model;
Spectral analysis; Tikhonov
regularization

1. Introduction

Improving the bias-variance tradeoff is crucial in many high-dimensional forecasting applications. The classical least-squares regression can suffer from substantial predictive variance when the number of predictors is large compared with the sample size. One useful variance reduction strategy is summarizing massive raw features into a few leading principal components and then predicting the response with these low-dimensional predictors. When the large-dimensional covariates obey a factor structure, the principal component estimator can provide consistent forecasts of the regression function if the factors suffice to predict the target variable. The principal component regression (PCR) often shows better performance than the LASSO regression and ensemble algorithms such as bagging for macroeconomic forecasting; see, for example, Stock and Watson (2002), De Mol, Giannone, and Reichlin (2008), Stock and Watson (2012), Castle, Clements, and Hendry (2013), and Carrasco and Rossi (2016). The PCR is also widely used when studying DNA microarrays (Alter, Brown, and Botstein 2000), medical shapes (Cootes et al. 1995), climate (Preisendorfer 1998), robust synthetic control (Agarwal et al. 2021) and many other fields (see Chapter 6 of James et al. 2021). While the PCR model is dense at the variable level (see, e.g., the discussions in Ng 2013), it relies on only low-dimensional principal components in forecasting. However, when many idiosyncratic components are useful, even though each is negligible, the PCR is not optimal and leaves room for improvement in many empirical applications.

Consider a toy example with $p = 100$ covariates $x_i = (x_{i,1}, \dots, x_{i,p})' \in \mathbb{R}^p$, where we denote the transpose of any matrix or vector A by A' , and $n = 100$ observations from a factor model given by

$$x_i = \Lambda_1 f_{i,1} + \Lambda_2 f_{i,2} + e_i, \quad i = 1, \dots, n. \quad (1.1)$$


The latent factors $f_{i,1} \in \mathbb{R}$ and $f_{i,2} \in \mathbb{R}$, the entries of idiosyncratic variables $e_i = (e_{i,1}, \dots, e_{i,p})'$, and the entries of loading coefficients $\Lambda_1 = (\Lambda_{1,1}, \dots, \Lambda_{p,1})'$ and $\Lambda_2 = (\Lambda_{1,2}, \dots, \Lambda_{p,2})'$ are all generated independently from the standard normal distribution. One may think of the data vector x_i as a set of economic variables at a given time. We then generate the target variables y_i from a factor-augmented predictive regression model given by

$$y_i = 0.6f_{i,1} + 0.2f_{i,2} + e_i'\beta + \varepsilon_i =: \mu_i + \varepsilon_i, \quad (1.2)$$

where the regression errors ε_i are generated from the standard normal distribution independently of the regression means μ_i . We generate the direction of the coefficient vector $\beta = (\beta_1, \dots, \beta_p)'$ uniformly over the \mathbb{R}^p unit sphere, and thus each entry β_i is stochastically negligible for large p . Then we scale the total idiosyncratic signal $\|\beta\|^2 = \sum_{i=1}^p \beta_i^2$ through the grid $\{0, 0.1, \dots, 1\}$. The PCR model is the special case with $\|\beta\|^2 = 0$. The idiosyncratic information, at aggregate level, becomes more important as $\|\beta\|^2$ increases.

We plot the median, over 5000 replications, of the training predictive loss $n^{-1} \sum_{i=1}^n (\mu_i - \hat{\mu}_i)^2$ as a function of idiosyncratic signal length $\|\beta\|^2$ on the left-hand-side of Figure 1, where

CONTACT Yi He  y.he2@uva.nl  Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

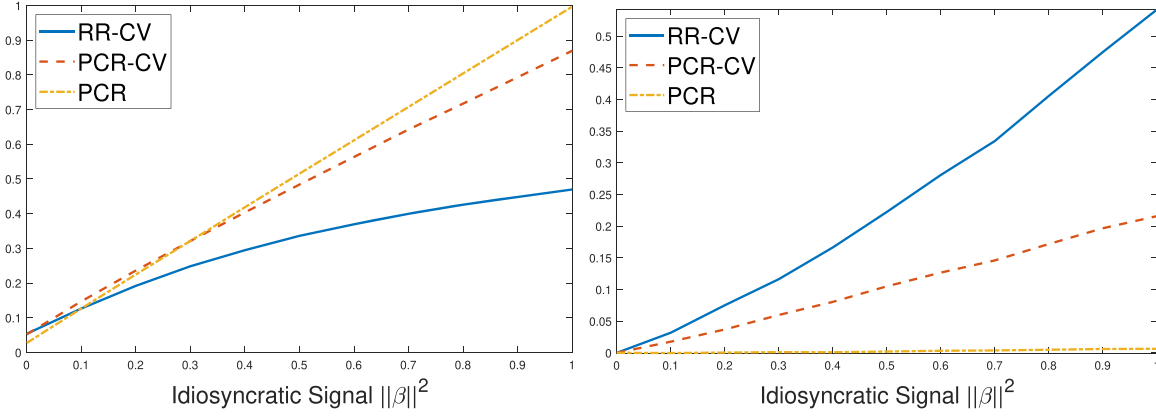


Figure 1. Median predictive loss (left) and change in median sample variance of estimated means (right) as functions of idiosyncratic signal $\|\beta\|^2$.

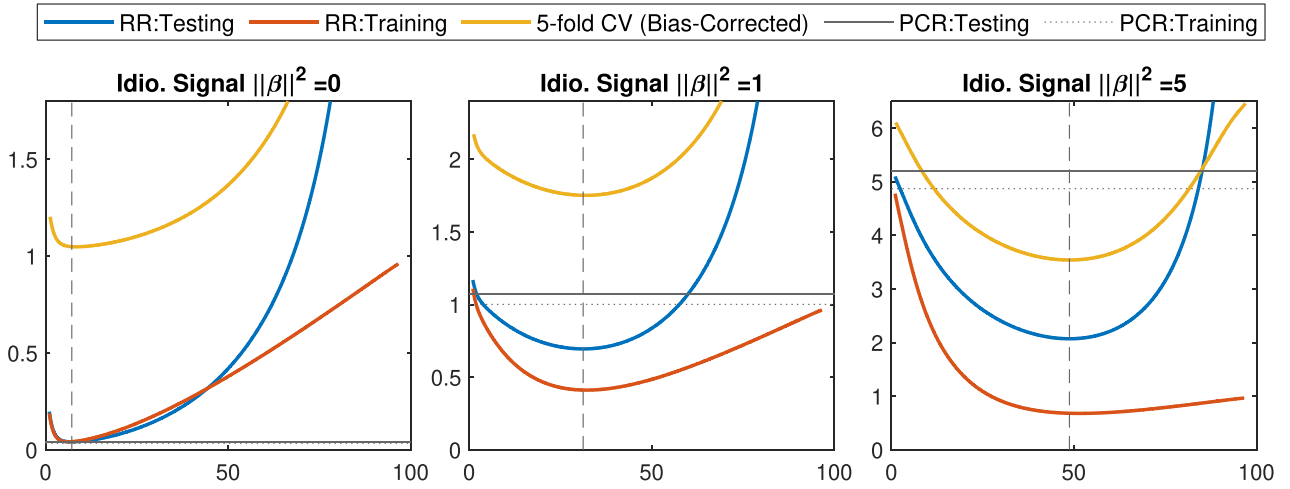


Figure 2. Average training and testing predictive losses for RR and Oracle PCR.

μ_i and $\hat{\mu}_i$ denote the population and estimated means of the target variable y_i , respectively. We consider both the oracle PCR estimator using the true number of factors $r = 2$, and the feasible PCR estimator using a number of PCs selected by 5-fold cross-validation. We compare these PCR estimators with the ridge regression estimator, hereinafter abbreviated as RR estimator, that uses all the raw variables x_i and does not require selecting the number of factors. We select the ridge penalty using 5-fold cross-validation. The median predictive loss of the PCR estimators grows (almost) linearly in the idiosyncratic signal length $\|\beta\|^2$, but that of the RR estimator grows significantly slower. The cross-validated RR estimator is only slightly worse than the oracle PCR estimator but comparable to the cross-validated PCR estimator for small $\|\beta\|^2$, whereas the advantage of RR estimator emerges quickly as $\|\beta\|^2$ grows.

To illustrate the sensitivity of RR to the idiosyncratic information, we plot the median of the sample variance of the estimated means (which equals to the sum of squared estimated coefficients on all standardized components) over the replications on the right-hand-side of Figure 1. To emphasize the changes, we only plot the differences to the values for $\|\beta\|^2 = 0$. The variance of RR estimates grow quickly with $\|\beta\|^2$, while that of the oracle PCR coefficients hardly change. The cross-validation selection improves the sensitivity of PCR to $\|\beta\|^2$ only slightly because a

few extra components are actually asymptotically negligible in our dense models.

Figure 2 compares the *training* and *testing* predictive losses for ridge regression, averaged over 1000 replications, as functions of the *effective degrees of freedom* (see Hastie, Tibshirani, and Friedman 2009, p. 64) implied by the ridge penalty. We define *testing* predictive loss as the *expected* squared forecast error of the out-of-sample regression mean $\mu_{\text{oos}} = f'_{\text{oos}}\theta + e'_{\text{oos}}\beta$ with new variables f_{oos} and e_{oos} using the estimated regression function and the new covariates x_{oos} , where $\theta = (0.6, 0.2)'$ and β are the population coefficients in (1.2). Again, the optimal ridge estimator (dashed line) for testing predictive loss matches the performance of PCR when $\|\beta\|^2 = 0$ but outperforms PCR when $\|\beta\|^2 > 0$. More importantly, these plots illustrate an important relation between training and testing predictive losses for performance evaluation: they share (almost) the same optimal ridge penalty. We formally establish this universal property in Section 2 using a general notation of predictive loss, and generalize the results to time series data in Section 3.

Our toy example illustrates that omitting idiosyncratic information can lead to nontrivial loss in predictive efficiency, and the ridge regression is more robust against PCR capturing both factor and idiosyncratic information. The contribution of this article is a comprehensive theory of the optimality, robustness,

and cross-validation consistency of the ridge regression under a general factor-augmented regression model given by

$$X = F\Lambda' + E\Gamma' \quad (1.3)$$

$$Y = F\theta + E\beta + \epsilon \quad (1.4)$$

where the left-hand side of the equations represent a large sample of observations in high dimensions: $X = [x_1 - \bar{x}, \dots, x_n - \bar{x}]' \in \mathbb{R}^{n \times p}$ is the demeaned design matrix of predictors $x_i \in \mathbb{R}^p$ with large n and p , and $Y = (y_1 - \bar{y}, \dots, y_n - \bar{y})'$ is the demeaned vector of univariate targets y_i . Throughout, $\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$ denotes the sample mean of any variable a . The low-rank matrix $F = [f_1 - \bar{f}, \dots, f_n - \bar{f}]'$ collects low-dimensional latent factors $f_i \in \mathbb{R}^r$, r finite, and $E = [e_1 - \bar{e}, \dots, e_n - \bar{e}]'$ is a large-dimensional random matrix of idiosyncratic components $e_i \in \mathbb{R}^p$. The matrix Γ is a rotation matrix, possibly unknown. The fixed-effects $\theta = (\theta_1, \dots, \theta_r)'$ on the factors and the random-effects $\beta = (\beta_1, \dots, \beta_p)'$ on the idiosyncratic components are both unknown. The noise component $\epsilon = (\epsilon_1 - \bar{\epsilon}, \dots, \epsilon_n - \bar{\epsilon})'$ is a demeaned vector of uncorrelated (but possibly dependent) regression errors ϵ_i . One may also interpret (1.3) and (1.4) as an error-in-variables regression model (6.1) but we postpone the discussions to Section 6.

