Annual Review of Financial Economics

Recent Developments in Factor Models and Applications in Econometric Learning

Jianqing Fan,[1] Kunpeng Li,[2] and Yuan Liao[3]

[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, USA; email: jqfan@princeton.edu

[2]International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China

[3]Department of Economics, Rutgers University, New Brunswick, New Jersey 08901, USA

Annu. Rev. Financ. Econ. 2021. 13:401–30

The *Annual Review of Financial Economics* is online at financial.annualreviews.org

https://doi.org/10.1146/annurev-financial-091420-011735

JEL codes: C58, C01

Keywords

factor models, spiked low-rank matrix, matrix completion, unbalanced panel, factor adjustments, robustness, model selection, multiple testing, high-dimensional statistics

Abstract

This article provides a selective overview of the recent developments in factor models and their applications in econometric learning. We focus on the perspective of the low-rank structure of factor models and particularly draw attention to estimating the model from the low-rank recovery point of view. Our survey mainly consists of three parts. The first part is a review of new factor estimations based on modern techniques for recovering low-rank structures of high-dimensional models. The second part discusses statistical inferences of several factor-augmented models and their applications in statistical learning models. The final part summarizes new developments dealing with unbalanced panels from the matrix completion perspective.

1. INTRODUCTION

The recent decade has witnessed a blossoming of developments in statistical learning theories and practice, fueled by the exciting progress of large-scale optimizations and dimension reduction techniques. Factor models, as one of the central machineries for summarizing and extracting information from large-scale data sets, have received much attention in this revolutionary era of data science, and many breakthrough methodologies and applications have been developed in this exciting area.

This article makes a selective overview of the recent developments in factor models and their applications in econometric learning. Our review focuses on the perspective of the low-rank structure of factor models and draws particular attention to estimating the model from the low-rank recovery point of view. A central focus in the progress of this literature is the understanding and recovering of low-rank structures of high-dimensional models. Many new learning theories and methods have been developed and have revolutionized the modern understanding of econometric modeling. Meanwhile, the low-rank structure is one of the key properties of factor models. While researchers have long been aware of this structure, studying the factor model from the perspective of low-rank matrix recovery is relatively new and has led to many exciting new discoveries and understandings.

The survey mainly consists of three parts. The first part is a review on new factor estimation based on modern techniques for recovering low-rank structures of high-dimensional models. The second part discusses statistical inferences of several factor-augmented models and applications in statistical learning models. The final part summarizes new developments dealing with unbalanced panels from the matrix completion perspective.

We concentrate on recent developments in methodologies and applications in econometric learning. For a more comprehensive account on this topic, see chapters 9–11 of the book by Fan et al. (2020). Meanwhile, several important topics that are not covered in this survey have generated extensive research in the literature. Those topics include selecting the number of factors, weak factors, identification, continuous-time and time-varying models, nonstationarity and structural breaks, Bayesian methods, bootstrap factors, as well as more sophisticated panel data models. Several excellent reviews have been written with emphasis on these topics. We refer readers to the reviews by Stock & Watson (2016) for dynamic factor models with applications on macroeconomics, Bai & Wang (2016) for time series and panel data models, and Gagliardini, Ossola & Scaillet (2019) for a recent review on conditional factor models with applications to finance. Another class of estimation is a hybrid of the principal components analysis (PCA) method and the state space approach (for more discussions on this topic, see Giannone, Reichlin & Small 2008; Doz, Giannone & Reichlin 2011). In addition, the generalized dynamic factor model is another important strand of literature in which factors are often estimated using the dynamic principal components, the frequency domain analog of principal components, developed by Brillinger (1964). Forni et al. (2000, 2005) provide rates of convergence of the common component estimated by dynamic principal components. Finally, for more detailed developments, we refer readers to the following papers, among others: Bai & Ng (2002); Onatski (2010, 2012); Chudik, Pesaran & Tosetti (2011); Bai & Li (2012, 2016); Ahn & Horenstein (2013); Cheng, Liao & Schorfheide (2016); Gagliardini, Ossola & Scaillet (2016); Aït-Sahalia & Xiu (2017); Baltagi, Kao & Wang (2017); Li, Li & Shi (2017); Massacci (2017); Su & Wang (2017); Barigozzi, Cho & Fryzlewicz (2018); Liao & Yang (2018); Li, Todorov & Tauchen (2019); Pelger (2019); Chen, Mykland & Zhang (2020); and Goncalves & Perron (2020).

We use the following notation: For a matrix $\mathbf{A}$, let $\lambda_i(\mathbf{A})$ denote the ith largest singular value of $\mathbf{A}$ and use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote its smallest and largest eigenvalues. We define

the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}'\mathbf{A})}$, the operator norm $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}'\mathbf{A})}$, the element-wise norm $\|\mathbf{A}\|_\infty = \max_{ij} |A_{ij}|$, and the matrix ℓ_1-norm $\|\mathbf{A}\|_{\ell_1} := \max_{i \le N} \sum_{j=1}^{N} |A_{ij}|$. In addition, define projection matrices $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}$ and $\mathbf{M}_\mathbf{A} = \mathbf{I} - \mathbf{P}_\mathbf{A}$ when $\mathbf{A}'\mathbf{A}$ is invertible. Finally, for two (random) sequences a_T and b_T, we write $a_T \ll b_T$ (or $b_T \gg a_T$) if $a_T = o_P(b_T)$.

2. SPIKED INCOHERENT LOW-RANK MODELS

2.1. The Model

Modern high-dimensional factor models can be viewed as a type of spiked incoherent low-rank model, a broad class of models that have drawn active research in the last decade. A spiked incoherent low-rank model typically refers to a large matrix $\mathbf{\Sigma}$ (either observable or not), having the following decomposition:

$$\mathbf{\Sigma} = \mathbf{L} + \mathbf{S}. \qquad 1.$$

Such decomposition requires the following three assumptions:

1. **Low-rank.** The rank of $\mathbf{L}$ is either bounded or grows very slowly compared with its dimensions.
2. **Spikedness.** The nonzero singular values of $\mathbf{L}$ grow fast, while the largest singular value of $\mathbf{S}$ is either bounded or grows much slower.
3. **Incoherence.** The left and right singular vectors of $\mathbf{L}$, corresponding to the nonzero singular values, should have diversified elements, which means that elements of the rescaled singular vectors should be uniformly bounded.

The low-rank structure achieves dimension reductions: Suppose the matrix $\mathbf{\Sigma}$ is of $N \times N_1$ dimensions, while the rank of $\mathbf{L}$ is r. Then, the low-rank structure reduces the dimension from $O(NN_1)$ to $O((N + N_1)r)$; the latter is the magnitude of the number of parameters in $\mathbf{L}$. Meanwhile, the spikedness helps separate $\mathbf{L}$ from $\mathbf{S}$, approximately, and ensures that the large signals concentrate on $\mathbf{L}$, the low-rank component. In addition, the incoherence, a condition that excludes matrices being low-rank and sparse simultaneously, enables us to estimate well the singular eigenvectors.

We explain these three properties using the matrix form of factor models. Consider

$$y_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad i \le N, \quad t \le T, \qquad 2.$$

where $\mathbf{f}_t$ is an r-dimensional vector of factors, $\mathbf{b}_i$ is the loading vector, and u_{it} is the idiosyncratic noise. Specifically, Equation 1 applies to two decompositions of this model.

2.1.1. Factor decomposition.

The matrix form of the factor model gives

$$\mathbf{Y} = \mathbf{M} + \mathbf{U}, \quad \mathbf{M} := \mathbf{B}\mathbf{F}',$$

where $\mathbf{Y}$ and $\mathbf{U}$ are $N \times T$ matrices of y_{it} and u_{it}, $\mathbf{B}$ is the $N \times r$ matrix of $\mathbf{b}_i$, while $\mathbf{F}$ is the $T \times r$ matrix of $\mathbf{f}_t$. Then, corresponding to the notation in Equation 1, $\mathbf{\Sigma} = \mathbf{Y}$, $\mathbf{L} = \mathbf{M}$, and $\mathbf{S} = \mathbf{U}$. In this decomposition, $\mathbf{\Sigma}$ is observable. Apparently, $\mathbf{M}$ is a low-rank matrix with rank r. The nonzero singular values of $\mathbf{M}$, under the strong factor assumption, grow much faster than those of $\mathbf{U}$, which gives rise to the spikedness property. To discuss the incoherence assumption, let $\boldsymbol{\xi}$ be the $N \times r$ matrix whose columns are the left singular vectors of $\mathbf{M}$, and let $\boldsymbol{\xi}_i'$ denote its ith row. The incoherence of singular vectors requires

$$\max_{i \le N} \|\sqrt{N}\boldsymbol{\xi}_i\| \le \sqrt{Cr} \qquad 3.$$

for some constant C that can possibly grow. The right singular vectors can be bounded similarly.

Remark 1 (Relation with pervasiveness). A more familiar condition that is often imposed on factor models is known as pervasiveness, which assumes that all the eigenvalues of the r by r matrix $\frac{1}{N}\mathbf{B}'\mathbf{B}$ are well bounded from both below and above (some rate conditions should be imposed to make this statement formal). Under some mild conditions, the pervasiveness implies the spikedness condition with more specific growth rate. It is used to identify, approximately, the factor component and idiosyncratic component. On the other hand, the incoherence is another condition that helps the identification issue. It relaxes the rates on the spikeness: The slower the growth of the incoherent constant C in Equation 3, the more relaxed the spikedness condition.

2.1.2. Covariance decomposition. It is also well known from the factor model (Equation 2) that the covariance matrix of $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$, denoted by $\mathbf{\Sigma}_y$, can be decomposed as follows:

$$\mathbf{\Sigma}_y = \mathbf{L} + \mathbf{\Sigma}_u, \quad \mathbf{L} := \mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}', \tag{4.}$$

where $\mathbf{\Sigma}_u$ denotes the covariance matrix of $\mathbf{u}_t$. The above decomposition is well known for portfolio allocations and risk managements, where the total volatility is decomposed into the systematic risk $\mathbf{L}$, plus the (sparse) idiosyncratic risk $\mathbf{\Sigma}_u$. It also leads to the spiked incoherent low-rank model, but $\mathbf{\Sigma}_y$ is unknown and needs to be estimated.

2.2. Estimation

There are two general approaches to estimating the model in Equation 1: (*a*) PCA and (*b*) low-rank regularization. Here, we present a general PCA estimation setting and defer the discussion of low-rank regularization to Section 3.2. We shall assume $\mathrm{rank}(\mathbf{L}) = r$ to be known.

For any matrix $\mathbf{A}$, let $\mathbf{A} = \mathbf{U}_A\mathbf{D}_A\mathbf{V}_A'$ denote the singular value decomposition (SVD) of $\mathbf{A}$. Define the singular value hard-thresholding operator as

$$H_R(\mathbf{A}) := \mathbf{U}_A\bar{\mathbf{D}}_R\mathbf{V}_A', \tag{5.}$$

where $\bar{\mathbf{D}}_R$ is a diagonal matrix that keeps the top R diagonal elements of $\mathbf{D}_A$ and replaces the remaining elements by zeros. So $H_R(\mathbf{A})$ is the best rank R matrix approximation to $\mathbf{A}$.

Suppose an estimator of $\mathbf{\Sigma}$, denoted by $\widehat{\mathbf{\Sigma}}$, is available, satisfying

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = O_P(\eta_N), \quad \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty = O_P(c_N) \tag{6.}$$

for some sequences η_N and c_N. We use $\widehat{\mathbf{\Sigma}}$ as the input matrix, which can be the sample covariance matrix or its robust versions (Fan, Wang & Zhong 2019). The goal is to estimate $\mathbf{L}$ in Equation 1 and its $N \times r$ matrix of the left singular vectors, denoted by $\boldsymbol{\xi}$ (also let $\boldsymbol{\zeta}$ denote its right singular vectors). We use, respectively, $\widehat{\mathbf{L}} := H_R(\widehat{\mathbf{\Sigma}})$ with $R = r$, which is the rank r projection of $\widehat{\mathbf{\Sigma}}$, and the $N \times r$ matrix $\widehat{\boldsymbol{\xi}}$, whose columns are the left singular vectors of $\widehat{\mathbf{\Sigma}}$. The following theorem, adapted from Fan, Wang & Zhong (2018), provides deviation bounds of the estimators. To make this article self-contained, we also provide a simpler proof with slightly different conditions.

Theorem 1. Consider the general model given in Equation 1 with bounded $r := \mathrm{rank}(\mathbf{L})$. Suppose that $\min_{2\le i\le r+1} |\lambda_{i-1}(\mathbf{L}) - \lambda_i(\mathbf{L})| \asymp \max_{2\le i\le r+1} |\lambda_{i-1}(\mathbf{L}) - \lambda_i(\mathbf{L})| := g_N$ and $\eta_N + \|\mathbf{S}\| = o_P(g_N)$. Then, under the condition from Equation 6, we have result 1:

$$\|\widehat{\mathbf{L}} - \mathbf{L}\| = O_P\left(\eta_N + \|\mathbf{S}\|\right), \quad \|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}\| = O_P\left(\frac{\eta_N + \|\mathbf{S}\|}{g_N}\right).$$

And result 2 states: If, additionally, $\|\mathbf{S}\|_\infty + \|\mathbf{L}\|_\infty = O_P(1)$ and $N_1 c_N = o_P(g_N)$, then

$$\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|_\infty \le O_P\left(\frac{N_1}{\sqrt{N}} + \sqrt{N_1}\right)(\eta_N + \|\mathbf{S}\|)g_N^{-2} + O_P\left(c_N\frac{N_1}{\sqrt{N}} + c_N\sqrt{N_1} + \|\mathbf{S}\boldsymbol{\zeta}_d\|_\infty \vee \|\mathbf{S}'\boldsymbol{\xi}_d\|_\infty\right)g_N^{-1}.$$

Proof. See the online **Supplemental Materials**. □

Supplemental Material >

This theorem is relatively general and is applicable to low-rank models that are not necessarily consequences from factor models. The proof relies on perturbation bounds for singular vectors/values, and the achieved rates are sharp. Result 1 is simple and gives asymptotic bounds under the operator norm. Result 2 gives an element-wise deviation bound for the singular vectors, which requires more dedicated technical arguments.

3. ESTIMATION UNDER FACTOR MODELS

We observe an $N \times T$ data matrix $\mathbf{Y}$, which can be decomposed as

$$\mathbf{Y} = \mathbf{M} + \mathbf{U} = \mathbf{B}\mathbf{F}' + \mathbf{U},$$

where $\mathbf{B}$ is $N \times r$ factor loadings matrix, $\mathbf{F}$ is $T \times r$ factors matrix, and $\mathbf{U}$ is $N \times T$ idiosyncratic errors, which are uncorrelated with $\mathbf{M} := \mathbf{B}\mathbf{F}'$. All the three parts, $\mathbf{B}$, $\mathbf{F}$, and $\mathbf{U}$, are unobserved. The tth column of this expression can be written as

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t. \qquad 7.$$

3.1. Principal Components Analysis and Maximum Likelihood Estimation

This section gives the least-squares estimation and maximum likelihood estimation (MLE).

3.1.1. Principal components analysis. Under the model's specification, we have the covariance structure (Equation 4). One of the most widely used estimation methods for the factor loading matrix and latent factors is PCA. Define the sample covariance $\mathbf{S}_y = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t' = \frac{1}{T}\mathbf{Y}\mathbf{Y}'$. Let $\widehat{\boldsymbol{\xi}}_j$ be the jth eigenvector corresponding to the largest jth eigenvalue of $\mathbf{S}_y$. The PCA estimates $\mathbf{B}$ by taking $\widehat{\mathbf{B}} = \sqrt{N}(\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_R)$, which estimates $\mathbf{B}$ up to a diagonal transformation. Given $\widehat{\mathbf{B}}$, the factors can be estimated via the least squares:

$$\widehat{\mathbf{F}} = \mathbf{Y}'\widehat{\mathbf{B}}(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1} = \frac{1}{N}\mathbf{Y}'\widehat{\mathbf{B}}.$$

This also leads to the estimated low-rank component $\frac{1}{T}\widehat{\mathbf{B}}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\widehat{\mathbf{B}}'$ for $\mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'$.

PCA is equivalent to the singular value hard-thresholding by taking the input matrix $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}_y$. Then $\frac{1}{T}\widehat{\mathbf{B}}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\widehat{\mathbf{B}}' = H_R(\mathbf{S}_y)$. One can then apply Theorem 1 to infer the rates of convergence of the PCA estimators, which were obtained by Stock & Watson (2002a). Bai (2003) proved the asymptotic normality of PCA estimators for the factors and loadings. Results with general input $\hat{\boldsymbol{\Sigma}}$ can be found in chapter 10 of Fan et al. (2020).

3.1.2. Maximum likelihood estimations. Another popular method to estimate a factor model is the maximum likelihood (ML) method (see, e.g., Lawley & Maxwell 1971; Bai & Li 2012;

Doz, Giannone & Reichlin 2012). Under the independence and normality assumptions, the log-likelihood function based on $\mathbf{y}_t$ is, for some constant C,

$$\log L_{\mathrm{ML}}(\mathbf{B}, \mathrm{cov}(\mathbf{f}_t), \mathrm{diag}(\mathbf{\Sigma}_u)) = C - \frac{T}{2}\ln|\mathbf{\Sigma}_y| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{y}_t'\mathbf{\Sigma}_y^{-1}\mathbf{y}_t.$$

The log-likelihood function is then maximized with respect to the matrix parameters $(\mathbf{B}, \mathrm{cov}(\mathbf{f}_t), \mathrm{diag}(\mathbf{\Sigma}_u))$ under additional restrictions that $\mathbf{\Sigma}_u$ is diagonal (Bai & Li 2012, 2016) or sparse with regularizations (Bai & Liao 2016, Wang, Yang & Yao 2019). Recently, Barigozzi & Luciani (2019) explicitly accounted for autocorrelations of the factors in the likelihood function.

The factors can be estimated by two methods, one of which is the projection method. Under the joint normality assumptions of $\mathbf{f}_t$ and $\mathbf{u}_t$, we have

$$\mathbb{E}(\mathbf{f}_t|\mathbf{y}_t) = \mathbf{B}'(\mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u)^{-1}\mathbf{y}_t = (\mathbf{I}_r + \mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{y}_t.$$

This provides the basis of estimating factors. The other approach is the generalized least squares (GLS): for given $\mathbf{B}$ and $\mathbf{\Sigma}_u^{-1}$, the GLS estimator for $\mathbf{f}_t$ is

$$\widehat{\mathbf{f}}_t = (\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{y}_t.$$

Replacing the unknown parameters with their ML estimators, one obtains two estimators for the latent factors. Under large-N setup, the differences of the two methods (PCA and MLE) for estimating factors are asymptotically negligible.

3.2. Low-Rank Estimation

As an alternative to PCA, one can estimate $\mathbf{M}$ directly, taking advantage of its low-rank structure, based on the nuclear-norm regularization, the ℓ_1-norm of singular values, that encourages the sparseness in singular values and hence low-rankness. For an $n \times m$ matrix $\mathbf{A}$, let $\|\mathbf{A}\|_n := \sum_{i=1}^{\min\{m,n\}}\psi_i(\mathbf{A})$ be its nuclear-norm, where $\psi_i(\mathbf{A})$ is the ith largest singular value of $\mathbf{A}$.

3.2.1. Singular value thresholding. Given the low-rank structure of $\mathbf{M}$ (sparsity in singular value of $\mathbf{M}$), we can estimate the model via solving the following penalized optimization:

$$\widehat{\mathbf{M}} = \arg\min_{M}\frac{1}{2}\|\mathbf{Y} - \mathbf{M}\|_F^2 + \nu\|\mathbf{M}\|_n \qquad 8.$$

for some tuning parameter $\nu > 0$. The solution is $\widehat{\mathbf{M}} = S_\nu(\mathbf{Y})$, where $S_\nu(\cdot)$ is the singular value thresholding operator (Ma, Goldfarb & Chen 2011), defined as follows. Let $\mathbf{Y} = \mathbf{U}_y\mathbf{D}\mathbf{V}_y'$ be its SVD. Then $S_\nu(\mathbf{Y}) := \mathbf{U}_y\mathbf{D}_\nu\mathbf{V}_y'$, where $\mathbf{D}_\nu = \mathrm{diag}(\{D_{ii} - \nu\}_+)$, with D_{ii} being the diagonal entries of $\mathbf{D}$. So, $S_\nu(\mathbf{Y})$ applies soft-thresholding on the singular values of $\mathbf{Y}$. One can additionally estimate the factors and loadings using the singular vectors.

We note that this method is closely related to the principal components estimator, except the soft-thresholding is replaced by hard-thresholding. Let R denote the working number of factors, which is the number of principal components one takes when applying the principal components method. We note that the principal components estimator for $\mathbf{M}$ with R factors is given by (see also Section 2.2)

$$\widehat{\mathbf{M}}_{\mathrm{PC}} = H_R(\mathbf{Y}), \quad H_R(\mathbf{Y}) := \mathbf{U}_y\bar{\mathbf{D}}_R\mathbf{V}_y'.$$

This estimator is the solution to the penalized least squares problem (Equation 8) except that the nuclear-norm is replaced by $\sum_{i=1}^{\min\{N,T\}} p_\nu(\psi_i(\mathbf{M}))$, where $p_\nu(\theta) = \nu^2 - (\nu - |\theta|)_+^2$ is the hard-thresholding penalty (Fan et al. 2020) and $\psi_i(\mathbf{M})$ is the ith singular value of $\mathbf{M}$.

Therefore, the difference between Equation 8 and PCA is more fundamentally about that of hard- and soft-thresholding. In spite of many good properties, the soft-thresholding estimator possesses shrinkage bias, while the hard-thresholding estimator reduces the bias. As a matter of fact, the shrinkage bias is on the singular values, rather than on the singular vectors. Indeed, the singular vectors of the two estimators are the same and equal to the top R singular vectors of $\mathbf{Y}$. An important implication is that the factor estimator building on $\widehat{\mathbf{M}}$ is numerically equivalent to the principal components estimators for the factors, which do not suffer from any shrinkage bias. A formal statement and proof of the unbiasedness of eigenvectors can be found in the work by Fan et al. (2019b).

3.2.2. Low-rank plus sparse decomposition. Recall that $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_u$ denote the $N \times N$ covariance matrices of $\mathbf{y}_t$ and $\mathbf{u}_t$ in the model in Equation 7 and that we have the following decomposition:

$$\mathbf{\Sigma}_y = \mathbf{L} + \mathbf{\Sigma}_u, \quad \mathbf{L} := \mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'. \tag{9.}$$

We now demonstrate that this decomposition also provides a nice structure for estimating the covariance components. A key assumption is conditionally sparsity, namely, $\mathbf{\Sigma}_u$ is sparse. While the definition of sparsity may differ in different contexts, here we mean

$$J := \sum_{i \neq j} 1\{\mathbb{E}\, u_{it} u_{jt}\}$$

should not grow too fast, as $N \to \infty$. This requirement can be weakened to approximate sparsity, replacing ℓ_0-norm, the indicator function, in the definition of sparsity by the ℓ_q-norm ($q < 1$). In addition, $\mathbf{L}$ is a low-rank matrix. Thus, we can directly estimate the above covariance decomposition via solving the following penalized optimization:

$$(\widehat{\mathbf{L}}, \widehat{\mathbf{\Sigma}}_u) := \arg\min_{\mathbf{L}, \mathbf{\Sigma}_u} \frac{1}{2} \|\mathbf{S}_y - (\mathbf{L} + \mathbf{\Sigma}_u)\|_F^2 + \nu_1 \|\mathbf{L}\|_n + \nu_2 \|\mathbf{\Sigma}_u\|_1, \tag{10.}$$

where ν_1 and ν_2 are tuning parameters. Note that here we use the notation $\|\mathbf{A}\|_1 = \sum_{i,j} |A_{ij}|$ as the matrix element-wise 1-norm, distinguished from the usual matrix ℓ_1-norm $\|\mathbf{A}\|_{\ell_1} := \max_{i \leq N} \sum_{j=1}^N |A_{ij}|$. The above optimization has been employed by many authors to study the low-rank plus sparse decomposition, while some authors exclude the diagonal elements of $\mathbf{\Sigma}_u$ from the penalization and additionally impose positive-definite and other constraints on $\mathbf{L}$ and $\mathbf{\Sigma}_u$ (Agarwal, Negahban & Wainwright 2012; Klopp, Lounici & Tsybakov 2017). Finally, given $\widehat{\mathbf{L}}$, we can estimate the factors and loadings by extracting its eigenvectors.

The above optimization can be solved by alternating the estimation of $\mathbf{L}$ and $\mathbf{\Sigma}_u$, and closed form solutions are available in both iterations. Given $\mathbf{\Sigma}_u$, solving for $\mathbf{L}$ leads to the singular value soft-thresholding, $\widehat{\mathbf{L}} = S_{\nu_1}(\mathbf{S}_y - \mathbf{\Sigma}_u)$, and given $\mathbf{L}$, solving for $\mathbf{\Sigma}_u$ leads to the element-wise soft-thresholding, $\widehat{\mathbf{\Sigma}}_u = \widetilde{S}_{\nu_2}(\mathbf{S}_y - \mathbf{L})$. While both iterations solve convex problems, standard convergence analysis can be applied to show that the iterative algorithm converges in polynomial time.

Agarwal, Negahban & Wainwright (2012) and Klopp, Lounici & Tsybakov (2017) study the statistical convergence properties of Equation 10. Let columns of $\mathbf{U}_{L,2}$ be the singular vectors of the true $\mathbf{L}$ corresponding to the zero singular values. Define projections $\mathcal{P}(\mathbf{A}) := \mathbf{U}_{L,2}\mathbf{U}_{L,2}'\mathbf{A}\mathbf{U}_{L,2}\mathbf{U}_{L,2}'$ and $\mathcal{M}(\mathbf{A}) := \mathbf{A} - \mathcal{P}(\mathbf{A})$. In addition, let $(\mathbf{A})_J$ and $(\mathbf{A})_{J^c}$ be the submatrices of $\mathbf{A}$, whose elements respectively correspond to $\mathbb{E}\, u_{it} u_{jt} \neq 0$ and $\mathbb{E}\, u_{it} u_{jt} = 0$. Additionally, define

$$\mathcal{C}(\nu_1, \nu_2) := \{(\mathbf{A}_1, \mathbf{A}_2) : \nu_1 \|\mathcal{P}(\mathbf{A}_1)\|_n + \nu_2 \|(\mathbf{A}_2)_{J^c}\|_1 \leq 3\nu_1 \|\mathcal{M}(\mathbf{A}_1)\|_n + 3\nu_2 \|(\mathbf{A}_2)_J\|_1\}.$$

A key quantity is the restricted strong convexity (RSC) constant, which is defined as follows:

$$\kappa(\nu_1, \nu_2) := \sup\{c > 0 : \|\mathbf{A}_1 + \mathbf{A}_2\|_F^2 \geq c\|\mathbf{A}_1\|_F^2 + c\|\mathbf{A}_2\|_F^2 \text{ for all } (\mathbf{A}_1, \mathbf{A}_2) \in \mathcal{C}(\nu_1, \nu_2)\}.$$

Supplemental Material >

We then have the following theorem, adapted from Agarwal, Negahban & Wainwright (2012). To make the article self-contained, we also provide a proof with slightly different conditions (see the online **Supplemental Materials**).

Theorem 2. Conditioning on events $4\|\mathbf{S}_y - \mathbf{\Sigma}_y\| \le \nu_1$ and $4\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_\infty \le \nu_2$, there is $C > 0$ that depends only on rank($\mathbf{L}$), so that

$$\frac{1}{N^2}\|\widehat{\mathbf{L}} - \mathbf{L}\|_F^2 + \frac{1}{N^2}\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\|_F^2 \le \frac{C}{\kappa^2(\nu_1, \nu_2)} \frac{(\nu_1^2 + (J+N)\nu_2^2)}{N^2}.$$

Proof. See the online **Supplemental Materials**. □

The optimal tuning parameters can be set to satisfy $\nu_1 \asymp \frac{N}{\sqrt{T}}$ and $\nu_2 \asymp \sqrt{\frac{\log N}{T}}$, respectively, accounting for estimating errors under two matrix norms:

$$\|\mathbf{S}_y - \mathbf{\Sigma}_y\| \le \nu_1, \quad \|\mathbf{S}_y - \mathbf{\Sigma}_y\|_\infty \le \nu_2.$$

Both can be shown to hold with high probability under weak serial dependence and sub-Gaussian conditions. In addition, if $\kappa(\nu_1, \nu_2)$ is bounded away from zero, with the choice of tunings, the convergence rate in Theorem 1 is $O_P(1 + \frac{J\log N}{N^2})\frac{1}{T}$, which is sufficient to guarantee the convergence of the estimated factors and loadings. We refer to Lemma 2 of Agarwal, Negahban & Wainwright (2012) for a more refined lower bound of $\kappa(\nu_1, \nu_2)$.

The above problem is also termed robust PCA (Candès et al. 2011). For recent advances and references, see the work by Chen et al. (2020b), in which factorization methods are also discussed.