Many datasets support the low-rank perturbation form (1.3); see the survey by Johnstone and Paul (2018) for many examples in electrocardiogram tracing, microarrays, satellite images, medical shapes, climate, signal detection, and finance. We assume that the low-rank $p \times r$ coefficient matrix $\Lambda = [\Lambda_1, \dots, \Lambda_r]$ has a divergent strength $a_p = \text{tr}(\Lambda' \Lambda) \rightarrow \infty$ at an arbitrary rate to separate the spectrum of the factor component $F\Lambda'$ from the stochastically bounded spectrum of the idiosyncratic component $E\Gamma'$ in (1.3). Therefore, we can analyze the spectral regularization using the standard eigenbasis without suffering from PCA inconsistency issues (Paul 2007; Johnstone and Lu 2009; Onatski 2012). Following Cai, Han, and Pan (2020), we allow the unobservable idiosyncratic components to enter the factor model (1.3) subject to a possibly unknown rotation $\Gamma \in \mathbb{R}^{p \times p}$ for generality. We discuss the possibility of extending our results for $a_p = O(1)$, especially when PCA becomes inconsistent, as future works in Appendix A.9 in the supplementary materials.

Our regression model (1.4) is the factor-augmented model proposed by Kneip and Sarda (2011). Substituting $E\Gamma' = X - F\Lambda'$ under model (1.3) and using the identity $\Gamma'\Gamma = I_p$, we can reparameterize the regression model (1.4) as follows:

$$Y = F\theta + X\beta + \epsilon, \quad \beta = \Gamma\beta, \quad \theta = \theta - \Lambda'\beta, \quad (1.5)$$

where the entries of θ and β are called common and specific effects, respectively. One might also interpret the last equation as a misspecified model where we only observe the design matrix X but not the latent factor matrix F . Throughout the article, we refer factor-augmented regression models to these two equivalent representations (1.4) and (1.5) exchangeably. Note that our model augments the purely variable-based model (with $\theta = \theta - \Lambda'\Gamma\beta = (0, \dots, 0)'$) in, for example, Castle, Clements, and Hendry (2013), Fan, Ke, and Wang (2020) by adding extra factor information.

Unlike the aforementioned papers, we study the cases where the coefficient vector β is possibly dense. Even if the true regression vector β on the idiosyncratic components is sparse, the

regression vector β on the variables may contain little (or even no) zero entries due to a nontrivial rotation Γ . In general, we are interested in the models that are possibly dense at both variable and principal component levels. We consider every idiosyncratic component to be asymptotically negligible but allow them to have nontrivial aggregate explanatory power; see Jolliffe (1982), Dobriban and Wager (2018), Giannone, Lenza, and Primiceri (2021), Hastie et al. (2022), and many references therein for examples of applications.

Based on spectral analysis of large-dimensional random matrices, this article shows that the ridge regression is optimal in capturing factor and idiosyncratic information simultaneously among the entire class of spectral regularized estimators under the mixed-effects hypothesis. Ridge regression has a long tradition in statistics tracing back at least to Tikhonov (1963a, 1963b) and Hoerl and Kennard (1970); see Liu and Dobriban (2020) for an excellent survey. To our knowledge, we are the first to study the ridge regression model with factor augmentation. Our analysis of large-dimensional covariance matrices under factor models requires a different approach from that in Dicker (2016) and Dobriban and Wager (2018). In particular, our results extend to more general data generating processes of covariates by analyzing limiting predictive loss instead of predictive risk (i.e., the expected predictive loss); see Appendix A.1 in the supplementary materials for comparisons with Dobriban and Wager (2018).

We prove that the k -fold cross-validation selects the asymptotically optimal ridge penalty, as illustrated in Figure 2, subject to a straightforward bias correction. The bias-corrected selection is uniformly consistent for all positive ridge penalties bounded away from zero, including those diverging to infinity. We extend the scope of the k -fold cross-validation to non-iid data and provide theoretical guarantee for dealing with martingale difference regression errors. Furthermore, we propose a new method called “double” ridge regression and an asymptotically correct k -fold cross-validation method in the presence of nuisance variables, in particular autoregressors, under mild conditions.

The rest of the article is organized as follows. We develop the asymptotic theories for independent errors in Section 2 and then generalize the results to martingale difference errors and autoregressive models in Section 3. More simulated and real-life examples are discussed in Sections 4 and 5, respectively. We conclude the article in Section 6. All the proofs of the theorems are postponed to the supplementary materials. The supplementary materials also provides more technical discussions and remarks about the conditions and auxiliary results, together with additional simulation results and some important lemmas that may be of independent interest.

2. Main Results

First, consider the factor model (1.3) in introduction. We are interested in the asymptotic regime where the number of predictors is comparable to the sample size:

Assumption 1 (Large-dimensional Asymptotics). The number of covariates $p = p(n)$ and the sample size n are comparable in such a way that $p/n \rightarrow c \in (0, \infty)$ as $n \rightarrow \infty$.

The concentration ratio p/n , possibly larger than 1, plays a central role in our asymptotic theory. Unless specified otherwise, all asymptotic results hold as $n, p \rightarrow \infty$ simultaneously in the probability space, enlarged if necessary, with the largest sigma-algebra.

Without loss of generality, we treat both the in-sample factor scores matrix F and the loading matrix Λ in (1.3) as fixed parameters. When factors are random, our approach is equivalent to conditioning on a particular realization of F . The number of factors r is finite but not necessarily known.

Assumption 2 (Factor Model). The following conditions hold:

- (a) The total factor loading $a_p := \text{tr}(\Lambda' \Lambda) \rightarrow \infty$, and $a_p = O(p)$.
- (b) $\Sigma_\Lambda := \Lambda' \Lambda / a_p$ converges to some $r \times r$ diagonal matrix with diagonal elements $\sigma_{ii} \geq \sigma_{jj} > 0$ for $i < j$. Without loss of generality, $F'F/n = I_r$ and $\Gamma' \Gamma = I_p$ where I_d denotes $d \times d$ identity matrix.
- (c) $\lambda_{\max}(\Omega_E) = O_{\mathbb{P}}(1)$ where $\Omega_E = \frac{1}{n} E' E$ is the (unobservable) sample covariance matrix of the idiosyncratic components.

Conditions (a)–(c) are the *weaker* factor models in Bai and Ng (2021) allowing an arbitrarily slow rate of total factor strength a_p . When all entries, say, of the loading matrix $\Lambda = \{\Lambda(i, j)\}$ are bounded away from infinity, we readily have $a_p = \sum_{i=1}^p \sum_{j=1}^r \Lambda^2(i, j) = O(pr) = O(p)$ in condition (a). We allow the loading matrix Λ to have many vanishing entries and one may use $a_p \propto p^\alpha$ for any $\alpha \in (0, 1]$; see, for example, Bai (2003), Hallin and Liska (2007), and Agarwal et al. (2021) for the linear case $\alpha = 1$ and Onatski (2009) for the case $\alpha > 2/3$. Via the factor decomposition (1.3) of $X = F\Lambda' + E\Gamma'$, these conditions ensure that singular values of the factor part $F\Lambda'$ to be much larger than (hence, separated from) those of the idiosyncratic part $E\Gamma'$; see, for example, Cai, Han, and Pan (2020) for discussions and comparisons with the generalized spiked model in Johnstone (2001).

Remarkably, our condition (b) relaxes some identification conditions in the previous papers. We allow ties in the proportion of total loading $\{\sigma_{ii} : i = 1, \dots, r\}$ and require only the subspace consistency of PCA in the spirit of Jung and Marron (2009). When there are ties, all bases of the span of corresponding latent factors satisfy our conditions and therefore we are not able to identify the true ones in the data generating process. Nevertheless, we could estimate the subspace as a whole consistently and unify the further factor analysis for cross-validation and autoregressive models, as train-test splits and residual-based estimation can make latent factors indistinguishable in subsamples or after projections. One limitation of condition (b) is that all eigenvalues of Λ must diverge at the same rate of a_p , but we can easily relax this requirement and allow different rates if necessary.

Condition (c) holds for a separable random matrix $E = B^{1/2} \zeta A^{1/2}$ in (1.3) where $A \in \mathbb{R}^{p \times p}$ and $B \in \mathbb{R}^{n \times n}$ are positive-definite matrices with bounded spectral norms and $\zeta = \{\zeta_{ij}\} \in \mathbb{R}^{n \times p}$ is a random matrix of iid standardized entries with a bounded fourth moment; see, for example, Paul and Silverstein (2009) and Chapter 5 of Bai and Silverstein (2010). Condition (c) also holds for high-dimensional linear time series satisfying

the conditions in Remark 8 of Onatski and Wang (2021). See Appendix A.2 of the supplementary materials for more detailed discussions.

For out-of-sample analysis, we assume that the data obeying (1.3) and (1.4) are from the following data generating process (DGP):

$$x_i = \Lambda f_i + \Gamma e_i, \quad (2.1)$$

$$y_i = \phi_0 + f_i' \theta + e_i' \beta + \varepsilon_i. \quad (2.2)$$

where $\phi_0 = \phi_0(n)$ is an unspecified intercept possibly indexed by n and the regression errors ε_i satisfy the following condition:

Assumption 3 (Independent Errors). The regression errors $\{\varepsilon_i\}$ are mutually independent with zero mean $\mathbb{E}\varepsilon_i = 0$ and a common variance $\sigma^2 = \mathbb{E}\varepsilon_i^2$ bounded away from zero and infinity, and they are independent of $\{f_i, e_i\}$. They are uniformly integrable in second moment such that $\sup_{i \in \mathbb{Z}} \mathbb{E}|\varepsilon_i|^{2+2\iota}$ is bounded away from infinity for some $\iota > 0$.

This assumption is common in statistical learning literature. We will relax it to handle uncorrelated but dependent errors in Section 3. See also Remark 2 for the relaxation of the homoscedasticity condition. We consider strongly exogenous predictors $\{x_i\}$ to allow for a standard factor estimation using the entire sample of raw predictors as in, for example, Stock and Watson (2012).

Throughout the article, we consider a mixed-effects model where the factor coefficients θ are arbitrary fixed-effects and the entries of idiosyncratic coefficient vector β are stochastic. In particular, we make the following assumption throughout the article for $\beta = (\beta_1, \dots, \beta_p)'$:

Assumption 4 (Random Effects). The idiosyncratic coefficient vector $\beta = \tau p^{-1/2} b$, where $b = (b_1, \dots, b_p)'$ is independent of the predictors $\{x_i\}$ and the regression errors $\{\varepsilon_i\}$, and has uncorrelated entries such that

$$\mathbb{E}[b_i | b_j, j \neq i] = 0, \quad \mathbb{E}[b_i^2 | b_j, j \neq i] = 1,$$

and $\max_{1 \leq i \leq p} \mathbb{E}b_i^{2+2\iota}$ is bounded away from infinity for some $\iota > 0$. The total idiosyncratic signal length $\tau^2 = \tau_n^2$ is unknown, possibly dependent on $\{x_i\}$, and $\tau^2 = O_{\mathbb{P}}(1)$. When $\tau^2 = 0$, the augmented model (2.2) reduces to the principal component regression model.