3.3. Covariance Estimation

Fan, Liao & Mincheva (2013) propose a nonparametric estimator of $\mathbf{\Sigma}_y$, named POET (Principal Orthogonal complEment Thresholding), when the factors are unobservable. It is basically a one-step solution to the optimization given in Equation 10 with initialization $\mathbf{\Sigma}_u = 0$. To motivate the estimator, suppose $r = R$. Then, heuristically,

$$\mathbf{L} \approx H_R(\mathbf{\Sigma}_y), \quad \mathbf{\Sigma}_u \approx \mathbf{\Sigma}_y - H_R(\mathbf{\Sigma}_y).$$

Thus, one estimates $\mathbf{L}$ by $H_R(\mathbf{S}_y)$ and sets $\mathbf{S}_u := \mathbf{S}_y - H_R(\mathbf{S}_y)$. To account for the sparsity assumption on $\mathbf{\Sigma}_u$, Fan, Liao & Mincheva (2013) estimate $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_y$ as

$$\widehat{\mathbf{\Sigma}}_u = (h(S_{u,ij}, \lambda_{ij}))_{N\times N}, \quad \widehat{\mathbf{\Sigma}}_y = H_R(\mathbf{S}_y) + \widehat{\mathbf{\Sigma}}_u, \tag{11.}$$

where $h(x, \lambda_{ij})$ denotes the element-wise thresholding operator with thresholding value λ_{ij}. Here, we emphasize element-dependent thresholding λ_{ij} to adapt to varying scales of covariance. For correlation thresholding at level λ, we take $\lambda_{ij} = \lambda\sqrt{s_{u,ii}s_{u,jj}}$, with $s_{u,ii}$ as a diagonal element of $\mathbf{S}_u$ (Fan, Liao & Mincheva 2013); we can also take other forms, such as the adaptive thresholding from Cai & Liu (2011). In general, the thresholding function should satisfy the following conditions:

1. $h(x, \lambda) = 0$ if $|x| < \lambda$,
2. $|h(x, \lambda) - x| \le \lambda$, and
3. there are constants $a > 0$ and $b > 1$, such that $|h(x, \lambda) - x| \le a\lambda^2$ if $|x| > b\lambda$.

Note that condition 3 requires that the thresholding bias should be of higher order. It is not necessary for consistent estimations, but we recommend using nearly unbiased thresholding (Antoniadis & Fan 2001) for inference applications. One such example is known as SCAD

(smoothly clipped absolute deviation). As noted by Fan, Liao & Yao (2015), the unbiased thresholding is required to avoid size distortions in a large class of high-dimensional testing problems involving a plug-in estimator of $\mathbf{\Sigma}_u$. In particular, this rules out the popular soft-thresholding function, which does not satisfy condition 3 due to its first-order shrinkage bias.

3.4. Projected Principal Components Analysis

In empirical asset pricing, factor loadings are known to depend on individual-specific observables $\mathbf{X}_i$, which represent a set of time-invariant characteristics such as individual stocks' size, momentum, and values. To incorporate the information carried by the observed characteristics, Connor & Linton (2007) and Connor, Matthias & Linton (2012) model explicitly the loading matrix as a function of covariates $\mathbf{X}_i$. Fan, Liao & Wang (2016) extend the model to allowing components in factor loadings that are not explainable by characteristics:

$$\mathbf{b}_i = \mathbf{g}(\mathbf{X}_i) + \boldsymbol{\gamma}_i, \quad \mathbb{E}(\boldsymbol{\gamma}_i|\mathbf{X}_i) = 0. \tag{12.}$$

Here, $\mathbf{g}(\cdot)$ is a vector of nonparametric functions. With this model, they introduce an improved factor estimator, known as projected PCA.

The basic idea of projected PCA is to smooth the observations $\{y_{it}\}_{i=1}^N$ for each given t against their associated covariates $\{\mathbf{X}_i\}_{i=1}^N$ (cross-sectional smoothing) and apply PCA to the smoothed data (fitted values). Let $\{\boldsymbol{\phi}_j(\mathbf{x})\}_{j=1}^J$ be a set of basis functions. This can be either unstructured, such as kernel machines, or structured, such as a basis for additive models (Fan et al. 2020). Set $\boldsymbol{\phi}(\mathbf{X}_i)' = (\boldsymbol{\phi}_1(\mathbf{X}_i), \ldots, \boldsymbol{\phi}_J(\mathbf{X}_i))$ and $\Phi(\mathbf{X}) = (\boldsymbol{\phi}(\mathbf{X}_1), \ldots, \boldsymbol{\phi}(\mathbf{X}_N))'$, an $N \times J$ matrix. Then the projection matrix on characteristics can be taken as $\mathbf{P} = \Phi(\mathbf{X})(\Phi(\mathbf{X})'\Phi(\mathbf{X}))^{-1}\Phi(\mathbf{X})'$. The projected data $\mathbf{PY}$ is the fitted value of regressing $\mathbf{Y}$ on to the basis functions.

We make the following key assumption:

Assumption 1. We impose the following conditions.

1. **Relevance:** With probability approaching one, all the eigenvalues of $\frac{1}{N}(\mathbf{PB})'\mathbf{PB}$ are bounded away from both zero and infinity as $N \to \infty$.
2. **Orthogonality:** $\mathbb{E}(u_{it}|\mathbf{X}_i) = 0$ for all $i \le N, t \le T$.

The above conditions require that the strengths of the loading matrix should remain strong after the projection. Condition 2 implies that if we apply $\mathbf{P}$ to both sides of $\mathbf{Y} = \mathbf{BF}' + \mathbf{U}$, then

$$\mathbf{PY} \approx \mathbf{PBF}' = \mathbf{GF}',$$

where $\mathbf{G} = \mathbf{PB}$ is the $N \times r$ matrix, which $\approx (\mathbf{g}(\mathbf{X}_i))_{N\times r}$ under additional assumption $\mathbb{E}(\boldsymbol{\gamma}_i|\mathbf{X}_i) = 0$ for all $i \le N$. In other words, the noise $\mathbf{U}$ is suppressed, while signals remain. Hence, the scaled sample covariance $(\mathbf{PY})'\mathbf{PY} = \mathbf{Y}'\mathbf{PY} \approx \mathbf{FG}'\mathbf{GF}'$. For the identification purpose, let us assume $\mathbf{\Xi} := \mathbf{G}'\mathbf{G}$ is a diagonal matrix and $\mathbf{F}'\mathbf{F}/T = \mathbf{I}$. Then, from

$$\frac{1}{T}\mathbf{Y}'\mathbf{PYF} \approx \mathbf{F}\mathbf{\Xi},$$

we infer that the columns of $\mathbf{F}$ are approximately the eigenvectors of the $\mathbf{Y}'\mathbf{PY}$, scaled by a factor $\sqrt{T}$. This motivates estimating factors by using the top R eigenvectors of $\mathbf{Y}'\mathbf{PY}$.

Fan, Liao & Wang (2016) derive the rates of convergence of the projected PCA method. A nice feature is that the consistency of latent factors is achieved even when the sample size T is finite, so long as N goes to infinity. Intuitively, the idiosyncratic noise is removed from cross-sectional projections, which does not require a long time series.

Similarly, in many applications, while we do not know the latent factors $\mathbf{f}_t$, we do know that factors are related to some proxy variables $\mathbf{W}_t$. For example, the latent factors are unknown for equity markets, but they are related to Fama–French factors (Fama & French 2015); latent factors for disaggregated macroeconomics time series are unknown, but they are related to aggregated ones (McCracken & Ng 2016). Switching the roles of rows and columns, longitudinal regression of each series $\{y_{it}\}_{t=1}^T$ on $\{\mathbf{W}_t\}_{t=1}^T$ yields the projected data matrix, from which latent factors and loadings can be extracted similarly. For details on how latent factor learning is augmented by instruments $\{\mathbf{W}_t\}_{t=1}^T$, see Fan, Ke & Liao (2021).

3.5. Diversified Projection

In this section, we continue denoting R as the number of factors we use and r as the true number of factors. Fan & Liao (2020) propose a simpler factor estimator that does not rely on eigenvectors by using cross-sectional diversified projections (DPs). Let $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_R)$ be a given exogenous (or deterministic) $N \times R$ matrix, where each of its R columns $\mathbf{w}_k$ is an $N \times 1$ vector of diversified weights, the definition of which is made clear below. We estimate $\mathbf{f}_t$ by simply taking

$$\widehat{\mathbf{f}}_t = \frac{1}{N}\mathbf{W}'\mathbf{y}_t.$$

By substituting $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ into the definition, immediately we have

$$\widehat{\mathbf{f}}_t = \mathbf{H}\mathbf{f}_t + \frac{1}{N}\mathbf{W}'\mathbf{u}_t, \quad \mathbf{H} = \frac{1}{N}\mathbf{W}'\mathbf{B}. \tag{13.}$$

Thus, $\widehat{\mathbf{f}}_t$ (consistently) estimates $\mathbf{f}_t$ up to an $R \times r$ affine transform $\mathbf{H}$, with the estimation error $\mathbf{e}_t := \frac{1}{N}\mathbf{W}'\mathbf{u}_t$. The assumption that $\mathbf{W}$ should be diversified ensures that as $N \to \infty$, $\mathbf{e}_t$ is diversified away (converging to zero in probability). More specifically, we impose the following assumption.

Assumption 2. There is a constant $c > 0$, so that as $N \to \infty$. In addition, we impose the following conditions.

1. The $R \times R$ matrix $\frac{1}{N}\mathbf{W}'\mathbf{W}$ satisfies $\lambda_{\min}(\frac{1}{N}\mathbf{W}'\mathbf{W}) > c$.
2. $\mathbf{W}$ is independent of $\{\mathbf{u}_t : t \le T\}$.
3. Suppose $R \ge r$, $\operatorname{rank}(\mathbf{H}) = r$ and $\psi^2_{\min}(\mathbf{H}) \gg \frac{1}{N}$, where $\psi_{\min}(\mathbf{H})$ denotes the minimum nonzero singular value of $\mathbf{H} = \frac{1}{N}\mathbf{W}'\mathbf{B}$.

Conditions 1 and 2 define the diversified weights $\mathbf{W}$. When $(u_{1t}, \ldots, u_{Nt})$ are cross-sectionally weakly dependent, they ensure that $\mathbf{e}_t$ is diversified away. Condition 3 of Assumption 2 is a key condition: It requires that $\mathbf{W}$ should not diversify away the factor components in the time series. Several choices of $\mathbf{W}$ can be recommended to satisfy this condition. For instance, if factor loadings satisfy Equation 12, then fix R components of sieve basis functions, $(\phi_1(\cdot), \ldots, \phi_R(\cdot))$, we can define

$$\mathbf{W} := (w_{i,k})_{N\times R}, \quad \text{where } w_{i,k} = \boldsymbol{\phi}_k(\mathbf{X}_i).$$

Alternatively, we can also use transformations of the initial observation $\mathbf{x}_t$ for $t = 0$, which was considered by Juodis & Sarafidis (2020). If $\mathbf{y}_0$ is independent of $\{\mathbf{u}_t : t \ge 1\}$, we can apply $w_{i,k} = \phi_k(y_{i,0})$. These weights are correlated with $\mathbf{B}$ through $\mathbf{y}_0 = \mathbf{B}\mathbf{f}_0 + \mathbf{u}_0$.

An important benefit of the DP is that it is robust to overestimating the number of factors. Theoretical studies of factor models have been crucially depending on the assumption that the number of factors, r, should be consistently estimated. This usually requires strong conditions on the strength of factors and serial conditions. Recently, Barigozzi & Cho (2018) proposed a PCA-based method to estimate factors that are robust to overestimated r. They provide rates of convergence of the estimated common components when $R \ge r$.

Fan & Liao (2020) apply DP to several inference problems in factor-augmented models, including the postselection inference, high-dimensional covariance estimation, and factor specification tests. They formally justify the robustness to overestimating the number of factors in these applications. In particular, DP admits $r = 0$ but $R \geq 1$ as a special case. That is, the inference is still valid even if no common factors are present, but factors are nevertheless estimated for insurance. In addition, Karabiyik, Urbain & Westerlund (2019) apply DP to the context of panel data models in the presence of common factors.

3.6. Factor Estimators Robust to Heavy Tails

To apply either the PCA or the MLE to estimate the model, we need an initial covariance estimator $\mathbf{S}_y$, the application of which requires that elements of $\mathbf{y}_t$ have sufficient moments. Some technical results of factor estimations even require sub-Gaussian conditions on the data's tail distributions. However, heavy-tailed data are not uncommon in economic applications. For instance, approximately 30% of the 131 disaggregated macroeconomic variables in the work by Ludvigson & Ng (2010) have excess kurtosis greater than six, so their distributions are fatter than the t distribution with degrees of freedom five. Indeed, heavy tails are a stylized feature of high-dimensional data, as it is unlikely that all variables have sub-Gaussian tails.

Because the presence of heavy-tailed data invalidates many conditions required for estimating factor models, the recent literature has proposed several methods that are robust to the tail distributions. Here, we describe two of them: truncation and robust M-estimation.

Suppose we have independently and identically distributed (i.i.d.) data X_i, with mean μ and standard deviation σ. Consider the truncation (Winsorization) data

$$\widetilde{X}_i := \operatorname{sgn}(X_i)\min\{|X_i|, \tau\}, \tag{14.}$$

with predetermined $\tau > 0$ and estimate μ by the truncated mean. Then, Fan, Wang & Zhu (2021) show that when $\tau \asymp \sigma\sqrt{n}$, the truncated mean has Gaussian concentration:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i - \mu\right| \geq x\frac{\sigma}{\sqrt{n}}\right) \leq 2\exp(-cx^2), \quad \text{universal constant } c.$$

In contrast, without Winsorization ($\tau = 0$), it is bounded by x^{-2} by the Markov inequality. In other words, the truncated mean behaves like the sample mean from the Gaussian data, whereas the untruncated one (sample mean) has Cauchy tails.

Catoni (2012) constructs a robust M-estimator that shares the same Gaussian concentration. Fan, Li & Wang (2017) and Fan, Wang & Zhong (2019) use the Huber loss to define the mean estimator:

$$\widehat{\mu} = \arg\min_{\mu}\sum_{i=1}^{n}\rho_\tau(X_i - \mu), \qquad \rho_\tau(z) = \begin{cases} z^2/\tau, & |z| < \tau \\ 2|z| - \tau, & |z| \geq \tau, \end{cases} \tag{15.}$$

for $\tau \asymp \sigma\sqrt{n}$, called the adaptive Huber loss as it requires τ to depend on n and σ for asymmetric error distributions. These authors also establish a similar concentration inequality.

In the high-dimensional setting, suppose X_{it} is i.i.d. across t and $\max_{i\leq N}\mathbb{E}X_{it}^2 < \sigma^2$, then for both robustified methods with properly chosen tuning parameters, it can be shown that when $\log N = o(T)$, there is $C > 0$, with probability at least $1 - 4N^{-3}$,

$$\max_{i\leq N}|z_i - \mathbb{E}X_{it}| \leq C(\sigma + 1)\sqrt{\frac{\log N}{T}}, \tag{16.}$$

Supplemental Material >

where z_i is the robust estimator of $\mathbb{E}X_{it}$ using either the truncated data as in Equation 14 or the adaptive Huber estimation given in Equation 15. A formal proof of result (Equation 16) and associated conditions are presented in the online **Supplemental Materials** (Theorem B.2).