This condition is based on the hypothesis of the random effect in, for example, Dobriban and Wager (2018), but here we relax the dependence structure between the entries β_i and allow them to be non-identically distributed. See Appendix A.3 in the supplementary materials for more examples and comparisons. We emphasize that this assumption is only needed for modeling the idiosyncratic coefficients β , and we allow for arbitrary fixed-effects θ on the latent factors in our factor-augmented model (2.2).

We split our main results into two sections: the first one shows the efficiency of the ridge regression among the general class of spectral regularized estimators; the second one establishes the uniform consistency of the bias-corrected k -fold cross-validation for selecting the best ridge penalty.

2.1. On the Optimality of Ridge Regression

In this section, we show that the ridge regression is asymptotically optimal among the spectral regularized estimators in capturing idiosyncratic information and it is more robust than principal component regression against factor augmentation.

Consider the singular value decomposition $n^{-1/2}X = \sum_{i=1}^K \lambda_i^{1/2} u_i v_i'$, where λ_i denotes the i th largest eigenvalue (counting multiplicities) and $K \leq \{n-1, p\}$ denotes the rank of X . Note that $\{u_i\}$ are all orthogonal to the n -dimensional unit vector proportional to the all-ones vector, denoted by $\iota_n := n^{-1/2}(1, \dots, 1)' \in \mathbb{R}^n$, due to data centering. To reduce the estimation variance of regression means $\mu = F\theta + E\beta$ in (1.4), we shrink the projections of the (uncentered) target vector $y := (y_1, \dots, y_n)' \in \mathbb{R}^n$ onto the eigenbasis $\{\iota_n\} \cup \{u_i \in \mathbb{R}^n : 1 \leq i \leq K\}$ via a shrinkage function $\delta : (0, \infty) \rightarrow [0, 1]$:

$$\hat{\mu}(\delta) = \iota_n \iota_n' y + \sum_{i=1}^K \delta(\lambda_i) u_i u_i' y = (\hat{m}_{\delta}(x_1), \dots, \hat{m}_{\delta}(x_n))' \quad (2.3)$$

where \hat{m}_{δ} is the estimated regression function given by

$$\hat{m}_{\delta}(x) = \bar{y} + (x - \bar{x})' \hat{\beta}(\delta), \quad (2.4)$$

and $\hat{\beta}(\delta)$ is a penalized least-squares estimator (Tikhonov 1963a, 1963b) given by

$$\hat{\beta}(\delta) = (X'X + nW(\delta))^+ X'Y \quad (2.5)$$

with A^+ denoting the Moore-Penrose inverse of any matrix A and the weight matrix

$$W(\delta) = \sum_{i=1}^K \delta(\lambda_i) v_i v_i', \quad \tilde{\delta}(x) = x \cdot \frac{1 - \delta(x)}{\delta(x)}. \quad (2.6)$$

In particular, choosing $\tilde{\delta}(x) \equiv \gamma$ and $\delta(x) = \frac{x}{x+\gamma}$ for some constant $\gamma > 0$ in (2.5) yields the ridge (Hoerl and Kennard 1970) estimator $\hat{\beta}(\gamma) = (X'X + n\gamma I_p)^{-1} X'Y$. We allow for infinite penalties for dimension reduction purpose: one should interpret $\tilde{\delta}(x) = \infty$ if and only if $\delta(x) = 0$, and then we must set β orthogonal to v_i in estimation. For more discussions we refer to Hastie, Tibshirani, and Friedman (2009, chap. 3). Henceforth we call (2.3) a *spectral regularized estimator* of the regression means.

We define the predictive loss as the reducible mean squared forecast error (see James et al. 2021 for decomposition into reducible and irreducible errors), conditional on a realization of the training set, given by

$$\tilde{L}(\delta) = \mathbb{E} \left[(\mu_{\text{oos}} - \hat{m}_{\delta}(x_{\text{oos}}))^2 \mid \mathcal{S}_n \right],$$

where $x_{\text{oos}} = \Lambda f_{\text{oos}} + \Gamma e_{\text{oos}}$ and $\mu_{\text{oos}} = \phi_0 + f'_{\text{oos}} \theta + e'_{\text{oos}} \beta$ are out-of-sample (OOS) random covariates and the regression mean from the DGPs (2.1) and (2.2) with the same factor model parameters and regression coefficients. The sigma-algebra \mathcal{S}_n is generated by the in-sample variables $\{f_i, e_i, \varepsilon_i : i \leq n\}$ and the random coefficients β .

We use the following paradigm to evaluate the spectral regularized estimators.

Assumption 5 (Evaluation Paradigm). The new variables f_{oos} and e_{oos} are independent of both in-sample regression errors $\{\varepsilon_i : i \leq n\}$ and random coefficients β , given the in-sample random variables $\{f_i, e_i : i \leq n\}$. The conditional covariance matrices of a_{oos} centered around the sample mean given by $\tilde{\Omega}_a = \mathbb{E}[(a_{\text{oos}} - \bar{a})(a_{\text{oos}} - \bar{a})' \mid \mathcal{S}_n]$ have spectral norms bounded in n for both $a \in \{f, e\}$.

This paradigm is flexible enough to include both the training and testing predictive losses used in Figure 2:

- If one draws the pair of new variables f_{oos} and e_{oos} uniformly over the sample $\{f_i, e_i : 1 \leq i \leq n\}$, then we obtain the training predictive loss given by

$$\tilde{L}(\delta) = \frac{1}{n} \sum_{i=1}^n (\mu_i - m_{\delta}(x_i))^2 = \frac{1}{n} \|\mu - \hat{\mu}(\delta)\|^2, \quad (2.7)$$

with the sample covariance matrices $\tilde{\Omega}_a = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})(a_i - \bar{a})'$ for $a \in \{f, e\}$.

- A more typical implementation is to generate f_{oos} and e_{oos} , independent of \mathcal{S}_n , from some population distribution or a random draw on a test set in order to assess the testing predictive loss. Then $\tilde{\Omega}_a = \Sigma_a + (\mu_a - \bar{a})(\mu_a - \bar{a})'$, where μ_a and Σ_a denote the mean and covariance matrix of a_{oos} for $a \in \{f, e\}$.

Our estimated regression function (2.4) uses oracle estimators $\Lambda' \hat{\beta}(\delta)$ and $\Gamma' \hat{\beta}(\delta)$ of θ and β , respectively via the decomposition:

$$\hat{m}_{\delta}(x_{\text{oos}}) = \bar{y} + (f_{\text{oos}} - \bar{f})' \Lambda' \hat{\beta}(\delta) + (e_{\text{oos}} - \bar{e})' \Gamma' \hat{\beta}(\delta).$$

The following theorem relates $\tilde{L}(\delta)$ to the estimation error of linear coefficients and gives further stochastic approximations based on spectral analysis.

Theorem 1. Under Assumptions 1–5, for any sequence of candidate shrinkage function $\delta = \delta_n : (0, \infty) \rightarrow [0, 1]$ may or may not depend on the covariates in the training set,

$$\tilde{L}(\delta) - \|\theta - \Lambda' \hat{\beta}(\delta)\|_{\tilde{\Omega}_f}^2 - \|\beta - \Gamma' \hat{\beta}(\delta)\|_{\tilde{\Omega}_e}^2 \xrightarrow{\mathbb{P}} 0, \quad (2.8)$$

where $\|v\|_{\tilde{\Omega}}^2 = v' \tilde{\Omega} v$ denotes the weighted norm of any vector v . Furthermore,

$$\|\theta - \Lambda' \hat{\beta}(\delta)\|_{\tilde{\Omega}_f}^2 - \theta' D_{\delta} \tilde{\Omega}_f D_{\delta} \theta \xrightarrow{\mathbb{P}} 0, \quad (2.9)$$

$$\|\beta - \Gamma' \hat{\beta}(\delta)\|_{\tilde{\Omega}_e}^2 - \int_0^{\lambda_{r+1}} \left\{ \tau^2 (1 - \delta(x))^2 + \sigma^2 \frac{p}{n} \frac{\delta^2(x)}{x} \right\} d\tilde{F}_p(x) - \tau^2 \tilde{F}_p(0) \xrightarrow{\mathbb{P}} 0. \quad (2.10)$$

where $D_{\delta} = \text{diag}(1 - \delta(\lambda_1), \dots, 1 - \delta(\lambda_r))$ and $\tilde{F}_p(x) = \frac{1}{p} \sum_{i=1}^p v_i' \Gamma' \tilde{\Omega}_e \Gamma' v_i \mathbb{1}(\lambda_i \leq x)$, if there is no tie values in $\{\sigma_{ii} : i = 1, \dots, r\}$ and the penalty function $\tilde{\delta}$ is bounded away from 0. The probability mass $\tilde{F}_p(0) > 0$ when the data matrix X is column-rank deficient with $p > n$. More generally, the result remains true if $\delta(\lambda_i) = \delta(\lambda_j) + o_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0$ for every tie values $\sigma_{ii} = \sigma_{jj}$ with $1 \leq i < j \leq r$.

The estimation errors of fixed and random effects are asymptotically orthogonal in (2.8) because their interaction vanishes through aggregation of the uncorrelated mean-zero random coefficients and regression errors. Via the approximation (2.8) one can interpret different predictive loss functions using different weight matrices: (a) the training errors weight the coefficient estimation errors by the sample covariance matrices; (b) the testing errors weight the coefficient estimation errors by the population covariance matrices (centered around sample means); (c) the L_2 norm of coefficient estimation error $\|\theta - \Lambda' \tilde{\beta}(\delta)\|^2 + \|\beta - \Gamma' \tilde{\beta}(\delta)\|^2$ uses identity weight matrices $\tilde{\Omega}_a = I_p$ for $a \in \{f, e\}$. The asymptotic limit in (2.9) shows the bias of shrinkage estimator of fixed-effects θ , and that in (1) shows the bias-variance tradeoff for random-effects estimation.

Remark 1 (PCR). Plugging in the shrinkage function $\delta(x) = \mathbb{1}[x > \lambda_{r+1}]$ for the oracle PCR estimator yields a limit increasing linearly in $\|\beta\|^2$ as depicted in Figure 1:

$$\begin{aligned} \tilde{L}(\delta) &= 0 \cdot \|\theta\|^2 + \int_0^{\lambda_{r+1}} \left\{ \tau^2(1-0)^2 + \sigma^2 \frac{p}{n} \frac{0^2}{x} \right\} d\tilde{F}_p(x) \\ &\quad + \tau^2 \tilde{F}_p(0) + o_{\mathbb{P}}(1) \\ &= \tau^2 \tilde{F}_p(\lambda_{r+1}) + o_{\mathbb{P}}(1) = \|\beta\|^2 \cdot \frac{1}{p} \text{tr}(\tilde{\Omega}_e) + o_{\mathbb{P}}(1) \end{aligned}$$

where the last equation is due to the facts that $\|\beta\|^2 = \tau^2 + o_{\mathbb{P}}(1)$ and the random slope $\tilde{F}_p(\lambda_{r+1}) = p^{-1} \text{tr}(\tilde{\Omega}_e) + o_{\mathbb{P}}(1)$. In fact, this linear pattern generalizes for any PCR estimator using a finite number of $\hat{r} \geq r$ principal components as adding a few idiosyncratic components does not change its asymptotic behavior in dense models.