The Gaussian concentration is fundamental to high-dimensional econometrics, as we estimate many means simultaneously and the maximum estimation error accumulates slowly with Gaussian tails. It also applies to estimate covariance, as its (i, j) element is of form $\mathbb{E}y_{it}y_{jt}$. When the high-dimensional data have heavy-tailed components, we can replace the sample covariance by its robust version $\widehat{\mathbf{S}}_y$ before estimating the factors. By the Gaussian concentration inequality, the robustly estimated covariance $\widehat{\mathbf{S}}_y$ satisfies

$$\|\widehat{\mathbf{S}}_y - \mathbf{\Sigma}_y\|_{\max} = O_P\left(\sqrt{\frac{\log N}{T}}\right),$$

so long as $\mathbb{E}\, y_{it}^2 y_{jt}^2$ is uniformly bounded (and serial independence is assumed). Fan, Wang & Zhu (2021) propose another robust covariance input,

$$\hat{\mathbf{\Sigma}}_U = \frac{1}{\binom{n}{2}} \sum_{j \neq k} \min\left(\|\mathbf{y}_j - \mathbf{y}_k\|_2^2, \tau\right) \frac{(\mathbf{y}_j - \mathbf{y}_k)(\mathbf{y}_j - \mathbf{y}_k)'}{\|\mathbf{y}_j - \mathbf{y}_k\|_2^2},$$

which shares similar robust properties and is semipositive-definite. Note that $\tau = \infty$ corresponds to the sample covariance matrix.

Based on the above robust covariance inputs, we can create factor estimators and derive their theoretical properties following the guidance of Section 2.2. See chapter 10 of Fan et al. (2020) for further generalizations.

3.7. Use of Cross-Covariance

When factors are highly persistent but $\mathbb{E}\mathbf{u}_t\mathbf{u}_{t-h}^T = 0$, then the cross-covariance

$$\mathbf{\Sigma}_h = \mathbb{E}\mathbf{y}_t\mathbf{y}_{t-h}' = \mathbf{B}(\mathbb{E}\mathbf{f}_t\mathbf{f}_{t-h})\mathbf{B}', \qquad h \geq 1$$

contains valuable information about $\mathbf{B}$. This motivates one to estimate loadings by applying PCA to aggregated $\{\mathbf{\Sigma}_h : h = 1, \ldots\}$ and was studied by Lam & Yao (2012). A related idea has been extended to matrix-variate PCA (Wang, Liu & Chen 2019; Chen, Tsay & Chen 2020). Fan & Zhong (2018) also provide a procedure to efficiently aggregate the cross-covariance information with the covariance information when $h = 0$.

3.8. Which Method to Use?

Many references have documented the comparisons among various estimation methods. Westerlund & Urbain (2013) made a comparison between PCA and cross-sectional averages in the panel data setting. Meanwhile, the PCA and low-rank penalized regressions are practically very similar. So, we do not distinguish their use in practice. In general, because of the simplicity for implementations and relatively weak required conditions, the PCA still seems to be the most widely used method in applied research. Meanwhile, robust covariance inputs can also be integrated with the surveyed low-rank recovery methods.

In addition, when either factors or loadings can be partially explained by observed characteristics, the projected PCA is recommended. This is particularly useful in asset pricing applications where the explanatory power of asset characteristics has been well documented in the literature.

4. FACTOR-AUGMENTED INFERENCE AND ECONOMETRIC LEARNING

4.1. Inverse Regression Forecasts

Forecasting in a data-rich environment has been an important research topic in economics and finance. Typical examples include forecasts of the aggregate output or inflation rate using a large number of the categorized macroeconomic variables.

Stock & Watson (2002a) and Bai & Ng (2006) consider the following factor-augmented regression model for the h-step ahead forecast:

$$y_{t+h} = \boldsymbol{\alpha}'\mathbf{f}_t + \boldsymbol{\beta}'\mathbf{w}_t + \varepsilon_t \quad 17.$$

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t. \quad 18.$$

Here, $\mathbf{w}_t$ in Equation 17 is the observed predictors, which may include lagged dependent variables. Equation 18 is a high-dimensional factor model that includes a vector of latent factors $\mathbf{f}_t$. The forecast can be implemented by regressing y_{t+h} onto $\mathbf{w}_t$ and estimated factors. The factor model given in Equation 18 serves as an important dimension reduction tool.

Fan, Xue & Yao (2017) generalize Equation 17 to the nonlinear model with multiple indices. Consider the following forecasting model:

$$y_{t+1} = h(\boldsymbol{\phi}_1'\mathbf{f}_t, \ldots, \boldsymbol{\phi}_R'\mathbf{f}_t, \boldsymbol{\varepsilon}_{t+1}), \quad 19.$$

where $h(\cdot)$ is an unknown link function and $\boldsymbol{\varepsilon}_{t+1}$ is the error independent of $\mathbf{f}_t$ and $\mathbf{u}_t$. Vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_R$ are r-dimensional linear-independent prediction indices. In contrast with linear forecasting, the above model specifies that the predicting function is nonlinear and depends on multiple indices of extracted factors. If we specify $R < r$, further dimension reductions are achieved.

A prominent result related to the model in Equation 19 is given by Li (1991), whose work shows that under some regularity conditions, such as $\mathbf{f}_t$ is elliptically symmetric, we have

$$\mathbb{E}(\mathbf{f}_t|y_{t+1}) = \boldsymbol{\Phi}\mathbf{a}(y_{t+1}) \quad 20.$$

for an R-dimensional vector $\mathbf{a}(y_{t+1})$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_R]$ is an $r \times R$ matrix. In other words, the inverse regression vector $\mathbb{E}(\mathbf{f}_t|y_{t+1})$ falls in the column space spanned by $\boldsymbol{\Phi}$, which can be extracted by PCA. Indeed, since $\mathbb{E}(\mathbb{E}(\mathbf{f}_t|y_{t+1})) = \mathbb{E}(\mathbf{f}_t) = 0$,

$$\mathrm{cov}(\mathbb{E}(\mathbf{f}_t|y_{t+1})) = \boldsymbol{\Phi}\mathbb{E}[\mathbf{a}(y_{t+1})\mathbf{a}(y_{t+1})']\boldsymbol{\Phi}'.$$

The above matrix has R nonvanishing eigenvalues if $\mathbb{E}[\mathbf{a}(y_{t+1})\mathbf{a}(y_{t+1})']$ is nondegenerate. Their corresponding eigenvectors have the same linear span as $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_R$ do. If one can consistently estimate $\mathrm{cov}(\mathbb{E}(\mathbf{f}_t|y_{t+1}))$, then the subspace spanned by $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_R$, which is of our primary interests, can be obtained by extracting the top R eigenvectors of the estimated covariance matrix that correspond to the R largest eigenvalues.

However, it is not an easy task to directly estimate the covariance of $\mathbb{E}(\mathbf{f}_t|y_{t+1})$. Li (1991) suggests the sliced covariance estimate, a widely used technique for dimension reductions: The sliced covariance matrix also satisfies the fundamental property given in Equation 20, namely, $E(\mathbf{f}_t|y_{t+1} \in \mathbf{I}_k)$ falls in the column space spanned by $\boldsymbol{\Phi}$ for any given partition of the range of y_{t+1} into H slices $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_H$. Correspondingly, let

$$\widehat{\mathrm{cov}(\mathbb{E}(\mathbf{f}_t|y_{t+1}))} = \frac{1}{H}\sum_{h=1}^{H}\left[\frac{1}{\sum_{t=1}^{T}1(y_{t+1} \in \mathbf{I}_h)}\sum_{t=1}^{T}\mathbf{f}_t 1(y_{t+1} \in \mathbf{I}_h)\right] \times \left[\frac{1}{\sum_{t=1}^{T}1(y_{t+1} \in \mathbf{I}_h)}\sum_{t=1}^{T}\mathbf{f}_t 1(y_{t+1} \in \mathbf{I}_h)\right]', \quad 21.$$

which is a nonparametric covariance estimator. The above sliced covariance estimator is based on the observable factors. If the factors are unknown, they are replaced by their estimators, which leads to the following sufficient forecasting algorithm based on the factor models.

Algorithm 1. Sufficient forecasting algorithm based on the factor models.

Step 1. Estimate factors in the model given in Equation 18 for $t = 1, \ldots, T$.
Step 2. Construct the covariance estimator as in Equation 21, with $\widehat{\mathbf{f}}_t$ in place of $\mathbf{f}_t$.
Step 3. Obtain $\widehat{\boldsymbol{\phi}}_1, \widehat{\boldsymbol{\phi}}_2, \ldots, \widehat{\boldsymbol{\phi}}_R$ by the top R eigenvectors of the covariance in step 2.
Step 4. Construct the predictive indices $\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_t, \ldots, \widehat{\boldsymbol{\phi}}_R'\widehat{\mathbf{f}}_t$.
Step 5. Nonparametrically estimate $b(\cdot)$ with indices from step 4, and forecast y_{t+1}.

Implementing the above algorithm requires the number of slices H, the number of predictive indices R, and the number of factors r. In practice, H has little influence on the estimated directions, as pointed out by Li (1991) and explained above that Equation 20 holds. In regard to the choice of R, the first R eigenvalues of $\mathrm{cov}(\mathbb{E}(\mathbf{f}_t|y_{t+1}))$ must be significantly different from zero compared with the estimation error. Several methods such as those from Li (1991) and Schott (1994) have been proposed to determine R. For instance, the average of the smallest $r - L$ eigenvalues would follow χ^2 distribution if the underlying factors are normally distributed. The number of factors can be determined by a number of methods.

4.2. Factor-Adjusted Regularized Model Selection

Consider a high-dimensional regression model

$$\begin{aligned} y_t &= \boldsymbol{\beta}'\mathbf{g}_t + \boldsymbol{\nu}'\mathbf{x}_t + \eta_t, \\ \mathbf{g}_t &= \boldsymbol{\theta}'\mathbf{x}_t + \boldsymbol{\varepsilon}_{g,t}, \end{aligned} \tag{22.}$$

where $\mathbf{g}_t$ is a treatment variable whose effect $\boldsymbol{\beta}$ is of the main interest. The model contains high-dimensional exogenous control variables $\mathbf{x}_t = (x_{1t}, \ldots, x_{Nt})$ that determine both the outcome and the treatment variables. Having many control variables creates challenges for statistical inferences; as such, we assume that $(\boldsymbol{\nu}, \boldsymbol{\theta})$ are sparse vectors.

Control variables are often strongly correlated due to the presence of confounding factors

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t. \tag{23.}$$

This invalidates conditions of using penalized regressions to directly select among $\mathbf{x}_t$. Instead, if we substitute Equation 23 into Equation 22, we reach a factor-adjusted regression model:

$$\begin{aligned} y_t &= \boldsymbol{\alpha}_y'\mathbf{f}_t + \boldsymbol{\gamma}'\mathbf{u}_t + \boldsymbol{\varepsilon}_{y,t}, \\ \mathbf{g}_t &= \boldsymbol{\alpha}_g'\mathbf{f}_t + \boldsymbol{\theta}'\mathbf{u}_t + \boldsymbol{\varepsilon}_{g,t}, \\ \boldsymbol{\varepsilon}_{y,t} &= \boldsymbol{\beta}'\boldsymbol{\varepsilon}_{g,t} + \eta_t, \end{aligned} \tag{24.}$$

where $\boldsymbol{\alpha}_g' = \boldsymbol{\theta}'\mathbf{B}$, $\boldsymbol{\alpha}_y' = \boldsymbol{\beta}\boldsymbol{\alpha}_g' + \boldsymbol{\nu}'\mathbf{B}$, and $\boldsymbol{\gamma}' = \boldsymbol{\beta}\boldsymbol{\theta}' + \boldsymbol{\nu}'$. Here, $(\boldsymbol{\alpha}_y, \boldsymbol{\alpha}_g, \boldsymbol{\beta})$ are low-dimensional coefficient vectors, while $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ are high-dimensional sparse vectors. Importantly, the model contains high-dimensional latent controls $\mathbf{u}_t$, which are weakly dependent due to the nature of idiosyncratic noises. The use of $\mathbf{u}_t$ instead of $\mathbf{x}_t$ validates conditions for many high-dimensional variable selection methods.

Fan, Ke & Wang (2020) and Hansen & Liao (2018) show that the penalized regression can be successfully applied to Equation 24 to select components in $\mathbf{u}_t$, which are cross-sectionally

weakly correlated. Motivated by Belloni, Chernozhukov & Hansen (2014), the algorithm can be summarized as follows. For notational simplicity, we focus on the univariate case $\dim(\boldsymbol{\beta}) = 1$.

Algorithm 2. Estimate $\boldsymbol{\beta}$ as follows.

Step 1. Estimate $\{(\mathbf{f}_t, \mathbf{u}_t) : t \le T\}$ from Equation 23 to obtain $\{(\widehat{\mathbf{f}}_t, \widehat{\mathbf{u}}_t) : t \le T\}$.
Step 2. Run penalized variable selections on $\widehat{\mathbf{u}}_t$:

$$(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\alpha}}_y) = \arg\min_{\boldsymbol{\gamma}, \alpha_y} \frac{1}{T} \sum_{t=1}^{T} (y_t - \boldsymbol{\alpha}_y' \widehat{\mathbf{f}}_t - \boldsymbol{\gamma}' \widehat{\mathbf{u}}_t)^2 + P_\tau(\boldsymbol{\gamma}),$$

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}_g) = \arg\min_{\boldsymbol{\theta}} \frac{1}{T} \sum_{t=1}^{T} (\mathbf{g}_t - \boldsymbol{\alpha}_g' \widehat{\mathbf{f}}_t - \boldsymbol{\theta}' \widehat{\mathbf{u}}_t)^2 + P_\tau(\boldsymbol{\theta}).$$

Obtain residuals: $\widehat{\boldsymbol{\varepsilon}}_{y,t} = y_t - (\widehat{\boldsymbol{\alpha}}_y' \widehat{\mathbf{f}}_t + \widehat{\boldsymbol{\gamma}}' \widehat{\mathbf{u}}_t)$ and $\widehat{\boldsymbol{\varepsilon}}_{g,t} = \mathbf{g}_t - (\widehat{\boldsymbol{\alpha}}_g' \widehat{\mathbf{f}}_t + \widehat{\boldsymbol{\theta}}' \widehat{\mathbf{u}}_t)$.
Step 3. Estimate $\boldsymbol{\beta}$ by residual-regression: $\widehat{\boldsymbol{\beta}} = (\sum_{t=1}^{T} \widehat{\boldsymbol{\varepsilon}}_{g,t}^2)^{-1} \sum_{t=1}^{T} \widehat{\boldsymbol{\varepsilon}}_{g,t} \widehat{\boldsymbol{\varepsilon}}_{y,t}$.