Note that the candidate shrinkage function δ is arbitrary on the positive line as long as it is “honest” in the sense that it only depends on the covariates but not the target variable *before* selection. Our limit theorem shows that shrinking the projections of the target vector onto the first r principal components can only increase the predictive loss asymptotically via the squared bias (2.9). We should keep the leading principal components unshrunk asymptotically by choosing a shrinkage function such that $1 - \delta(\lambda_i) \xrightarrow{\mathbb{P}} 0$ for all $i = 1, \dots, r$ and thus the fixed-effects estimator is consistent in the sense that $\|\theta - \Lambda' \tilde{\beta}(\delta)\|^2 \xrightarrow{\mathbb{P}} 0$.

We will discuss how to choose the optimal (ridge) estimator satisfying this condition later on. Therefore, now we focus on this subset of shrinkage functions and study the limiting predictive loss in the following corollary:

Corollary 1. Suppose that $1 - \delta(\lambda_i) \xrightarrow{\mathbb{P}} 0$ for all $i = 1, \dots, r$. Theorem 1 implies that

$$\tilde{L}(\delta) - \int_0^\infty \left\{ \tau^2(1 - \delta(x))^2 + \sigma^2 \frac{p}{n} \frac{\delta^2(x)}{x} \right\} d\tilde{F}_p(x) - \tau^2 \tilde{F}_p(0) \xrightarrow{\mathbb{P}} 0, \quad (2.11)$$

where we integrate the limiting error over the entire domain of \tilde{F}_p for notational convenience although it is asymptotically equivalent to integrate over only the sub-interval $[0, \lambda_{r+1}]$.

If further provided that, with probability one, $\lambda_{\max}(\tilde{\Omega}_e)$ is bounded by some absolute constant and \tilde{F}_p converges vaguely

to a nonrandom limit \tilde{F} , then Corollary 1 remain true with \tilde{F}_p replaced by the nonrandom limit \tilde{F} for every continuous shrinkage function δ . In particular, we have the following special cases.

Corollary 2. Let the signal length τ^2 and error variance σ^2 be constants, and generate the new variables e_{00s} independent of the training data. Suppose that all the entries of e_i and e_{00s} are from an iid array with mean zero and variance σ_e^2 . Then, Corollary 1 implies that for every continuous shrinkage functions δ such that $\lim_{x \rightarrow \infty} \delta(x) = 1$,

$$\tilde{L}(\delta) \xrightarrow{\mathbb{P}} \int_0^\infty \left\{ \tau^2(1 - \delta(x))^2 + \sigma^2 \frac{\delta^2(x)}{x} \right\} dF(x) + \tau^2 F(0), \quad (2.12)$$

where F is the Marčenko and Pastur (1967) law with a point mass $F(0) = \max\{0, 1 - 1/c\}$ at the origin, and the positive density function $F'(x) = \frac{1}{2\pi xy\sigma_e^2} \sqrt{(b-x)(x-a)}$ on $[a, b]$ (and zero density elsewhere) with $a = \sigma_e^2(1 - \sqrt{c})^2$ and $b = \sigma_e^2(1 + \sqrt{c})^2$. More examples satisfying (2.12) with various limits F are available in Appendix A.4 in the supplementary materials.

For generality, we use the stochastic approximation (2.11) according to the random nature of the predictive loss. We shall show that the optimal (ridge) estimator is asymptotically invariant to the random measure F_p (or its limit F if any), and therefore the optimality of ridge regression does not rely on the spectral convergence of large-dimensional covariance matrices.

First, consider the degenerate scenario in that the PCR model is correct with $\tau^2 = 0$, or more generally $\tau^2 \xrightarrow{\mathbb{P}} 0$. One may still use a ridge estimator to minimize the limiting error (2.11) with the shrinkage function $\delta(x) = \frac{x}{x+\gamma}$ such that the ridge penalty $\gamma = \gamma_n$ is diverging at a slower rate of the smallest spiked value $\lambda_r \xrightarrow{\mathbb{P}} \infty$. To our knowledge, this finding first appeared formally in De Mol, Giannone, and Reichlin (2008) who studied the PCR model with a_p diverging at a linear rate of p . Such ridge estimator is a smoothed approximation of the PCR estimator performing an asymptotic selection of principal components in the sense that $\delta(\lambda_i) = \frac{\lambda_i}{\lambda_i + \gamma} \geq \frac{\lambda_r}{\lambda_r + \gamma} \xrightarrow{\mathbb{P}} 1$ as $\gamma/\lambda_r = O_{\mathbb{P}}(\gamma/a_p) \xrightarrow{\mathbb{P}} 0$ for $i = 1, \dots, r$, while $\delta(\lambda_i) = \frac{\lambda_i}{\lambda_i + \gamma} \leq \frac{\lambda_{r+1}}{\lambda_{r+1} + \gamma} \xrightarrow{\mathbb{P}} 0$ for all the non-spiked eigenvalues $\{\lambda_i : i \geq r+1\}$ that are stochastically bounded as $\gamma \xrightarrow{\mathbb{P}} \infty$. In practice, one may select such a ridge penalty coefficient by using cross-validation and we postpone the details to next section. The optimal predictive loss vanishes to zero in probability as we can estimate the fixed-effects consistently and the random-effects are negligible.

A more interesting case is the augmented model with an idiosyncratic signal length $\tau^2 > 0$ bounded away from zero (stochastically), minimizing the asymptotic limit in (2.11) is equivalent to minimizing the integrand

$$\tau^2(1 - \delta(x))^2 + \sigma^2 \frac{p}{n} \frac{\delta^2(x)}{x} \text{ for each } x \in (0, \infty).$$

That is, for each $x \in (0, \infty)$, we choose $\delta(x)$ as the minimum point of the convex quadratic function

$$\Delta_x(\delta) := \tau^2(1 - \delta)^2 + \sigma^2 \frac{p}{n} \frac{\delta^2}{x}, \quad \delta \in [0, 1].$$

By solving the first-order condition $\Delta'_x(\delta) = 0$, it is elementary to show that the optimal shrinkage function is given by

$$\delta^*(x) = \frac{x}{x + \gamma^*}, \quad \gamma^* = \frac{p}{n} \frac{\sigma^2}{\tau^2}, \quad \Rightarrow \quad \tilde{\delta}^*(x) \equiv \gamma^* \quad (2.13)$$

which corresponds to a ridge estimator with the penalty coefficient γ^* , that is, with $W(\tilde{\delta}^*) = \gamma^* I_p$ in (2.5). One can interpret γ^* as the noise-to-signal ratio σ^2/τ^2 scaled by the aspect ratio p/n . Note that the optimal predictive loss is not vanishing in this case as the high-dimensional random-effects β cannot be estimated consistently. Since the ridge coefficient γ^* is now (stochastically) bounded, the optimal shrinkage function (2.13) naturally satisfies the conditions that $1 - \delta(\lambda_i) \xrightarrow{\mathbb{P}} 0$ as the spiked eigenvalues $\lambda_i \xrightarrow{\mathbb{P}} \infty$ for $1 \leq i \leq r$, leading to a consistent estimation of the fixed-effects θ .

Remarkably, the optimal shrinkage function (2.13) is the same for all our predictive loss functions sharing the approximate form (2.11). We summarize the optimality of ridge regression into the following corollary:

Corollary 3 (Optimality of ridge regression). Consider any predictive loss \tilde{L} admitting the asymptotic approximation in Theorem 1. For an arbitrarily small $\zeta > 0$ and any candidate shrinkage function $\delta = \delta_n$ from Theorem 1,

$$\mathbb{P}(\tilde{L}(\delta) > \tilde{L}(\delta^*) - \zeta) \rightarrow 1,$$

where $\delta^*(x) = \frac{x}{x + \gamma^*}$ denotes the shrinkage function for the optimal ridge estimator with the penalty coefficient $\gamma^* = \frac{p}{n} \frac{\sigma^2}{\tau^2}$ when τ^2 is bounded away from 0 with probability approaching 1, or otherwise an arbitrary diverging penalty $\gamma^* \xrightarrow{\mathbb{P}} \infty$ but $\gamma^* = o_{\mathbb{P}}(a_p)$ if $\tau^2 = o_{\mathbb{P}}(1)$. In the latter case, $\tilde{L}(\delta^*) \xrightarrow{\mathbb{P}} 0$ and we may take $\gamma^* = \exp\left(\sum_{i=1}^{r_{\max}+1} w_i \log \lambda_i\right)$ where $w_i \in (0, 1)$ are arbitrary fixed weights such that $\sum_{i=1}^{r_{\max}+1} w_i = 1$ for any pre-specified bound $r_{\max} \geq r$. However, in general, only the fixed-effects are consistently estimated in the sense that $\|\theta - \Lambda' \hat{\beta}(\tilde{\delta})\|^2 \xrightarrow{\mathbb{P}} 0$ and the optimal ridge estimator trades off the estimation bias and variance of random-effects.

Remark 2. Corollary 3 generalizes for heteroscedastic errors with different variance $\sigma_i^2 = \text{var}(\varepsilon_i)$ under more strengthened conditions on the covariates $\{x_i\}$, where one can relax expression (2.13) by replacing σ^2 with the average variance $\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i^2$. Due to the space limit, we postpone the details to Appendix A.8 in the supplementary materials.

2.2. Uniform Consistency of Bias-Corrected k -Fold Cross-Validation

This section establishes the uniform consistency of the k -fold cross-validation (CV) algorithm for selecting the best ridge estimators, subject to a straightforward bias correction. Our asymptotic limits of the cross-validated mean squared error is *uniform* for all positive ridge penalties bounded away from zero, including those diverging to infinity. Throughout we use a fixed number of folds $k \geq 2$. Generally speaking, there are both computational and statistical advantages to use k -fold CV rather

than the leave-one-out CV; see, for example, Section 5.1.3 of James et al. (2021). On the other hand, using a small number of folds may introduce a bias in the selection if the optimal parameter depends on the sample size of the training set.

To fix ideas, the procedure starts from partitioning the entire index sets $\{1, \dots, n\}$ into k mutually exclusive subsets $\mathcal{I}_1, \dots, \mathcal{I}_k$. For generality, we treat the partition $\{\mathcal{I}_j : j = 1, \dots, k\}$ as fixed for any given n . When the folds are randomly created, our setup is equivalent to conditioning on a particular realization. One simple choice is to divide \mathcal{I} into blocks, using consecutive indexes in each fold. In general, it is unnecessary to use blocks, and we may assign indexes randomly to the folds. Following the common practice, we generate folds of approximately equal size such that $n_j/n \rightarrow 1/k$ where $n_j := |\mathcal{I}_j|$ denotes the cardinality of the index subset \mathcal{I}_j .

Let $\tilde{y}_i = y_i - \bar{y}$ and $\tilde{x}_i = x_i - \bar{x}$ be the demeaned observations, \bar{y} and \bar{x} are the sample means using all data. For any fold \mathcal{I}_j treated as a validation set, we fit a ridge estimator $\hat{\beta}_{-j}(\gamma)$ on the remaining $k-1$ folds by minimizing the L_2 penalized error given by

$$\hat{\beta}_{-j}(\gamma) = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n-j} \sum_{i \notin \mathcal{I}_j} (\tilde{y}_i - \tilde{x}_i' \beta)^2 + \gamma \|\beta\|^2 \right\}, \quad (2.14)$$

where $n-j = n - n_j$ denotes the number of training observations, and $\gamma > 0$ is a candidate ridge penalty coefficient. Recall the demeaned design matrix $X = [\tilde{x}_1, \dots, \tilde{x}_n]'$ and demeaned target vector $Y = [\tilde{y}_1, \dots, \tilde{y}_n]'$. Let $D_{-j} = \text{diag}(\mathbb{1}[1 \in \mathcal{I}_{-j}], \dots, \mathbb{1}[n \in \mathcal{I}_{-j}])$ be the $n \times n$ diagonal matrices that selects the observations on the training set $\mathcal{I}_{-j} := \mathcal{I}_j^c$. Solving the quadratic optimization problem (2.14) yields a close-form solution given by

$$\hat{\beta}_{-j}(\gamma) = (X' D_{-j} X + n_{-j} \gamma I_p)^{-1} X' D_{-j} Y. \quad (2.15)$$

The mean squared error on the observations in the held-out fold \mathcal{I}_j is given by

$$\text{MSE}_j(\gamma) = \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} (\tilde{y}_i - \tilde{x}_i' \hat{\beta}_{-j}(\gamma))^2.$$

We repeat the procedure for $j = 1, \dots, k$, and average these errors to compute the k -fold cross-validation MSE given by

$$\text{CV}^{(k)}(\gamma) = \sum_{j=1}^k \frac{n_j}{n} \text{MSE}_j(\gamma) = \frac{1}{n} \sum_{j=1}^k \sum_{i \in \mathcal{I}_j} (\tilde{y}_i - \tilde{x}_i' \hat{\beta}_{-j}(\gamma))^2.$$

Then we choose the value of γ over a sufficiently fine grid that minimizes $\text{CV}^{(k)}(\gamma)$. Note that we do not need to estimate the factor scores nor to determine the number of factors with ridge regression.

To apply factor analysis to the training subsamples rather than the full sample, we assume one additional regularity condition:

Assumption 6. Let $\Omega_{f,-j} := n_{-j}^{-1} \sum_{i \notin \mathcal{I}_j} (f_i - \bar{f})(f_i - \bar{f})'$ denote the training covariance matrix latent factor scores. For every $j = 1, \dots, k$, the eigenvalues of the $r \times r$ matrix $\Omega_{f,-j}^{1/2} \Sigma_{\Lambda} \Omega_{f,-j}^{1/2}$ converge to some limits $\sigma_{11,-j} \geq \dots \geq \sigma_{rr,-j} > 0$.

The condition allows us to reparameterize the factor scores and loadings on the training set to maintain similar orthogonal structures in [Assumption 2](#). More specifically, consider the spectral decomposition $\Omega_{f,-j}^{1/2} \Sigma_{\Lambda} \Omega_{f,-j}^{1/2} = Q_{-j} \Sigma_{-j} Q_{-j}'$ where Σ_{-j} is a diagonal matrix and Q_{-j} is an orthogonal matrix. Define the demeaned matrix of the idiosyncratic information $E := [e_1 - \bar{e}, \dots, e_n - \bar{e}]'$. The training design matrix obeys the approximate factor model

$$D_{-j}X = D_{-j}F\Lambda' + D_{-j}E\Gamma' = F_{-j}\Lambda_{-j}' + D_{-j}E\Gamma'$$

where $F_{-j} := D_{-j}F\Omega_{f,-j}^{-1/2}Q_{-j}$ and $\Lambda_{-j} := \Lambda\Omega_{f,-j}^{1/2}Q_{-j}$ satisfy the orthogonal conditions:

$$\begin{aligned} F_{-j}'F_{-j}/n_{-j} &= I_r, \quad \text{and} \\ \Lambda_{-j}'\Lambda_{-j}/a_p &= \Sigma_{-j} \rightarrow \text{diag}(\sigma_{11,-j}, \dots, \sigma_{rr,-j}). \end{aligned}$$

[Assumption 6](#) holds under a stronger condition that $n_{-j}^{-1} \sum_{t \notin \mathcal{I}_j} (f_i - \bar{f})(f_i - \bar{f})' \rightarrow I_r$ for all $j = 1, \dots, k$, in which case the limiting eigenvalue $\sigma_{ii,-j} = \sigma_{ii}$. If f_i are iid random vectors with an identity covariance matrix, the conditions holds almost surely by strong law of large numbers. When $\{f_i\}$ is a (multivariate) time series and the folds are blocks, that is, each time index subset \mathcal{I}_j corresponds to a continuous time period, the condition follows from ergodicity:

$$\begin{aligned} \frac{1}{n_{-j}} \sum_{i \notin \mathcal{I}_j} (f_i - \bar{f})(f_i - \bar{f})' &= \frac{n}{n_{-j}} \frac{1}{n} \sum_{i=1}^n (f_i - \bar{f})(f_i - \bar{f})' \\ &\quad - \frac{n_j}{n_{-j}} \frac{1}{n_j} \sum_{t \in \mathcal{I}_j} (f_t - \bar{f})(f_t - \bar{f})' \\ &= \frac{n}{n_{-j}} I_r - \frac{n_j}{n_{-j}} I_r + o_{\mathbb{P}}(1) \\ &= I_r + o_{\mathbb{P}}(1) \end{aligned}$$

for any stationary ergodic process $\{f_i : i \in \mathbb{Z}\}$ with zero mean and identity covariance matrix, where \mathbb{Z} denotes the set of integers; see Blum and Eisenberg (1975) for more discussions. [Assumption 6](#) then holds with probability approaching 1.

Furthermore, when the folds are randomly generated, the results remain true if the projection process $\{w/f_i : i \in \mathbb{Z}\}$ is stationary and strong mixing for every unit vectors $w \in \mathbb{R}^r$ by combining the theorem in Blum and Hanson (1960) and Cramér–Wold device. See Bradley (1986) and Bradley (2005) for the various types of sufficient conditions for strong mixing.

Theorem 2. Suppose that [Assumptions 1–4](#) and [6](#) hold. Uniformly for all positive ridge penalties γ bounded away from 0,

$$\begin{aligned} \text{CV}^{(k)}(\gamma) - \sigma^2 - \left(\frac{\gamma}{a_p + \gamma} \right)^2 \cdot \theta' W_{\gamma} \theta \\ - \int_0^{\infty} \left\{ \tau^2 \frac{\gamma^2}{(x + \gamma)^2} + \sigma^2 \frac{p}{n_{\text{eff}}} \frac{x}{(x + \gamma)^2} \right\} dF_p^{(k)}(x) \\ - \tau^2 F_p^{(k)}(0) \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

where $F_p^{(k)}$ is some (improper) empirical distribution depending only on the covariates $\{x_i\}$, the effective training sample size is given by $n_{\text{eff}} = \frac{1}{k} \sum_{j=1}^k n_{-j}$ provided that n_j/n converges to the

same limit for all j , and $W_{\gamma} \in \mathbb{R}^{r \times r}$ is a random weight matrix being stochastically bounded and positive-definite with probability tending to 1 when $r > 0$. The closed-form expressions of $F_p^{(k)}$ and W_{γ} are available in Section E.3 of the supplementary materials.

The first term σ^2 is irreducible but irrelevant to the choice of γ . The second term $\left(\frac{\gamma}{a_p + \gamma} \right)^2 \cdot \theta' W_{\gamma} \theta$ represents the cross-validated estimation error of the fixed-effects at the same (stochastic) order of $\left(\frac{\gamma}{a_p + \gamma} \right)^2 \|\theta\|^2$. The rest represents the cross-validated error for the random-effects where we use $n_{\text{eff}} \approx \frac{k-1}{k}n$ due to the train-test split.

Let $\hat{\gamma}_{\text{cv}}$ denote the optimal ridge penalty selected by the regular k -fold CV. When $\tau^2 \xrightarrow{\mathbb{P}} 0$, the k -fold CV chooses some divergent penalty $\hat{\gamma}_{\text{cv}} \xrightarrow{\mathbb{P}} \infty$ to shrink the random-effects estimator toward zero as much as possible but $\hat{\gamma}_{\text{cv}} = o_{\mathbb{P}}(a_p)$ (unless $\|\theta\|^2 = 0$) such that the fixed-effect estimation errors vanish as well. Note that any divergent penalty sequence proportional to $\hat{\gamma}_{\text{cv}}$ is also asymptotically optimal. When τ^2 is bounded away from 0, by the same arguments above [Corollary 3](#), one should again minimize the integrand $\tau^2 \frac{\gamma^2}{(x + \gamma)^2} + \sigma^2 \frac{p}{n_{\text{eff}}} \frac{x}{(x + \gamma)^2}$ by using the optimal penalty $\gamma_{\text{cv}}^* = \frac{p}{n_{\text{eff}}} \frac{\sigma^2}{\tau^2} = \frac{n}{n_{\text{eff}}} \gamma^* \approx \frac{k}{k-1} \gamma^*$ for all $x > 0$ where $\gamma^* = \frac{p}{n} \frac{\sigma^2}{\tau^2}$ denotes the optimal ridge penalty in [Corollary 3](#). To conclude, the bias-corrected estimator of the optimal ridge penalty given by

$$\hat{\gamma} := \frac{k-1}{k} \hat{\gamma}_{\text{cv}}$$

satisfies the asymptotic optimality criterion in [Corollary 3](#).

3. Extension of Main Results

This section discusses two extensions of the main results for handling dependent data. Since our applications are often (although not necessarily) based on time series data, we shall index the observation by $t = 1, \dots, n$ instead of i to emphasize the concept of time.

Our first extension is to relax the independence [Assumption 3](#) on regression errors and only require them to be martingale differences.

Assumption 7 (MD Errors). The regression errors $\varepsilon_t = \sigma \eta_t$, where $\{\eta_t, \mathcal{F}_{nt} : t \in \mathbb{Z}\}$ is a martingale difference sequence for each n such that $\mathbb{E}_{t-1} \eta_t = 0$ and $\mathbb{E}_0 \eta_t^2 = 1$, and $\mathbb{E}_j[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{nj}]$ denotes the conditional expectation given the sigma-algebra \mathcal{F}_{nj} generated by $\{\eta_s : s \leq j\}$, $\{f_t\}$ and $\{e_t\}$. The variance $\sigma^2 = \sigma_n^2 = O_{\mathbb{P}}(1)$ is bounded away from zero almost surely, and may or may not depend on the covariates $\{x_t\}$. For some $\iota > 0$, $\sum_{t=1}^n \mathbb{E} |\eta_t|^{2+2\iota} = O(n)$.