Note that $\boldsymbol{\gamma} :\to P_\tau(\boldsymbol{\gamma})$ is a sparse-induced penalty function with a tuning parameter τ. When $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are sufficiently sparse and the principal components estimator is used in step 1 with the correct selection of the number of factors, the above procedure is asymptotically valid:

$$\sigma_{\eta,g}^{-1} \sigma_g^2 \sqrt{T} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, 1), \tag{25.}$$

where σ_g^2 and $\sigma_{\eta,g}^2$ are the asymptotic variances of $\boldsymbol{\varepsilon}_{g,t}$ and $\eta_t \boldsymbol{\varepsilon}_{g,t}$.

More recently, Fan & Liao (2020) show that the assumption of correct selection of the number of factors can be relaxed if we use the DP in step 1 instead, and Equation 25 is still valid as long as we select $R \ge r$ factors (over selection). Importantly, this admits $r = 0$ and $R \ge 1$ as a special case, i.e., there are no factors so that $\mathbf{x}_t = \mathbf{u}_t$ itself is cross-sectionally weakly dependent, but nevertheless we estimate $R \ge 1$ number of factors to run postselection inference to alleviate the dependence among $\mathbf{x}_t$. This setting is empirically relevant, as it allows one to avoid pretesting the presence of common factors for inference.

Figure 1, taken from Fan & Liao (2020), plots the histograms of the t-statistics based on estimated $\boldsymbol{\beta}$ over 200 simulations, superimposed with the standard normal density, where R DPs are used to estimate factors in step 1. Here, the weights are the initial transformations ($t = 0$), so that the ith row of $\mathbf{W}$ is $(x_{it}, x_{it}^2, \ldots, x_{it}^R)$ at $t = 0$. The double selection is the algorithm used by Belloni, Chernozhukov & Hansen (2014) that directly selects among $\mathbf{x}_t$, corresponding to the case $R = 0$. The factor-augmented algorithm works well even if $r = 0$; however, when $r \ge 1$ factors are present, double selection leads to severely biased estimations.

Therefore, as a practical guidance, we recommend that one should always run factor-augmented postselection inference, with $R \ge 1$, to guard against confounding factors among the control variables.

4.3. Factor-Adjusted Robust Multiple Testing

Large-scale multiple testing finds many applications in economics and finance. The test statistics are frequently dependent and should be adjusted in order to control the false discovery rate and gain the power of the tests.

4.3.1. False discovery rate control. Controlling the false discovery proportion (FDP) in large-scale hypothesis testing based on strongly dependent tests has been an important problem in many scientific discoveries across disciplines. For applications in empirical asset pricing, readers are

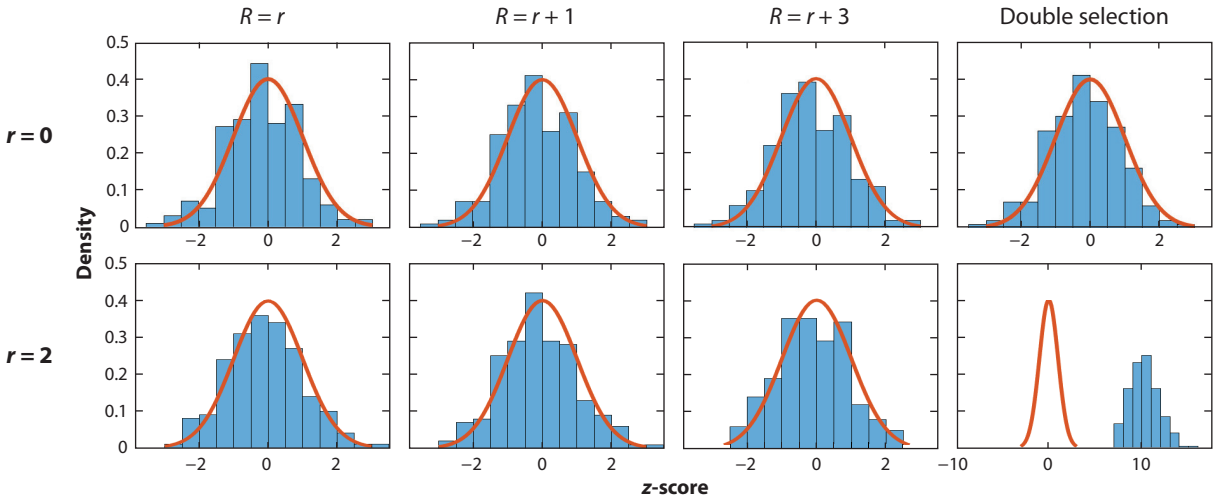

Figure 1

Histograms of the standardized estimates in Equation 25 over 200 replications, superimposed with the standard normal density (*red lines*). The panels showing double selection correspond to directly selecting among control variable $\mathbf{x}_t$ (corresponding to $R = 0$, no factor adjustment), while all other panels correspond to using diversified factor estimators with R number of working factors. The top four panels correspond to $r = 0$, and the bottom four panels correspond to $r = 2$. When $R \geq r$, Equation 25 holds, whereas when $R < r$, Equation 25 is violated. R denotes as the number of factors we used, and r denotes as the true number of factors. The z-score refers to the t-statistics. Figure adapted with permission from Fan & Liao (2020).

referred to works by Barras, Scaillet & Wermers (2010); Harvey, Liu & Zhu (2015); Harvey & Liu (2018); Fan et al. (2019a) and references therein; and Giglio, Liao & Xiu (2021).

Suppose we observe realizations of a random vector $\{\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'\}_{t=1}^T$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ denote its mean vector. We are interested in testing individual hypotheses:

$$H_0^i : \alpha_i = 0, \quad i = 1, \ldots, N.$$

Let p_i denote the p-value for testing H_0^i based on a test statistic such as a t-test, which rejects if $p_i < x$ given some critical value x. Define the number of false discoveries (rejections) and the total number of rejections as follows:

$$\mathcal{F}(x) = \sum_{i=1}^N 1\{i : p_i < x \text{ and } H_0^i \text{ is true}\}, \qquad \mathcal{V}(x) = \sum_{i=1}^N 1\{i : p_i < x\}.$$

In large-scale multiple testing problems, researchers often aim to control the FDP and the false discovery rate (FDR), defined by

$$\text{FDP}(x) = \frac{\mathcal{F}(x)}{\max\{\mathcal{V}(x), 1\}}, \quad \text{FDR}(x) = \mathbb{E}\{\text{FDP}(x)\}.$$

The goal is to find the critical value x so that $\text{FDR}(x) \leq \tau$ for a desired level τ (e.g., 0.10), or more relevantly, $\text{FDP}(x) \leq \tau$ with high confidence. While $\mathcal{V}(x)$ is known, $\mathcal{F}(x)$ is not in practice. A general principle of finding x proceeds as the following two steps.

Algorithm 3. General principle for FDP/FDR control.

Step 1. Find $\bar{\mathcal{F}}(x)$, such that either it upper bounds $\mathcal{F}(x)$ for all $x \in (0, 1)$ or it estimates $\mathcal{F}(x)$ uniformly well.

Step 2. Set the critical value to $x^* = \sup\{x \in (0, 1) : \bar{\mathcal{F}}(x) \leq \tau \max\{\mathcal{V}(x), 1\}\}$.

One of the most popular procedures, proposed by Benjamini & Hochberg (1995), proceeds as follows. Denote $p_{(1)} \le \cdots \le p_{(N)}$ as the sorted p-values for the individual tests. Then the critical value is set to

$$x^* = \max\{p_{(i)} : p_{(i)} \le \tau i/N\}.$$

This method fits into Algorithm 3 with $\bar{\mathcal{F}}(x) = Nx$, which is an asymptotic upper bound for $\mathcal{F}(x)$ when the individual p-values are independent. One of the limitations of this upper bound is that it is too conservative if the number of true negatives is small compared with N. More fundamentally, it requires the test statistics be weakly dependent, a topic we shall discuss in more detail next. Other methods, such as those by Storey (2002), Fan, Han & Gu (2012), etc., aim to directly estimate $\mathcal{F}(x)$ in step 1 in the presence of strong dependence among test statistics and are also adaptive to the unknown number of true negatives. In addition, instead of Algorithm 3, Romano & Wolf (2007) and Romano, Shaikh & Wolf (2008) provide alternative procedures for FDR control.

4.3.2. Removing dependence by factor adjustments. The key to the success of FDR control is that the individual test statistic should be either weakly dependent or independent. This makes the FDR and FDP approximately the same and easier to control. Conversely, suppose the cross-sectional dependence of $\mathbf{y}_t$ is generated from a latent factor model:

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad \mathbb{E}(u_t|\mathbf{f}_t) = 0, \tag{26.}$$

where $\mathbb{E}\,\mathbf{f}_t = 0$ and $\boldsymbol{\alpha}$ is the mean vector. In empirical asset pricing, the model can be used to identify nonzero alphas out of a large number of assets and has been studied to identify skilled mutual fund managers, e.g., Barras, Scaillet & Wermers (2010) and Harvey, Liu & Zhu (2015). The presence of latent factors, however, leads to strong dependence among the t-statistics based on the naive sample means of $\mathbf{y}_t$, which invalidates the weak dependence assumptions. As well documented in the literature, strong dependence creates fundamental challenges to multiple testing, including large standard errors among the estimated α_i, unstable FDPs, and conservativeness of the test procedure. Learning dependence $\mathbf{B}\mathbf{f}_t$ and removing it from the model given in Equation 26 make the data not only weakly dependent but also less noisy (from $\mathbf{B}\mathbf{f}_t + \mathbf{u}_t$ to $\mathbf{u}_t$). This is the basic idea in factor-adjusted robust multiple tests (FarmTest) by using factor-adjusted data $\{\mathbf{y}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t\}_{t=1}^T$ (see Equation 26). Furthermore, Fan et al. (2019a) make adjustments so that it is also robust to heavy-tailed data.

To illustrate consequences of omitting adjusting latent factors as well as the effectiveness of the use of the factor-adjusted method (to be detailed below), let us consider a numerical example of a single factor model, where elements of $\mathbf{u}_t$, $\mathbf{f}_t$, and $\mathbf{B}_t$ are generated from the standard normal distribution. We take the true means to be $\alpha_i = 0.6$ for $1 \le i \le N/4$ and 0 otherwise, and we compare two estimated α_i: (*a*) the sample means of $\mathbf{y}_t$, without using factor adjustments and (*b*) the factor-adjusted estimator based on PCA. We apply the method of Benjamini & Hochberg (1995) for multiple testing, setting $\tau = 0.05$.

Figure 2*a*,*b* plots the histograms, from a single simulation, of the estimators for α_i, corresponding to those that satisfy the null hypotheses $\alpha_i = 0$ and those that satisfy the alternatives $\alpha_i = 0.6$. Clearly, there is a large overlap (*panel a*) between sample means from the null and the alternative, making it difficult to distinguish the alternatives from the nulls in tests based on sample means. In contrast, the PCA-based estimator can easily separate the nulls and alternatives, as shown in **Figure 2*b***.

Figure 2*c*,*d* plots the histograms of the true FDP over 1,000 simulations based on the two estimators. It is evident that the distribution of the FDP corresponding to the factor-adjusted

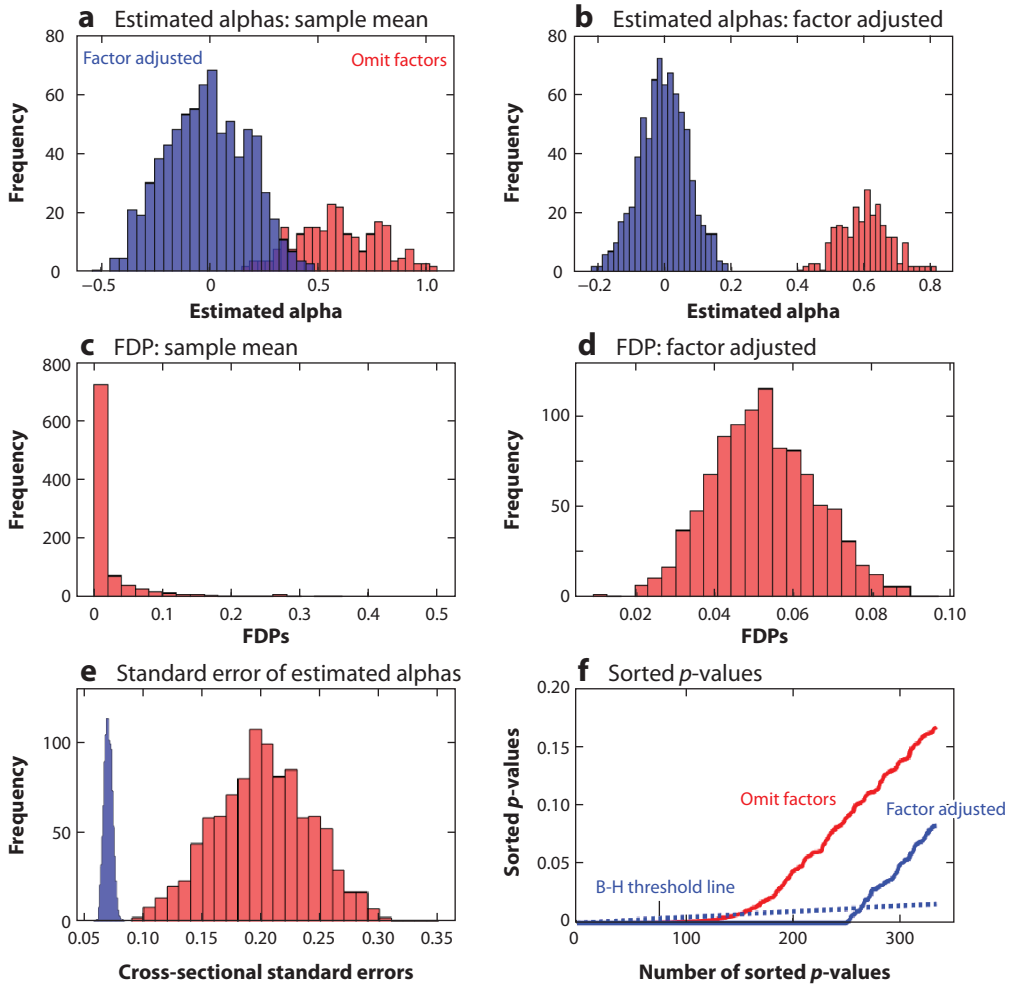

Figure 2

Comparison between the sample mean method (omit factors, *red*) and the factor-adjusted method (*blue*), with $T = 200$ and $N = 1{,}000$. Panels *a–d* plot the histograms of estimated individual alphas from a single simulation; panels *c–d* plot the individual false discovery proportions (FDPs) over 1,000 simulations. Panel *e* plots the cross-sectional histograms of standard errors of the estimated alphas over 1,000 simulations. Panel *f* plots the sorted *p*-values from a single simulation. The B-H threshold line refers to the Benjamini-Hochberg procedure. The B-H procedure rejects all the hypotheses if $p_{(i)}$ is below the B-H threshold line $f(i) := \tau i/N$.

estimator concentrates around the nominal level. In contrast, the one based on the sample mean has a noticeable long tail as well as a larger mean and variance, which demonstrates the challenge to control FPD in the presence of common factors, as explained above.