This MD assumption is required for the theory of k -fold cross-validation with uncorrelated (but dependent) errors using the usual random train-test splits. Note that the conditional variance $\mathbb{E}_{t-1} \eta_t^2$ can be stochastic, meaning that we allow for conditionally heteroscedastic errors with stochastic volatilities such as the GARCH-type error processes (see, e.g., Davis and Mikosch 2009

for an excellent survey). The last part of the assumption, allowing for heavy-tailed distributions, is needed for the uniform integrability of $\{\eta_t^2\}$. In general $\{\eta_t\}$ are even not necessarily to be identically distributed, but if they are then the last part simplifies to be $\mathbb{E}|\eta_t|^{2+2t} = O(1)$.

Theorem 3. All theorems and corollaries in Section 2 remain true by replacing the IID Assumption 3 with the MD Assumption 7, provided the temporal dependence conditions on $\{\eta_t\}$ in Appendix A.5 in the supplementary materials.

The most shocking finding is that, at least for ridge regression, we can in fact split the time series *arbitrarily* without worrying about the so-called “time leakage” issues in machine learning. In other words, the time order between the training and validation data is *unimportant* and the usual random train-test splits are allowed.

Our second extension is the factor-augmented autoregressive model given by

$$x_t = \Lambda f_t + \Gamma e_t \quad (3.1)$$

$$y_t = \phi_0 + \sum_{\ell=1}^q \phi_\ell y_{t-\ell} + f_t' \theta + e_t' \beta + \varepsilon_t, \quad (3.2)$$

where ε_t are martingale differences as in Assumption 7. The first equation is the same as (2.1), while the second equation adds the lagged target variables into (2.2). In forecasting applications, x_t are often lagged variables that are observable before time t . Let $z_t = (y_{t-1}, \dots, y_{t-q})'$ collect the lagged target values, and their coefficient vector be $\phi = (\phi_1, \dots, \phi_q)'$. We can rewrite the model into a general form given by

$$y_t = \phi_0 + z_t' \phi + f_t' \theta + e_t' \beta + \varepsilon_t. \quad (3.3)$$

For simplicity, we assume that the order of autoregression q is finite and known. Unless specified otherwise, we assume that the autoregressive model satisfies the standard stability condition (see, e.g., Chapter 6 of Hayashi 2000, p. 373):

Assumption 8. All the roots of the q th degree polynomial equation $1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_q z^q = 0$ are greater than 1 in absolute value, when $q > 0$. The target variables has finite second moment, that is, $\mathbb{E}[y_t^2]$ is bounded uniformly over t .

We extend the ridge regression to allow for the nuisance variables z_t and minimize the penalized mean squared errors of the parameters $(\alpha, \phi', \beta')'$ given by

$$\frac{1}{n} \sum_{t=1}^n (y_t - \alpha - z_t' \phi - x_t' \beta)^2 + \gamma \|\beta\|^2,$$

where $\gamma > 0$ is some candidate ridge penalty that may depend on $\{x_t\}$. When $q = 0$, this reduces to the standard ridge problem. Let $Z = [z_1 - \bar{z}, \dots, z_n - \bar{z}]'$ be the demeaned design matrix of autoregressors with $\bar{z} = n^{-1} \sum_{t=1}^n z_t$, and define the projection matrix onto its column space by $P_Z = Z(Z'Z)^{-1}Z'$. Solving the quadratic optimization problem yields a three-steps estimator

$$\begin{cases} \hat{\beta}(\gamma) = (X'(I - P_Z)X + n\gamma I_p)^{-1} X'(I - P_Z)Y, \\ \hat{\phi}(\gamma) = (Z'Z)^{-1} Z'(Y - X\hat{\beta}(\gamma)), \\ \hat{\alpha}(\gamma) = \bar{y} - \bar{z}'\hat{\phi}(\gamma) - \bar{x}'\hat{\beta}(\gamma). \end{cases} \quad (3.4)$$

We are interested in minimizing the predictive loss given by

$$L(\gamma) = \mathbb{E} \left[(\mu_{n+h} - \hat{m}_\gamma(z_{n+h}, x_{n+h}))^2 \mid \mathcal{S}_n \right],$$

for some given forecast horizon $h \geq 1$. (3.5)

The sigma-algebra \mathcal{S}_n is generated by the past variables $\{f_t, e_t, \varepsilon_t : t \leq n\}$ and the random coefficients β . An inevitable challenge here is that the target variables in training and test samples must be dependent due to autoregression (3.2). Nevertheless, we can generate new covariates in the future according to a similar paradigm in Section 2:

Assumption 9 (Testing Paradigm for AR Models). The testing paradigm in Assumption 5 holds with $f_{\text{oos}} = f_{n+h}$ and $e_{\text{oos}} = e_{n+h}$ for all horizons $h \in \{1, 2, \dots\}$. For all $1 \leq \ell \leq h-1$, $\mathbb{E}[(e_{n+h-\ell} - \bar{e})'(e_{n+h} - \bar{e}) \mid \mathcal{S}_n] = o_{\mathbb{P}}(p)$.

Theorem 4. Under the conditions of Theorem 3, Assumptions 8–9, and the conditions in Appendix A.6 in the supplementary materials, the optimal ridge penalty remains the same as in Corollary 3 for the testing predictive loss (3.5) at all forecast horizons $h \geq 1$.

Nevertheless, the classic results for the standard k -fold cross-validation do not extend to the autoregressive models with large-dimensional covariates. We notice that there are some recent findings in Bergmeir, Hyndman, and Koo (2018) for simple autoregressive models with no exogenous variable, but their results neither do not apply in the presence of large-dimensional covariates. Hence, we suggest to remove the autoregressive components before the cross-validation using the following algorithm:

Step 1 Let $\hat{\phi}$ be some preliminary estimator of the autoregressive coefficients ϕ , and construct the autoregressive residuals $r_t = (y_t - \bar{y}) - (z_t - \bar{z})'\hat{\phi}$. Apply the k -fold cross-validation to select the optimal ridge penalty parameter by regressing the autoregressive residuals r_t on the demeaned covariates $x_t - \bar{x}$ using the standard ridge estimator.

Step 2 Calculate the bias-corrected ridge penalty $\hat{\gamma} = \hat{\gamma}_{\text{cv}}(1 - \frac{1}{k})$, where $\hat{\gamma}_{\text{cv}}$ is the optimal ridge factor selected in Step 1.

Step 3 Calculate the estimators $\hat{\beta}$, $\hat{\phi}$, and $\hat{\alpha}$ in (3.4) using the penalty $\hat{\gamma}$ from Step 2.

Theorem 5. Under the conditions of Theorem 4, Theorem 2 remains true provided that the preliminary estimator $\hat{\phi}$ is consistent toward ϕ .

In the rest of this section, we discuss how to obtain a consistent estimator $\hat{\phi}$ under our mixed-effects models. We show in Appendix A.7 of the supplementary materials that the naive least-squares estimator using the autoregressors alone (i.e., neglecting the covariates) is consistent if the latent factors are asymptotically uncorrelated over time. To avoid this restrictive condition, we suggest to use the estimator $\hat{\phi}$ from a preliminary ridge regression (3.4) instead according to the following theorem. We call this approach “double” ridge regression because we need to apply ridge regression twice, once in Step 1 and one more time in Step 3 using different penalties.

Theorem 6 (Double ridge regression). Suppose that the conditions of Theorem 4 hold. The ridge estimator $\hat{\phi}$ defined in (3.4) is consistent toward ϕ if $\gamma = o_{\mathbb{P}}(a_p)$ and γ is bounded away from 0 with probability approaching 1, where $\gamma = \gamma_n$ may depend on the covariates $\{x_t\}$. When $\sum_{i=1}^r \theta_i^2 = 0$, the result remains true if $\gamma = \gamma_n$ diverges at the same or even higher rate of a_p . In particular, the following choices all satisfy the consistency condition:

1. Any constant ridge penalty $\gamma > 0$ not depending on n ;
2. Any diverging ridge penalty of the form $\gamma = \exp\left(\sum_{i=1}^{r_{\max}+1} w_i \log \lambda_i\right)$, where r_{\max} is some prespecified upper bound of the true number of factors r and $w_i \geq 0$ are arbitrary fixed constant weights such that $\sum_{i=1}^{r_{\max}+1} w_i = 1$ and $w_{r_{\max}+1} > 0$.

An interesting implication of this proposition is that, even with a possibly inconsistent estimator of $\hat{\phi}$ in Step 1 (e.g., the least-squares estimator omitting covariates), the final estimator $\hat{\phi}$ in Step 3 can recover consistency as long as the ridge penalty from Step 2 is not too large in the sense of Theorem 6. Unless specified otherwise, we use the ridge penalty $\gamma = \lambda_{r_{\max}+1}$ for our preliminary ridge regression in numerical analysis.

4. Monte Carlo Simulations

In this section, we compare the predictive performance of the ridge estimators and that of their competitors under factor augmented models using Monte Carlo experiments. We fix the number of predictors $p = 120$ but vary the sample size $n \in \{60, 120, 240\}$, yielding concentration ratios $p/n \in \{2, 1, 0.5\}$. These dimensions are calibrated from our empirical analysis where $n \approx p = 120$. The results for a fixed sample size $n = 120$ but a much larger number of predictors $p \in \{480, 840, 1200\}$ are available in Appendix C.3 in the supplementary materials.

We calibrate the data generating process of the latent factors $\{f_t\}$ from a vector autoregressive model (VAR) with four lags and the factor loading matrices Λ from a multivariate Gaussian distribution using the estimated scores and loadings of $r = 7$ factors using the FRED-MD database in Section 5. The details of the calibration procedure is available in Appendix B.1 of the supplementary materials. We then generate exogenous predictors from the factor model give by

$$x_t = \Lambda f_t + \Gamma e_t = \Lambda f_t + \Gamma \Sigma^{1/2} \varsigma_t,$$

where the entries $\varsigma_t = (\varsigma_{t,1}, \dots, \varsigma_{t,p})'$ are independently generated from the standard normal distribution, Σ is a diagonal covariance matrix from Bai and Silverstein (1998) (with 20% of its eigenvalues equal to λ , 40% equal to 3λ , and 40% equal to 10λ) such that $\text{tr}(\Sigma) = \text{tr}(\Lambda' \Lambda)$, the idiosyncratic matrix Γ is generated uniformly from the class of orthogonal matrices. The factor scores f_t are unobservable to the statistician, who always approximate the factor scores as \sqrt{n} times the leading eigenvectors of XX' when needed. We then generate the target variable from the factor augmented autoregressive model of order $q = 3$ given by

$$y_t = 0.29y_{t-1} + 0.07y_{t-2} + 0.11y_{t-3} + f_t' \theta + e_t' \beta + \varepsilon_t,$$

where we set a zero intercept $\phi_0 = 0$ without loss of generality, and the autoregressive coefficients $\phi = (0.29, 0.07, 0.11)'$ are calibrated from the monthly growth in U.S. industrial production index. In the supplementary materials we repeat the analysis for the case with no autoregressor (i.e., $q = 0$), and the conclusions remain qualitatively the same. The entries of factor coefficients $\theta = (\theta_1, \dots, \theta_7)'$ are generated from the uniform distribution on $(0, 1)$. All these coefficients are unknown to the statistician who always demeans the predictors in each sample and estimate all the coefficients. We consider two different data generating processes of the regression errors:

1. independent errors ε_i from the standardized student t_5 distribution;
2. GARCH(1,1) errors $\varepsilon_t = \sqrt{h_t} v_t$, $h_t = 1 - a - b + a\varepsilon_{t-1}^2 + b\varepsilon_{t-1}$ where $a = 0.1$, $b = 0.8$.

Without loss of generality, we maintain a long run variance $\sigma^2 = 1$ in both cases. The errors are independent of the covariates. The results are similar for these two types of errors. To save space, we only report the results for GARCH errors in the main document and postpone the results for independent t_5 errors to Appendix C in the supplementary materials.

To compare sparse and dense models in terms of idiosyncratic information, we generate the entries of the idiosyncratic regression coefficients $\beta = (\beta_1, \dots, \beta_p)'$ independently from a “spike-and-slab” distribution (Mitchell and Beauchamp 1988) such that

$$\beta_i \sim \begin{cases} \mathcal{N}(0, 5/p) & \text{with probability } \rho, \\ 0 & \text{with probability } 1 - \rho, \end{cases}$$

and the signal strength $\tau^2 = p\mathbb{E}\beta_i^2 = 5\rho \in \{0, 0.2, 0.5, 1, 2, 5\}$ controls the dense level; the larger τ^2 , the less zero entries in β in probability.

Table 1 reports the median, over 5000 replications, of the testing errors $(\mu_{n+h} - \hat{\mu}_{n+h})^2$ at the horizon $h = 1$ outside brackets and training errors $\frac{1}{n} \sum_{t=1}^n (\mu_t - \hat{\mu}_t)^2$ inside brackets for the following seven estimators:

1. PCR: Principal component regression using true number of factors $r = 7$.
2. KS2011: Factor-adjusted LASSO estimator proposed by Kneip and Sarda (2011) with L_1 penalty parameter selected by k -fold cross-validation;
3. PCR-CV: PCR using the number of factors selected by k -fold cross-validation.
4. SRR: Single ridge regression with the ridge penalty parameter selected by the bias-corrected k -fold cross-validation procedure proposed in the last section, using the least-squares estimator of the autoregressive coefficients in Step 1.
5. FaSRR: Factor-adjusted ridge regression that shrinks the eigenvalues as SRR except for the largest r eigenvalues.
6. DRR: Double ridge regression with the ridge penalty parameter selected by the bias-corrected k -fold cross-validation procedure, using the preliminary ridge regression of the autoregressive coefficients in Step 1 given in Theorem 6.
7. FaDRR: Factor-adjusted ridge regression that shrinks the eigenvalues as DRR except for the largest r eigenvalues.

Table 1. The testing errors (and training errors) over 5000 replications for GARCH errors.

$\frac{\tau^2}{\sigma^2}$	PCR	KS2011	PCR-CV	SRR	FaSRR	DRR	FaDRR
$p = 120, n = 60$							
0.0	0.24 (0.34)	0.61 (0.53)	0.24 (0.36)	0.27 (0.34)	0.24 (0.32)	0.27 (0.35)	0.23 (0.32)
0.2	0.98 (1.52)	3.53 (6.92)	1.00 (1.48)	0.83 (0.69)	0.89 (1.03)	0.81 (0.67)	0.87 (0.95)
0.5	2.58 (3.94)	9.24 (14.15)	2.63 (3.76)	1.81 (0.92)	2.06 (1.88)	1.82 (0.88)	2.03 (1.66)
1.0	5.68 (8.04)	19.70 (26.77)	5.56 (7.73)	3.34 (1.07)	4.00 (3.07)	3.32 (1.01)	3.84 (2.44)
2.0	11.61 (16.21)	37.93 (52.09)	11.10 (15.58)	6.64 (1.25)	7.95 (4.85)	6.65 (1.12)	7.72 (3.30)
5.0	30.79 (40.98)	98.54 (126.34)	28.99 (40.16)	15.53 (1.51)	18.42 (9.18)	14.71 (1.24)	17.84 (4.71)
$p = 120, n = 120$							
0.0	0.15 (0.25)	0.23 (0.33)	0.15 (0.27)	0.18 (0.27)	0.14 (0.24)	0.18 (0.27)	0.14 (0.24)
0.2	0.83 (1.65)	2.53 (5.41)	0.84 (1.56)	0.52 (0.55)	0.66 (1.08)	0.52 (0.54)	0.64 (1.03)
0.5	2.26 (4.58)	9.07 (18.79)	2.20 (4.28)	0.94 (0.72)	1.48 (1.85)	0.94 (0.71)	1.42 (1.73)
1.0	4.73 (9.23)	21.61 (38.15)	4.54 (8.67)	1.62 (0.84)	2.47 (2.59)	1.56 (0.82)	2.37 (2.33)
2.0	10.16 (19.12)	44.39 (80.74)	9.93 (17.86)	2.68 (0.99)	4.22 (3.66)	2.60 (0.94)	4.05 (3.10)
5.0	28.16 (47.50)	109.90 (200.60)	26.31 (44.92)	5.20 (1.29)	8.74 (6.17)	5.10 (1.17)	8.05 (4.71)
$p = 120, n = 240$							
0.0	0.10 (0.20)	0.14 (0.25)	0.10 (0.21)	0.13 (0.22)	0.10 (0.20)	0.13 (0.21)	0.10 (0.19)
0.2	0.75 (1.68)	2.54 (6.54)	0.71 (1.60)	0.30 (0.43)	0.48 (0.96)	0.30 (0.42)	0.48 (0.92)
0.5	2.20 (4.73)	8.82 (20.04)	2.15 (4.40)	0.48 (0.54)	0.98 (1.47)	0.48 (0.53)	0.93 (1.35)
1.0	4.88 (9.89)	21.13 (44.01)	4.73 (9.15)	0.67 (0.65)	1.39 (1.86)	0.67 (0.64)	1.31 (1.69)
2.0	10.14 (20.21)	44.77 (90.67)	9.77 (18.89)	0.96 (0.83)	2.03 (2.35)	0.96 (0.81)	1.87 (2.09)
5.0	27.59 (51.19)	123.21 (231.56)	26.03 (47.50)	1.84 (1.28)	3.82 (3.73)	1.79 (1.25)	3.34 (3.11)

NOTE: The smaller values for each row are highlighted in bold. The results for independent t_5 errors are similar and available in the supplementary materials.

The estimators SRR and DRR are what we propose in the last section, except they use different estimators of the autoregressive coefficients. The SRR estimator is slightly worse than DRR overall due to a larger estimation error of the autoregressive residuals in the cross-validation algorithm. The other estimators PCR, KS2011, FaSRR, and FaDRR are oracle in the sense that they use the true number of factors $r = 7$. They can be feasible in practice when the number of factors can be consistently selected; see, for example, Bai and Ng (2002). In contrast, the SRR and DRR estimators do not require estimating the number of factors but use all raw covariates. More implementation details of these algorithms are available in Appendix B.2 of the supplementary materials.

When the idiosyncratic signal $\tau^2 = 0$, the PCR and factor-adjusted Ridge estimators usually perform the best. Although the SRR and DRR estimator suffers from more estimation bias of fixed-effects due to the shrinkage, they still show good performance and remain robust against the absence of idiosyncratic signals. The LASSO estimator KS2011 is the worst and suffers from variable selection errors in dense models especially when $p > n$.

On the other hand, when $\tau^2 > 0$, our ridge estimators SRR and DRR consistently outperform all sparse learning methods PCR, KS2011 and PCR-CV in terms of both testing and training errors according to our asymptotic theory. The advantages of the ridge estimator becomes more significant as the signal length τ^2 grows.

It is noteworthy that direct factor adjustments cannot improve the ridge regression when $\tau^2 > 0$. FaSRR and FaDRR violate the full-rank condition (Appendix A.6) on the residual factor matrix by removing all the in-sample factor information asymptotically. As a result, the residual ridge regression fails to shrink the divergent loading matrix Λ (via Lemma A7 in the supplementary materials), leading to an inflation of fixed-effects estimation errors.

To conclude, the simulation results confirm that our double ridge regression with bias-corrected cross-validation is a robust

forecasting method for factor-augmented models with potential idiosyncratic information and nuisance variables. It delivers the best performance and nontrivial improvements in dense models with large τ^2 consistently, while remaining robust against sparse idiosyncratic information with small τ^2 .

5. A Real-life Example

One empirical application of dense learning is to forecast the monthly growth rate of the U.S. industrial production index using a large set of macroeconomic variables from the FRED-MD database publicly available at:

<https://research.stlouisfed.org/econ/mccracken/fred-databases/>.

This monthly database has similar predictive content as that in Stock and Watson (2002), and it is regularly updated through the Federal Reserve Economic Data. The industrial production index is an important indicator of macroeconomic activity. We calculate its monthly growth rate in percentage by

$$y_t = \log(IP_t/IP_{t-1}) \times 100$$

where IP_t denotes the U.S. industrial production index for month t . Our sample size span from January, 1959 to August, 2021 and all the data are available on a monthly basis. We transform the nonstationary economic variables into stationary forms and remove the data outliers using the MATLAB codes provided on the website mentioned above; see also McCracken and Ng (2016) for details.

In every month, we forecast the next-month growth ahead using $q = 0, 1, \dots, 6$ autoregressor(s) and (the principal components of) all the other covariates in the database. We estimate the factors and regression coefficients in rolling windows of a sample size $n = 120$, which equals to a time span of ten years. In each window we use the lagged variables with no missing values for training and forecasting. The covariates are normalized within each window, and we transform the latest observation accordingly.

Table 2. Out-of-sample relative mean squared forecast errors for the growth rate of U.S. industrial production index.

q	AR	PCR	KS2011	PCR-CV	SRR	DRR
0	1.0000	0.8091	0.8281	0.8269	0.8218	0.8218
1	0.9140	0.8216	0.8253	0.8159	0.8011	0.7971
2	0.8351	0.7950	0.8046	0.7837	0.7621	0.7541
3	0.7832	0.7503	0.7586	0.7433	0.7275	0.7168
4	0.7844	0.7397	0.7554	0.7356	0.7248	0.7135
5	0.7733	0.7334	0.7499	0.7377	0.7196	0.7082
6	0.7812	0.7373	0.7506	0.7333	0.7264	0.7136

NOTE: The bold entry marks the smallest error in each column.

Tables 2 compares the out-of-sample mean squared forecast errors of the rolling-window predictions using the following six methods: purely autoregressive (AR), principal component regression (PCR), factor-adjusted LASSO regression (KS2011), principal component regression with the number of components selected by cross-validation (PCR-CV), the “single” ridge regression (SRR), and the “double” ridge regression (DRR) in Theorem 6. We report only the results using the 5-fold cross-validation after bias correction, which are slightly better than that before correction. The implementation details are available in Appendix B.3 of the supplementary materials. The mean squared errors are normalized by that of the moving average forecasts (i.e., autoregressive estimator with $q = 0$) for comparison.