Finally, omitting confounding factors would lead to larger standard errors and conservative inference. **Figure 2*e,f*** plots the standard errors of individual estimated alphas and the sorted *p*-values for the two estimation methods. The sample-mean estimator has much fewer sorted *p*-values below the B-H threshold line (i.e., fewer rejections), compared with the factor-adjusted estimator. Hence, estimating and removing the latent factors is recommended before applying standard FDR control algorithms.

4.3.3. Identifying skilled hedge funds. Giglio, Liao & Xiu (2021) study the problem of identifying hedge funds that are able to produce positive alphas (i.e., have skill), among thousands of existing funds. They consider a linear pricing model, where hedge fund returns are

$$y_{it} = \alpha_i + \mathbf{b}_i'\boldsymbol{\lambda} + \mathbf{b}_i'(\mathbf{f}_t - \mathbb{E}\mathbf{f}_t) + u_{it}.$$

In the model, $\mathbf{f}_t$ contains both observable and latent factors. The model allows nontradable observable factors, and $\boldsymbol{\lambda}$ is the vector of factor risk premia.

At a broad level, their methodology proceeds as the Fama-MacBeth regression integrated with the PCA to extract latent factors:

Algorithm 4. Estimating alphas in the presence of latent and nontradable factors.

Step 1. Run fund-by-fund time series regressions to estimate fund exposures (betas) to observable factors.
Step 2. Apply PCA to the residuals to recover the latent factors and betas.
Step 3. Implement cross-sectional regressions like Fama-MacBeth to estimate the risk premia of the factors (including both observable and latent factors) and the alphas.

Because of many negative alphas from unskilled fund managers, the multiple testing problem should be properly formulated as one-sided hypotheses:

$$H_0^i : \alpha_i \le 0, \quad i = 1, \ldots, N.$$

Hence, rejecting H_0^i indicates skilled fund manager i. Conversely, the existence of potentially a very large number of negative alphas gives rise to the issue of power loss, only to add noises to the model. The loss of power associated with testing inequalities is well known as the problem of "deep in the null" and is often seen in the econometric literature. To address this issue, Giglio, Liao & Xiu (2021) propose to first screen off very bad funds, identified as:

$$\mathcal{I} = \{i \le N, \widehat{\alpha}_i/\mathrm{se}(\widehat{\alpha}_i) < -c_{NT}\},$$

where $c_{NT} > 0$ is a slowly growing sequence to ensure sure screening (Fan & Lv 2008): $P(\mathcal{I} \subseteq \mathcal{H}_0) \to 1$. They recommend applying FDR control algorithms on funds outside $\mathcal{I}$. Therefore, two ingredients are recommended for identifying skilled fund managers via multiple testing: (*a*) adjust the effect of latent factors and (*b*) remove the estimated alphas that are deep in the null. Both are playing the essential role of gaining good testing power.

4.4. Threshold Regression with Mixed Integer Optimization

Threshold regressions have been used in economic applications to capture potential structural changes on regression coefficients. The early literature models the threshold effect using some observable scalar variable q_t, as in

$$y_t = \mathbf{w}_t'\boldsymbol{\beta} + \mathbf{w}_t'\boldsymbol{\delta}1\{q_t > \gamma\} + \varepsilon_t,$$

where $\mathbf{w}_t$ and q_t are adapted to the filtration $\mathcal{F}_{t-1}$, $(\boldsymbol{\beta}, \boldsymbol{\delta}, \gamma)$ is a vector of unknown parameters, and ε_t satisfies the conditional mean restriction. Hence, when $q_t > \gamma$, the regression function becomes $\mathbf{w}_t'(\boldsymbol{\beta} + \boldsymbol{\delta})$; when $q_t \le \gamma$, it reduces to $\mathbf{w}_t'\boldsymbol{\beta}$ (Chan 1993, Hansen 2000). In practice, it might be controversial to choose which observed variable plays the role of q_t. For example, if the two different regimes represent the status of two environments of the population, arguably it is difficult to assume that the change of the environment is governed by just a single variable.

Seo & Linton (2007) and Lee et al. (2021) extend the model to multivariate threshold:

$$y_t = \mathbf{w}_t'\boldsymbol{\beta} + \mathbf{w}_t'\boldsymbol{\delta}1\{\boldsymbol{\gamma}'\mathbf{f}_t > 0\} + \varepsilon_t,$$

where $\mathbf{f}_t$ is a vector of factors and $\boldsymbol{\gamma}$ is the corresponding unknown coefficients. So the model introduces a regime change due to a single index of factors. Allowing multivariate thresholding is important, because it permits the structural change to be governed by a potentially much larger data set: $\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$, where $\dim(\mathbf{x}_t) = N \to \infty$. So, $\mathbf{f}_t$ can be unobserved factors that can be learned from $\mathbf{x}_t$. For the identification purpose, suppose $\frac{1}{T}\sum_t \mathbf{f}_t\mathbf{f}_t' = \mathbf{I}$ and $\mathbf{B}'\mathbf{B}$ is diagonal, then $\boldsymbol{\gamma}$ and $\mathbf{f}_t$ are separately identified. This gives rise to the factor-driven two-regime regression model.

A natural strategy to estimate the model is to rely on least squares:

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\gamma}} \sum_{t=1}^{T} (y_t - \mathbf{w}_t'\boldsymbol{\beta} - \mathbf{w}_t'\boldsymbol{\delta} 1\{\boldsymbol{\gamma}'\widehat{\mathbf{f}}_t > 0\})^2,$$

where $\widehat{\mathbf{f}}_t$ is the plugged-in principal components estimator of factors. Because the least squares problem is neither convex nor smooth in $\boldsymbol{\gamma}$, the computational task is demanding. Lee et al. (2021) recommend using algorithms based on mixed integer optimization (MIO). Introduce integers $d_t := 1\{\boldsymbol{\gamma}'\widehat{\mathbf{f}}_t > 0\} \in \{0, 1\}$. The goal is to introduce linear constraints with respect to variables of optimization. Suppose there are known upper and lower bounds for δ_j: $L_j \le \delta_j \le U_j$, where δ_j denotes the jth element of $\boldsymbol{\delta}$. Define $M_t \equiv \max_{\boldsymbol{\gamma}\in\Gamma} |\boldsymbol{\gamma}'\widehat{\mathbf{f}}_t|$, where Γ is the parameter space for $\boldsymbol{\gamma}$. Then it can be verified that the least squares problem is numerically equivalent to the following constraint MIO problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\gamma},\mathbf{d},\boldsymbol{\ell}} \sum_{t=1}^{T} (y_t - \mathbf{w}_t'\boldsymbol{\beta} - \mathbf{w}_t'\boldsymbol{\ell}_t)^2, \tag{27.}$$

subject to the following constraints: for any $\epsilon > 0$, for each $t = 1, \ldots, T$ and each $j = 1, \ldots, \dim(\mathbf{w}_t)$,

$$\begin{aligned} &\boldsymbol{\gamma} \in \Gamma,\ d_t \in \{0, 1\},\ L_j \le \delta_j \le U_j, \\ &(d_t - 1)(M_t + \epsilon) < \boldsymbol{\gamma}'\widehat{\mathbf{f}}_t \le d_t M_t, \\ &d_t L_j \le \ell_{j,t} \le d_t U_j, \\ &L_j(1 - d_t) \le \delta_j - \ell_{j,t} \le U_j(1 - d_t). \end{aligned} \tag{28.}$$

Then, we can apply modern MIO packages (e.g., Gurobi) to solve for the optimal $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma})$.

Finally, Lee et al. (2021) also derive the asymptotic distribution of the estimated coefficients and propose inferences based on bootstraps. Under the condition that $T = O(N)$, they show that the effect of estimating factors is negligible on the asymptotic distribution of the estimated $(\boldsymbol{\beta}, \boldsymbol{\delta})$ but would affect both the rate of convergence and the limiting distribution of the estimated $\boldsymbol{\gamma}$.

4.5. Community Detection

The stochastic block model has been a popular approach to modeling networks (for a recent review, see Abbe 2017). We observe a graph of N nodes. Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N\times N}$ be the adjancy matrix of edges, so that $a_{ij} = 1$ if nodes i and j are connected, and $a_{ij} = 0$ otherwise. Suppose each node belongs to one of r communities and the community that node i belongs to is denoted by an unknown $\pi_i \in \{1, \ldots, r\}$. In addition, elements of $\mathbf{A}$ are random variables. Then, the stochastic block model assumes that

$$P(a_{ij} = 1 | \pi_i = k, \pi_j = l) = w_{k,l},$$

where $w_{k,l}$ is an unknown probability. We observe the matrix $\mathbf{A}$ and aim to recover the membership π_i and the probabilities $w_{k,l}$ for all $k, l = 1, \ldots, r$.

Let $\mathbf{e}_1, \ldots, \mathbf{e}_r$ denote the canonical basis in $\mathbb{R}^r$, and $\mathbf{b}_i = \mathbf{e}_k$ where $\theta_i = k$. Then, $\mathbf{b}_i$ indicates the community membership of node i, and the membership matrix is

$$\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_N)', \quad N \times r,$$

whose rows represent nodes and columns represent communities. Let $\mathbf{W}$ denote the $r \times r$ matrix of $(w_{k,l})$, and let $\mathbf{L} := \mathbb{E}\mathbf{A}$. It can easily be seen that $\mathbf{L} = \mathbf{BWB}'$ is a low-rank matrix, whose rank equals r, leading to the following low-rank decomposition:

$$\mathbf{A} = \mathbf{L} + \mathbf{S}, \quad \mathbf{S} = \mathbf{A} - \mathbb{E}\mathbf{A}.$$

Therefore, $\mathbf{A}$ has the familiar decomposition, shown in Equation 4, with $\mathbf{L}$ being similar to the systematic risk and $\mathbf{B}$ as a low-rank loading matrix. Since the elements in $\mathbf{S}$ are independent with mean-zero (Wigner matrix), the operator norm $\|\mathbf{S}\|$ does not grow too fast, compared with that of $\mathbf{L}$. We can then apply PCA to $\mathbf{A}$ to estimate $\mathbf{B}$. Suppose r is known, then the estimator $\widehat{\mathbf{B}}$ is defined as $\sqrt{N}$ times the eigenvectors of $\mathbf{A}$, corresponding to the first r eigenvalues.

Theorem 1 can be applied to obtain a deviation bound for the estimated loading matrix. If there is a sequence $g_N \to \infty$ and constants $c_1, \ldots, c_r > 0$ such that the eigenvalues $\lambda_i(\mathbf{W}^{1/2}\mathbf{B}'\mathbf{B}\mathbf{W}^{1/2}) = c_i g_N(1 + o_P(1))$ for all $i \le r$, then there is an $r \times r$ matrix $\mathbf{H}$, so that

$$\|\widehat{\mathbf{B}} - \mathbf{BH}\|_\infty = O_P(g_N^{-2} N\|\mathbf{S}\| + g_N^{-1}\sqrt{N \log N}).$$

Therefore, elements of a rotated $\mathbf{B}$ can be estimated uniformly well. Moreover, because each community has many nodes to belong to, $\mathbf{BH}$ has many identical rows, which makes the cluster analysis a natural method for community detections. For instance, we can apply either the K-means cluster analysis or the homogeneous pursuit from Ke, Fan & Wu (2015) on the rows of $\widehat{\mathbf{B}}$ to consistently identify the communities.

5. UNBALANCED PANELS

Missing data and unbalanced panels are not uncommon in economic and financial studies. Addressing the missing data issue in statistical modeling belongs to a larger category of problems, known as matrix completion. Low-rank matrix completion refers to the problem of recovering missing entries from low-rank matrices. It is particularly relevant to empirical asset pricing factor models, because many time series of returns have short histories or missing records. In this section, we review several methods for matrix completions, which assume that the missing is at random, except for the work by Cai, Cai & Zhang (2016), Bai & Ng (2019b), and Fan & Kim (2019). In addition, the expectation-maximization (EM) algorithm is a classical approach to dealing with unbalanced panels. For detailed discussions on related issues, we refer the reader to works by Stock & Watson (2002b), Su, Miao & Jin (2019), and Zhu, Wang & Samworth (2019).

5.1. Inverse Probability Weighting

Recall that the covariance matrix of $\mathbf{y}_t$, under the factor model given in Equation 7, has the following decomposition, $\mathbf{\Sigma}_y = \mathbf{B}\,\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \mathbf{\Sigma}_u$, where columns of $\mathbf{B}$ are approximately equal to the eigenvectors of $\mathbf{\Sigma}_y$ corresponding to the first r eigenvalues. As such, let $\widehat{\mathbf{\Sigma}}_y$ be an input matrix, serving as an estimator for $\mathbf{\Sigma}_y$. Then, as described in Section 2.2, we can estimate the space spanned by $\mathbf{B}$ using the leading eigenvectors of $\widehat{\mathbf{\Sigma}}_y$.

In the presence of missing data with exogenous missing, let $x_{it} = 1\{y_{it} \text{ is observed}\}$, and we only observe $y_{it}x_{it}$ for all (i, t), in which unobserved data are set to zero. Suppose for now that

$w_i := P(x_{it} = 1)$ is known. We can construct an unbiased estimator $\widehat{\mathbf{\Sigma}}_y = (\widehat{\sigma}_{ij})$, with

$$\widehat{\sigma}_{ij} := \frac{1}{w_i w_j T} \sum_{t=1}^{T} y_{it} y_{jt} x_{it} x_{jt}.$$

In the matrix form, let $\mathbf{Y}$ and $\mathbf{X}$ be the $N \times T$ matrices of y_{it} and x_{it}. So, we only observe $\mathbf{Y} \circ \mathbf{X}$, where $\circ$ represents the element-wise matrix product, the Hadamard product. Also, let $\mathbf{W}$ be the diagonal matrix, with w_i being its ith diagonal entry. Then,

$$\widehat{\mathbf{\Sigma}}_y = \frac{1}{T}\mathbf{Z}\mathbf{Z}', \quad \mathbf{Z} := \mathbf{W}^{-1}(\mathbf{Y} \circ \mathbf{X}).$$

Therefore, columns of the loading matrix estimator $\widehat{\mathbf{B}}$ are equal to $\sqrt{N}$ times the top right singular vectors of $\mathbf{Z}$. This method simply replaces the missing entries of $\mathbf{Y}$ by zero and applies the inverse probability weighting (IPW) before applying PCA. The IPW has been popularly used in the causal inference literature (e.g., Imbens & Rubin 2015). Here, the same idea is applied to create an unbiased estimator for the covariance matrix.

In practice, we shall replace w_i by its consistent estimators, such as $\widehat{w}_i := \frac{1}{T}\sum_{t=1}^{T} x_{it}$. But in the case of homogeneous missing, that is, $w_1 = \cdots = w_N$, the IPW is not needed, because $\mathbf{W}$ equals the identity matrix up to a constant, which does not affect the PCA on $\mathbf{Y} \circ \mathbf{X}$. In addition, factors can be further estimated using least squares by regressing $y_{it}x_{it}$ on the estimated loadings.

Theoretical properties are studied by Abbe et al. (2020) and Su, Miao & Jin (2019) under the assumption of homogenous missing. Su, Miao & Jin (2019) use this estimator as their initial value for the EM algorithm. Xiong & Pelger (2019) allow heterogenous missing and prove that the estimators are also asymptotically normal (they estimate $w_i w_j$ directly by $\frac{1}{T}\sum_{t=1}^{T} x_{it}x_{jt}$). We can also quickly derive the rate of convergence by applying Theorem 1. However, the IPW is the least efficient approach among all the methods to be discussed in this section. We shall verify this in a simulation study in Section 5.5.

5.2. Regularized Matrix Completion

Regularized matrix completion is a powerful technique to recover missing entries from low-rank matrices. This approach is also much faster than the EM algorithm in handling large panels. Due to these nice properties, it has also attracted much attention in the recent econometrics literature (e.g., Athey et al. 2018; Moon & Weidner 2018; Bai & Ng 2019a; Giglio, Liao & Xiu 2021).

In the matrix form $\mathbf{Y} = \mathbf{M} + \mathbf{U}$, the goal is to recover the factor component $\mathbf{M} = \mathbf{B}\mathbf{F}'$ when $\mathbf{Y}$ has missing elements. The nuclear-norm regularization is directly applicable:

$$\widehat{\mathbf{M}} := \arg\min_{\mathbf{M}} \|(\mathbf{Y} - \mathbf{M}) \circ \mathbf{X}\|_F^2 + \lambda \|\mathbf{M}\|_n, \tag{29.}$$

with tuning parameter λ. The factors and loadings can be estimated by taking the singular vectors of $\widehat{\mathbf{M}}$. Negahban & Wainwright (2011) and Koltchinskii, Lounici & Tsybakov (2011) derive the rate of convergence under the Frobenius norm. Under suitable conditions (e.g., missing at random, RSC, sufficiently large noise), it can be proved that

$$\frac{1}{NT}\|\widehat{\mathbf{M}} - \mathbf{M}\|_F^2 = O_P\left(\frac{1}{T} + \frac{1}{N}\right).$$

Chen et al. (2020a) certifies further that the convex optimization found in Equation 29 is optimal for all noise levels under the Frobenius norm, operator norm, and element-wise infinity norm. The proof is based on a novel technical device that bridges the convex optimization with a nonconvex optimization problem. However, this estimator is not asymptotically normal due to the presence of shrinkage bias and thus is not suitable for statistical inferences.

5.3. Debiased Estimators

Recent progress in this literature focuses on debiasing the regularized regression in order to have valid confidence intervals (e.g., Chen et al. 2019; Chernozhukov et al. 2019; Xia & Yuan 2019). When the missing is homogeneous, $P(x_{it} = 1) = p$ for all (i, t), Chen et al. (2019) propose the following simple debiased estimator:

$$\widehat{\mathbf{M}}^d = H_R(\widehat{\mathbf{M}} + \widehat{p}^{-1}(\mathbf{Y} - \widehat{\mathbf{M}}) \circ \mathbf{X}), \tag{30.}$$

where $H_R(\cdot)$ is the best rank R approximation in Equation 5, $\widehat{\mathbf{M}}$ is given by Equation 29, and $\widehat{p}$ is the sample proportion of missing data. The idea is very intuitive. Ignoring the weak dependence between $\widehat{\mathbf{M}}$ and $\mathbf{X}$ and estimating error in $\widehat{p}$, we have

$$\mathbb{E}(\widehat{\mathbf{M}} + \widehat{p}^{-1}(\mathbf{Y} - \widehat{\mathbf{M}}) \circ \mathbf{X}) \approx \mathbb{E}\,\widehat{\mathbf{M}} + \mathbb{E}(\mathbf{Y} - \widehat{\mathbf{M}}) = \mathbf{M},$$

which is approximately unbiased. However, the estimator $\widehat{\mathbf{M}} + \widehat{p}^{-1}(\mathbf{Y} - \widehat{\mathbf{M}}) \circ \mathbf{X}$ is no longer of rank R, which increases the variances. This leads to using the projection as in Equation 30, which is asymptotically efficient in terms of both rate and preconstant.

Alternatively, the debiasing can be achieved through the two-step least squares (Chernozhukov et al. 2019). Suppose the true number of factors, r, is known.

Algorithm 5. Debias using two-step least squares.

Step 1. Obtain $\widehat{\mathbf{M}}$ as in Equation 29.

Step 2. Let the columns of $\frac{1}{\sqrt{N}}\widehat{\mathbf{B}}$ be the left singular vectors of $\widehat{\mathbf{M}}$, corresponding to the first r singular values.

Step 3. Estimate the latent factors at time t by $\widetilde{\mathbf{f}}_t := \left(\sum_{i=1}^N \widehat{\mathbf{b}}_i\widehat{\mathbf{b}}_i' x_{it}\right)^{-1} \sum_{i=1}^N \widehat{\mathbf{b}}_i y_{it} x_{it}$, and let $\widetilde{\mathbf{F}} = (\widetilde{\mathbf{f}}_1, \ldots, \widetilde{\mathbf{f}}_T)'$.

Step 4. Update loading estimates by $\widetilde{\mathbf{B}} = (\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_N)'$, where

$$\widetilde{\mathbf{b}}_i := \left(\sum_{t=1}^T \widetilde{\mathbf{f}}_t\widetilde{\mathbf{f}}_t' x_{it}\right)^{-1} \sum_{t=1}^T \widetilde{\mathbf{f}}_t y_{it} x_{it}.$$

Step 5. The asymptotically unbiased estimator for $\mathbf{M}$ is $\widetilde{\mathbf{M}} := \widetilde{\mathbf{B}}\widetilde{\mathbf{F}}'$.

A key technical argument is to ensure that the estimation error in $\widehat{\mathbf{B}}$ (step 2) has no impact on the factor estimator (step 3). Chen et al. (2019) achieved this using an auxiliary leave-one-out argument.

When the missing probability $P(x_{it} = 1)$ varies across i, there are two ways to revise the previous algorithm to achieve the asymptotic normality. One way is to replace Equation 29 with a weighted regularization:

$$\min_{\mathbf{M}} \|(\widehat{\mathbf{W}}^{-1/2}\mathbf{Y} - \widehat{\mathbf{W}}^{-1/2}\mathbf{M}) \circ \mathbf{X}\|_F^2 + \lambda\|\mathbf{M}\|_n, \tag{31.}$$

where $\widehat{\mathbf{W}}$ is a diagonal matrix, whose ith diagonal entry equals $\widehat{w}_i := \frac{1}{T}\sum_{t=1}^T x_{it}$. This debiases the least squares part of the loss function, adopting the same idea of IPW. The remaining steps of Algorithm 5 are the same. Then, the same auxiliary leave-one-out technical argument of Chen et al. (2019) still goes through. The other way is to apply sample splitting, which evenly splits the columns of $\mathbf{Y}$ into two parts: On one part, we run the penalized regression as in Equation 29 and obtain $\widehat{\mathbf{B}}$, and on the other part, we run iterative least squares. Then, we exchange the two parts and redo the estimations. The final estimator is taken as the average of the two. Supposing u_{it} is serially independent, the sample splitting then artificially creates independences among various statistics from the splitting sample. For detailed descriptions of this approach, the reader is referred to the work by Chernozhukov et al. (2019).

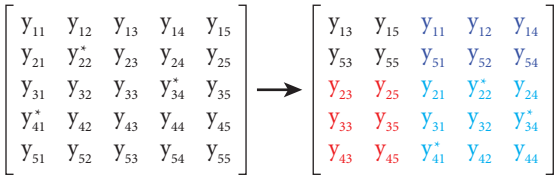

Figure 3

Missing data matrix rearrangement. This figure illustrates the block rearrangement from Bai & Ng (2019b) on the missing data matrix.

5.4. Block Rearrangements

In an attempt to handle endogenous missing, Bai & Ng (2019b) propose a block rearrangement method. At the cost of this generality, they require that the data matrix $\mathbf{Y}$ should have a sufficiently large balanced subblock after elementary rearrangements. (For related ideas, readers are referred to Cai, Cai & Zhang 2016; Fan & Kim 2019.)

Specifically, a preliminary step of their estimation is to rearrange the data in a shape in which all the factor loadings can be estimated in one subblock and all the factors can be estimated in another subblock. The example in **Figure 3** is adapted from Bai & Ng (2019b), which gives a good illustration on this manipulation, and shows the $N \times T$ matrix for y_{it}.

In **Figure 3**, the left matrix is the originally collected data, and the right is the rearranged one. The asterisks denote missing data. From the column perspective, the first, second, and fourth columns have missing values and therefore are rearranged as the last three columns in the right panel; from the row perspective, the second, third, and fourth rows have missing values and therefore are rearranged as the last three rows in the right panel. Bai & Ng (2019b) name the black blocks "bal," the black plus red blocks "tall," and the black plus blue blocks "wide."

Consider the missing value $\mathbf{y}^{*}_{22}$. We want to replace it with its expected value $\mathbb{E}(\mathbf{y}^{*}_{22}) = \mathbf{b}_2'\mathbf{f}_2$. Note that $\mathbf{y}^{*}_{22}$ shares the same factor loadings $\mathbf{b}_2$ with data points $\mathbf{y}_{23}$ and $\mathbf{y}_{25}$ in the wide block, and it shares the same factors $\mathbf{f}_2$ with data points $\mathbf{y}_{12}$ and $\mathbf{y}_{52}$ in the tall block. Meanwhile, $\mathbf{b}_2$ can be estimated using data in the tall block, and $\mathbf{f}_2$ can be estimated using data in the wide block. As a result, one might expect to recover $\mathbb{E}(\mathbf{y}^{*}_{22})$ with these two estimators. However, we must take into account the rotational indeterminacy inherent with the factor models. For a generic missing value $\mathbf{y}_{it}$,

$$\widehat{\mathbf{b}}_{\text{tall},i} = \mathbf{H}_{\text{tall}}'\mathbf{b}_i + o_P(1), \quad \widehat{\mathbf{f}}_{\text{wide},t} = \mathbf{H}_{\text{wide}}^{-1}\mathbf{f}_t + o_P(1).$$

Therefore,

$$\mathbf{b}_i'\mathbf{f}_t = \widehat{\mathbf{b}}_{\text{tall},i}'\mathbf{A}\widehat{\mathbf{f}}_{\text{wide},t} + o_P(1), \quad \mathbf{A} := \mathbf{H}_{\text{tall}}^{-1}\mathbf{H}_{\text{wide}}.$$

To estimate $\mathbf{A}$, by $\widehat{\mathbf{f}}_{\text{wide},t} = \mathbf{H}_{\text{wide}}^{-1}\mathbf{f}_t + o_P(1)$ and $\widehat{\mathbf{f}}_{\text{tall},t} = \mathbf{H}_{\text{tall}}^{-1}\mathbf{f}_t + o_P(1)$, we have

$$\widehat{\mathbf{f}}_{\text{tall},t} = \mathbf{A}\widehat{\mathbf{f}}_{\text{wide},t} + o_P(1).$$

So, one can run the regression of $\widehat{\mathbf{f}}_{\text{tall},t}$ on $\widehat{\mathbf{f}}_{\text{wide},t}$ to consistently estimate $\mathbf{A}$. This leads to the following estimation procedure.

Algorithm 6. Block rearrangement algorithm.

Step 1. Obtain estimators $(\widehat{\mathbf{b}}_{\text{wide}}, \widehat{\mathbf{F}}_{\text{wide}})$ using the tall block of $\mathbf{Y}$.
Step 2. Obtain estimators $(\widehat{\mathbf{b}}_{\text{tall}}, \widehat{\mathbf{F}}_{\text{tall}})$ using the wide block of $\mathbf{Y}$.
Step 3. Compute $\widehat{\mathbf{C}}_{\text{miss}} = \widehat{\mathbf{B}}_{\text{tall}}\mathbf{A}\widehat{\mathbf{F}}_{\text{wide}}'$, where $\mathbf{A}$ is obtained by regressing $\widehat{\mathbf{f}}_{\text{tall},t}$ on $\widehat{\mathbf{f}}_{\text{wide},t}$.
Step 4. Output $\widetilde{\mathbf{Y}}$, where $\widetilde{y}_{it} = y_{it}$ if y_{it} is observable and $\widetilde{y}_{it} = \widehat{c}_{\text{miss},it}$ if y_{it} is missing.

Once $\widetilde{\mathbf{Y}}$ is obtained, we apply the PCA again to the imputed data $\widetilde{\mathbf{Y}}$ to get more efficient estimates of $\mathbf{B}$ and $\mathbf{F}$. Suppose the size of the tall block is $N \times T_0$ and the size of the wide block is $N_0 \times T$. So, the size of the bal block is $N_0 \times T_0$. The whole sample size (including missing data points) is $N \times T$. Bai & Ng (2019b) require that

$$\max\{\sqrt{N}, \sqrt{T}\} = o(N_0) \quad \text{and} \quad \max\{\sqrt{N}, \sqrt{T}\} = o(T_0).$$

An implication of the above condition is that the missing data points should not be too frequent in the sense that the balanced subblock is large enough. Though this condition rules out the case of random missing (e.g., missing occurs as outcomes of Bernoulli trials), it is not stringent given the nature of endogenous missing.

5.5. A Simulation Study

We conduct a simulation study to compare six matrix completion approaches, namely:

1. **IPW.** The inverse probability weighting.
2. **ReUW.** Unweighted regularization. The eigenvectors of the estimator, as given in Equation 29.
3. **ReW.** Weighted regularization. The eigenvectors of the estimator, as given in Equation 31.
4. **ReDebias.** The debiased regularized estimator from Algorithm 5.
5. **EM.** The EM algorithm.