Our double ridge estimator, namely DRR, systematically improves the PCR estimators over different choices of the autoregressive orders $q \geq 1$, while the factor-adjusted LASSO estimator KS2011 consistently underperform the PCR. The cross-validation procedure hardly improves PCR predictions. These findings suggest that the underlying model tend to be dense rather than sparse in terms of idiosyncratic information. The DRR estimator consistently improves the prediction errors against SRR, which is consistent with our theory that DRR is more robust against complex covariates structures; see Theorem 6 and the discussions above it.

6. Concluding Remarks

This article studies the high-dimensional ridge regression methods under factor models. Our theory of the efficiency and robustness of ridge regression in capturing both factor and idiosyncratic information simultaneously helps explain its excellent performance in many complicated applications. Our results are applicable to large-dimensional datasets showing spiked eigenstructure, serial dependence, conditional heteroscedasticity, and heavy tails simultaneously. In particular, we extend the theory of k -fold cross-validation beyond iid setting and provide theoretical support for its applications to more complex data.

One can compare the factor-augmented models (1.3)–(1.4) with the following synthetic control models studied by Agarwal et al. (2021) (see also Abadie, Diamond, and Hainmueller 2010; Amjad, Shah, and Shen 2018; Arkhangelsky et al. 2021):

$$X = A + H, \quad Y = A\theta^* + \varepsilon + \phi, \quad (6.1)$$

where X is the *corrupted* covariate matrix, the latent factor matrix $A = FA' \in \mathbb{R}^{n \times r}$ is the true covariate matrix of low rank r with a strong signal $\|A\|_F \gg \|H\|_F$, ε contains response noises,

and ϕ are model misspecification errors. Rather than treating the misspecification error ϕ as deterministic and “inevitable” (see the interpretation below their Theorem 3.1 on p. 1735), here we model $\phi = H\Gamma\beta$ as a projection of noise matrix H through the random-effects β and an arbitrary rotation Γ . Therefore, the RR can outperform PCR by optimizing the bias-variance tradeoff on the misspecification errors ϕ . It is an exciting future research topic is to allow for weak factors in (6.1) such that $\|A\|_F \asymp \|H\|_F$.

One may wonder whether our findings generalize beyond the random matrix regime. In higher dimensions with $\kappa_n := \max\{p/n, 1\} = p/n \rightarrow \infty$, our results remain true by generalizing Assumption 2 as follows: $a_p/\kappa_n \rightarrow \infty$ (which is trivial if a_p is linear in p) and $\lambda_{\max}(\Omega_E) = O_{\mathbb{P}}(\kappa_n)$. To see this, first apply our results to the rescaled covariates $x_i \mapsto x_i/\sqrt{\kappa_n}$ with corresponding coefficients $\beta \mapsto \beta\sqrt{\kappa_n}$, and then transform back to the original scales. For lower dimensions with $p/n \rightarrow 0$ (but $p \rightarrow \infty$ sufficiently fast) and $\tau^2 > 0$, the optimal penalty $\gamma^* \rightarrow 0$ in (2.13), meaning that the optimal ridge estimator should approximate the least-squares estimator and outperform PCR. Otherwise, the optimal RR asymptotically reduces to the PCR when $\tau^2 = 0$; see Figure A.2 in Appendix A.9 for illustration. Non-asymptotic comparison between PCR and RR is possible using the concentration inequalities in martingale limit theory, but we leave it as future works.

Finally, our analysis assumes that the factor strength $a_p \rightarrow \infty$ so the standard PCA is (subspace) consistent. Is the RR still favorable otherwise? This is an interesting future research topic. Using the simulation example in the Introduction, we can illustrate that the ridge regression still consistently outperforms the PCR when the standard PCA becomes inconsistent. Due to the space limit, we leave the detailed illustrations to Appendix A.9 in the supplementary materials, where we also compare the regular and bias-corrected cross-validation and discuss a possible extension for “sparse + dense” idiosyncratic information.

Supplementary Materials

All the proofs of the theorems are available in the supplementary document, which also includes more technical discussions and remarks, extra simulation results, and useful lemmas that may be of independent interest.

Acknowledgments

The author thanks the Editor, Professor Marina Vannucci, an Associate Editor, and two reviewers for their valuable comments that led to this improved version of the article.

Disclosure Statement

The author reports there are no competing interests to declare.

ORCID

Yi He  <http://orcid.org/0000-0003-3540-4657>

References

- Abadie, A., Diamond, A., and Hainmueller, J. (2010). “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program,” *Journal of the American Statistical Association*, 105, 493–505. [12]

- Agarwal, A., Shah, D., Shen, D., and Song, D. (2021) "On Robustness of Principal Component Regression," *Journal of the American Statistical Association*, 116, 1731–1745. [1,4,12]
- Alter, O., Brown, P. O., and Botstein, D. (2000). "Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling," *Proceedings of the National Academy of Sciences*, 97, 10101–10106. [1]
- Amjad, M., Shah, D., and Shen, D. (2018). "Robust Synthetic Control," *Journal of Machine Learning Research*, 19, 802–852. [12]
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). "Synthetic Difference-In-Differences," *American Economic Review*, 111, 4088–4118. [12]
- Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [4]
- Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [11]
- (2021), "Approximate Factor Models with Weaker Loadings," ArXiv Working Paper arXiv:2109.03773. [4]
- Bai, Z. D., and Silverstein, J. W. (1998). "No Eigenvalues Outside the Support of the Limiting Spectral Distribution of Large-Dimensional Sample Covariance Matrices," *The Annals of Probability*, 26, 316–345. [10]
- (2010), *Spectral Analysis of Large Dimensional Random Matrices*, New York: Springer. [4]
- Bergmeir, C., Hyndman, R. J., and Koo, B. (2018), "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction," *Computational Statistics & Data Analysis*, 120, 70–83. [9]
- Blum, J., and Eisenberg, B. (1975), "The Law of Large Numbers for Subsequences of a Stationary Process," *The Annals of Probability*, 3, 281–288. [8]
- Blum, J., and Hanson, D. (1960), "On the Mean Ergodic Theorem for Subsequences," *Bulletin of the American Mathematical Society*, 66, 308–311. [8]
- Bradley, R. C. (1986), "Basic Properties of Strong Mixing Conditions," in *Dependence in Probability and Statistics: A Survey of Recent Results*, eds. E. Eberlein and M. S. Taqqu, pp. 165–192, Boston: Birkhäuser. [8]
- Bradley, R. C. (2005), "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144. [8]
- Cai, T. T., Han, X., and Pan, G. (2020), "Limiting Laws for Divergent Spiked Eigenvalues and Largest Nonspiked Eigenvalue of Sample Covariance Matrices," *The Annals of Statistics*, 48, 1255–1280. [3,4]
- Carrasco, M., and Rossi, B. (2016), "In-Sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business & Economic Statistics*, 34, 313–338. [1]
- Castle, J. L., Clements, M. P., and Hendry, D. F. (2013), "Forecasting by Factors, by Variables, by Both or Neither?," *Journal of Econometrics*, 177, 305–319. [1,3]
- Coates, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). "Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, 61, 38–59. [1]
- Davis, R. A., and Mikosch, T. (2009). "Probabilistic Properties of Stochastic Volatility Models," in *Handbook of Financial Time Series*, eds. T. G. Andersen, R. A. Davis, J. P. Kreiß, and T. V. Mikosch, pp. 255–267, Berlin: Springer. [8]
- De Mol, C., Giannone, D., and Reichlin, L. (2008), "Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?," *Journal of Econometrics*, 146, 318–328. [1,6]
- Dicker, L. H. (2016). "Ridge Regression and Asymptotic Minimax Estimation Over Spheres of Growing Dimension," *Bernoulli*, 22, 1–37. [3]
- Dobriban, E., and Wager, S. (2018), "High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification," *The Annals of Statistics*, 46, 247–279. [3,4]
- Fan, J., Ke, Y., and Wang, K. (2020), "Factor-Adjusted Regularized Model Selection," *Journal of Econometrics*, 216, 71–85. [3]
- Giannone, D., Lenza, M., and Primiceri, G. E. (2021), "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica*, 89, 2409–2437. [3]
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [2,5]
- Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022), "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," *The Annals of Statistics*, 50, 949–986. [3]
- Hayashi, F. (2000), *Econometrics* Princeton: Princeton University Press. [9]
- Hallin, M., and Liska, R. (2007). "The Generalized Dynamic Factor Model: Determining the Number of Factors," *Journal of the American Statistical Association*, 102, 603–617. [4]
- Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [3,5]
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning: with Applications in R* (2nd ed), New York: Springer. [1,5,7]
- Johnstone, I. M. (2001). "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of Statistics*, 29, 295–327. [4]
- Johnstone, I. M., and Lu, A. Y. (2009). "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682–693. [3]
- Johnstone, I. M., and Paul, D. (2018). "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, 106, 1277–1292. [3]
- Jolliffe, I. T. (1982). "A Note on the Use of Principal Components in Regression," *Journal of the Royal Statistical Society, Series C*, 31, 300–303. [3]
- Jung, S., and Marron, J. S. (2009). "PCA Consistency in High Dimension, Low Sample Size Context," *The Annals of Statistics*, 37, 4104–4130. [4]
- Kneip, A., and Sarda, P. (2011), "Factor Models and Variable Selection in High-Dimensional Regression Analysis," *The Annals of Statistics*, 39, 2410–2447. [3,10]
- Liu, S., and Dobriban, E. (2020), "Ridge Regression: Structure, Cross-Validation, and Sketching," in *International Conference on Learning Representations*. [3]
- Marčenko, V. A., and Pastur, L. A. (1967). "Distribution of Eigenvalues for Some Sets of Random Matrices," *Mathematics of the USSR-Sbornik*, 1, 457–483. [6]
- McCracken, M. W., and Ng, S. (2016), "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589. [11]
- Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [10]
- Ng, S. (2013), "Variable Selection in Predictive Regressions," in *Handbook of Economic Forecasting* (Vol. 2), eds. G. Elliott and A. Timmermann, pp. 752–789, Amsterdam: Elsevier. [1]
- Onatski, A. (2009). "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, 77, 1447–1479. [4]
- (2012). "Asymptotics of the Principal Components Estimator of Large Factor Models With Weakly Influential Factors," *Journal of Econometrics*, 168, 244–258. [3]
- Onatski, A., and Wang, C. (2021), "Spurious Factor Analysis," *Econometrica*, 89, 591–614. [4]
- Paul, D. (2007). "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model," *Statistica Sinica*, 17, 1617–1642. [3]
- Paul, D., and Silverstein, J. W. (2009). "No Eigenvalues Outside the Support of the Limiting Empirical Spectral Distribution of a Separable Covariance Matrix," *Journal of Multivariate Analysis*, 100, 37–57. [4]
- Preisendorfer, R. W. (1988). *Principal Component Analysis in Meteorology and Oceanography*, Amsterdam: Elsevier. [1]
- Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [1,11]
- (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business & Economic Statistics*, 30, 481–493. [1,4]
- Tikhonov, A. N. (1963a), "Regularization of Incorrectly Posed Problems," *Soviet Mathematics Doklady*, 4, 1624–1627 (English translation). [3,5]
- (1963b), "Solution of Incorrectly Formulated Problems and the Regularization Method," *Soviet Mathematics Doklady*, 4, 1035–1038 (English translation). [3,5]