We generate a two-factor model where loadings, factors, and u_{it} are independent standard normal. Under homogeneous missing, we generate $x_{it} \sim$ Bernoulli(0.5); under heterogeneous missing, we generate $x_{it}|w_i \sim$ Bernoulli(w_i) and $w_i \sim$ Uniform[0.1,1]. The three regularized methods require choosing λ, the tuning parameter. Write the penalized loss function to be $\|(\mathbf{W}^{-1/2}\mathbf{Y} - \mathbf{W}^{-1/2}\mathbf{M}) \circ \mathbf{X}\|_F^2 + \lambda\|\mathbf{M}\|_n$, where $\mathbf{W}$ is a diagonal weighting matrix. The theory requires that with a high probability, there is $c > 0$,

$$(2 + c)\|\mathbf{U} \circ (\mathbf{W}^{-1}\mathbf{X})\| < \lambda.$$

So we set λ to be the 0.95 quantile of $2.2\|\mathbf{Z} \circ (\mathbf{W}^{-1}\mathbf{X})\|$, where $\mathbf{Z}$ is an $N \times T$ matrix of standard normal variables. In practice, one can also simulate $\mathbf{Z}$ using the estimated idiosyncratic covariance matrix.

We compare the performance of estimating the loading space, measured by $\mathbf{P}_\mathbf{B} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. **Table 1** reports $\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_\mathbf{B}\|$ averaged over 100 replications for each method. In all scenarios, the IPW performs the worst among all estimators. Under homogeneous missing, all the other four

Table 1 Comparison among five matrix completion methods

N	T	IPW	ReUW	ReW	ReDebias	EM
		Homogeneous missing				
100	200	0.176	0.116	0.114	0.109	0.109
200	100	0.252	0.171	0.169	0.161	0.161
		Heterogeneous missing				
100	200	0.263	0.211	0.131	0.119	0.119
200	100	0.369	0.304	0.222	0.204	0.203

This table reports $\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_\mathbf{B}\|$ averaged over 100 replications for each method.
Abbreviations: EM, expectation-maximation algorithm; IPW, inverse probability weighting; ReDebias, debiased regularized estimator; ReUW, unweighted regularization; ReW, weighted regularization.

methods perform similarly, but the difference is much more noticeable under heterogeneous missing. The general ranking is that

$$\text{IPW} \prec \text{ReUW} \prec \text{ReW} \prec \text{ReDebias} \approx \text{EM}.$$

This ranking is as expected: IPW is the least efficient method among the five; ReUW uses the nuclear-norm regularized estimation that does not take into account the heterogeneous missing or debias; ReW accounts for the heterogeneous missing probabilities; and ReDebias further removes the regularization bias.

Finally, it is not surprising to see that ReDebias and EM perform similarly, because both start with an initial low-rank estimator (ReDebias initializes from ReW, while EM initializes from IPW) and then proceed via iterative least squares. But we note that ReDebias operates much faster because it only iterates once, so it is more attractive than EM in handling large-scale problems. We also implemented the early stop EM (which iterates only twice); it performs only slightly better than IPW and is worse than all the other estimators. Therefore, we conclude that ReDebias is the recommended method for handling large-scale low-rank matrix completion problems.

6. CONCLUSION

In this review, we have conducted a selective overview of the recent developments in factor models and their application on statistical learning. We focused on the perspective of the low-rank structure of factor models and particularly drew attention to estimating the model from the low-rank recovery point of view. New estimation and inference methods and matrix completion problems were discussed.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (NNSFC) grants 71991470, 71991471, and 71722011. Kunpeng Li and Yuan Liao are co-first authors, and Jianqing Fan is the corresponding author.

LITERATURE CITED

Abbe E. 2017. Community detection and stochastic block models: recent developments. *J. Mach. Learn. Res.* 18(1):6446–531

Abbe E, Fan J, Wang K, Zhong Y. 2020. Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Stat.* 48:1452–74

Agarwal A, Negahban S, Wainwright MJ. 2012. Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Stat.* 40(2):1171–97

Ahn S, Horenstein A. 2013. Eigenvalue ratio test for the number of factors. *Econometrica* 81:1203–27

Aït-Sahalia Y, Xiu D. 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *J. Econom.* 201(2):384–99

Antoniadis A, Fan J. 2001. Regularized wavelet approximations. *J. Am. Stat. Assoc.* 96:939–67

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K. 2018. *Matrix completion methods for causal panel data models*. NBER Work. Pap. 25132

Bai J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71:135–71

Bai J, Li K. 2012. Statistical analysis of factor models of high dimension. *Ann. Stat.* 40(1):436–65
Bai J, Li K. 2016. Maximum likelihood estimation and inference for approximate factor models of high dimension. *Rev. Econ. Stat.* 98(2):298–309
Bai J, Liao Y. 2016. Efficient estimation of approximate factor models via penalized maximum likelihood. *J. Econom.* 191(1):1–18
Bai J, Ng S. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70(1):191–221
Bai J, Ng S. 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74(4):1133–50
Bai J, Ng S. 2019a. Rank regularized estimation of approximate factor models. *J. Econom.* 212(1):78–96
Bai J, Ng S. 2019b. Matrix completion, counterfactuals, and factor analysis of missing data. arXiv:1910.06677 [econ.EM]
Bai J, Wang P. 2016. Econometric analysis of large factor models. *Annu. Rev. Econ.* 8:53–80
Baltagi BH, Kao C, Wang F. 2017. Identification and estimation of a large factor model with structural instability. *J. Econom.* 197(1):87–100
Barigozzi M, Cho H. 2018. Consistent estimation of high-dimensional factor models when the factor number is over-estimated. arXiv:1811.00306 [stat.ME]
Barigozzi M, Cho H, Fryzlewicz P. 2018. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *J. Econom.* 206(1):187–225
Barigozzi M, Luciani M. 2019. Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the EM algorithm. arXiv:1910.03821 [math.ST]
Barras L, Scaillet O, Wermers R. 2010. False discoveries in mutual fund performance: measuring luck in estimated alphas. *J. Finance* 65(1):179–216
Belloni A, Chernozhukov V, Hansen C. 2014. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* 81(2):608–50
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57(1):289–300
Brillinger DR. 1964. *A frequency approach to the techniques of principal components, factor analysis and canonical variates in the case of stationary time series.* Invited paper, Royal Statistical Society Conference, Cardiff, Wales, UK, Sept. 29–Oct. 1. **https://www.stat.berkeley.edu/~brill/Papers/rss1964.pdf**
Cai T, Cai TT, Zhang A. 2016. Structured matrix completion with applications to genomic data integration. *J. Am. Stat. Assoc.* 111(514):621–33
Cai T, Liu W. 2011. Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.* 106(494):672–84
Candès EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? *J. Assoc. Comput. Mach.* 58(3):1–37
Catoni O. 2012. Challenging the empirical mean and empirical variance: a deviation study. *Ann. l'IHP Probab. Stat.* 48:1148–85
Chan KS. 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Stat.* 21(1):520–33
Chen D, Mykland PA, Zhang L. 2020. The five trolls under the bridge: principal component analysis with asynchronous and noisy high frequency data. *J. Am. Stat. Assoc.* 115(532):1960–77
Chen EY, Tsay RS, Chen R. 2020. Constrained factor models for high-dimensional matrix-variate time series. *J. Am. Stat. Assoc.* 115(530):775–93
Chen Y, Chi Y, Fan J, Ma C, Yan Y. 2020a. Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* 30(4):3098–121
Chen Y, Fan J, Ma C, Yan Y. 2019. Inference and uncertainty quantification for noisy matrix completion. *PNAS* 116(46):22931–37
Chen Y, Fan J, Ma C, Yan Y. 2020b. Bridging convex and nonconvex optimization in robust PCA: noise, outliers, and missing data. arXiv:2001.05484 [stat.ML]
Cheng X, Liao Z, Schorfheide F. 2016. Shrinkage estimation of high-dimensional factor models with structural instabilities. *Rev. Econ. Stud.* 83(4):1511–43

Chernozhukov V, Hansen CB, Liao Y, Zhu Y. 2019. *Inference for heterogeneous effects using low-rank estimations.* Work. Pap. CWP31/19, Cent. Microdata Methods Pract., London
Chudik A, Pesaran MH, Tosetti E. 2011. Weak and strong cross-section dependence and estimation of large panels. *Econom. J.* 14(1):C45–90
Connor G, Linton O. 2007. Semiparametric estimation of a characteristic-based factor model of stock returns. *J. Empir. Finance* 14:694–717
Connor G, Matthias H, Linton O. 2012. Efficient semiparametric estimation of the Fama-French model and extensions. *Econometrica* 80:713–54
Doz C, Giannone D, Reichlin L. 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *J. Econom.* 164(1):188–205
Doz C, Giannone D, Reichlin L. 2012. A quasi-maximum likelihood approach for large, approximate dynamic factor models. *Rev. Econ. Stat.* 94:1014–24
Fama EF, French KR. 2015. A five-factor asset pricing model. *J. Financ. Econ.* 116(1):1–22
Fan J, Han X, Gu W. 2012. Estimating false discovery proportion under arbitrary covariance dependence. *J. Am. Stat. Assoc.* 107(499):1019–35
Fan J, Ke Y, Liao Y. 2021. Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. *J. Econom.* 222:269–94
Fan J, Ke Y, Sun Q, Zhou WX. 2019a. FarmTest: factor-adjusted robust multiple testing with approximate false discovery control. *J. Am. Stat. Assoc.* 114:1880–93
Fan J, Ke Y, Wang K. 2020. Factor-adjusted regularized model selection. *J. Econom.* 216(471):71–85
Fan J, Kim D. 2019. Structured volatility matrix estimation for non-synchronized high-frequency financial data. *J. Econom.* 209(1):61–78
Fan J, Li Q, Wang Y. 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. B* 79(1):247–65
Fan J, Li R, Zhang CH, Zou H. 2020. *Statistical Foundations of Data Science.* Boca Raton, FL: CRC Press
Fan J, Liao Y. 2020. *Learning latent factors from diversified projections and its applications to over-estimated and weak factors.* SSRN Work. Pap. 3446097
Fan J, Liao Y, Mincheva M. 2013. Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Stat. Soc. B* 75:603–80
Fan J, Liao Y, Wang W. 2016. Projected principal component analysis in factor models. *Ann. Stat.* 44(1):219–54
Fan J, Liao Y, Yao J. 2015. Power enhancement in high-dimensional cross-sectional tests. *Econometrica* 83:1497–541
Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. B* 70(5):849–911
Fan J, Wang D, Wang K, Zhu Z. 2019b. Distributed estimation of principal eigenspaces. *Ann. Stat.* 47(6):3009–31
Fan J, Wang W, Zhong Y. 2018. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* 18(207):1–42
Fan J, Wang W, Zhong Y. 2019. Robust covariance estimation for approximate factor models. *J. Econom.* 208(1):5–22
Fan J, Wang W, Zhu Z. 2021. A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. *Ann. Stat.* 49(3):1239–66
Fan J, Xue L, Yao J. 2017. Sufficient forecasting using factor models. *J. Econom.* 201(2):292–306
Fan J, Zhong Y. 2018. Optimal subspace estimation using overidentifying vectors via generalized method of moments. arXiv:1805.02826 [stat.ME]
Forni M, Hallin M, Lippi M, Reichlin L. 2000. The generalized dynamic factor model: identification and estimation. *Rev. Econ. Stat.* 82:540–54
Forni M, Hallin M, Lippi M, Reichlin L. 2005. The generalized dynamic factor model: one-sided estimation and forecasting. *J. Am. Stat. Assoc.* 100(471):830–40
Gagliardini P, Ossola E, Scaillet O. 2016. Time-varying risk premium in large cross-sectional equity data sets. *Econometrica* 84(3):985–1046
Gagliardini P, Ossola E, Scaillet O. 2019. *Estimation of large dimensional conditional factor models in finance.* Res. Pap. 19–46, Swiss Finance Inst., Geneva

Giannone D, Reichlin L, Small D. 2008. Nowcasting: the real-time informational content of macroeconomic data. *J. Monet. Econ.* 55(4):665–76
Giglio S, Liao Y, Xiu D. 2021. Thousands of alpha tests. *Rev. Financ. Stud.* 34(7):3456–96
Goncalves S, Perron B. 2020. Bootstrapping factor models with cross sectional dependence. *J. Econom.* 218:476–95
Hansen BE. 2000. Sample splitting and threshold estimation. *Econometrica* 68(3):575–603
Hansen C, Liao Y. 2018. The factor-lasso and *k*-step bootstrap approach for inference in high-dimensional economic applications. *Econom. Theory* 35:465–509
Harvey CR, Liu Y. 2018. *False (and missed) discoveries in financial economics.* Tech. Rep., Duke Univ., Durham, NC
Harvey CR, Liu Y, Zhu H. 2015. . . . and the cross-section of expected returns. *Rev. Financ. Stud.* 29(1):5–68
Imbens GW, Rubin DB. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* New York: Cambridge Univ. Press
Juodis A, Sarafidis V. 2020. *A linear estimator for factor-augmented fixed-T panels with endogenous regressors.* Tech. Rep., Dep. Econom. Bus. Stat., Monash Univ., Melbourne, Aust.
Karabiyik H, Urbain JP, Westerlund J. 2019. CCE estimation of factor-augmented regression models with more factors than observables. *J. Appl. Econom.* 34(2):268–84
Ke ZT, Fan J, Wu Y. 2015. Homogeneity pursuit. *J. Am. Stat. Assoc.* 110(509):175–94
Klopp O, Lounici K, Tsybakov AB. 2017. Robust matrix completion. *Probab. Theory Relat. Fields* 169(1–2):523–64
Koltchinskii V, Lounici K, Tsybakov AB. 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.* 39(5):2302–29
Lam C, Yao Q. 2012. Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Stat.* 40(2):694–726
Lawley D, Maxwell A. 1971. *Factor Analysis as a Statistical Method*. London: Butterworths. 2nd ed.
Lee S, Liao Y, Seo MH, Shin Y. 2021. Factor-driven two-regime regression. *Ann. Stat.* 49(3):1656–78
Li H, Li Q, Shi Y. 2017. Determining the number of factors when the number of factors can increase with sample size. *J. Econom.* 197(1):76–86
Li J, Todorov V, Tauchen G. 2019. Jump factor models in large cross-sections. *Quant. Econ.* 10(2):419–56
Li KC. 1991. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* 86(414):316–27
Liao Y, Yang X. 2018. *Uniform inference for characteristic effects of large continuous-time linear models.* SSRN Work. Pap. 3069985
Ludvigson S, Ng S. 2010. A factor analysis of bond risk premia. In *Handbook of Empirical Economics and Finance*, ed. A Ulah, D Giles, pp. 313–72. Boca Raton, FL: CRC Press
Ma S, Goldfarb D, Chen L. 2011. Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Program.* 128(1–2):321–53
Massacci D. 2017. Least squares estimation of large dimensional threshold factor models. *J. Econom.* 197(1):101–29
McCracken MW, Ng S. 2016. FRED-MD: a monthly database for macroeconomic research. *J. Bus. Econ. Stat.* 34(4):574–89
Moon HR, Weidner M. 2018. Nuclear norm regularized estimation of panel regression models. arXiv:1810.10987 [econ.EM]
Negahban S, Wainwright MJ. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Stat.* 39(2):1069–97
Onatski A. 2010. Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* 92(4):1004–16
Onatski A. 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econom.* 168(2):244–58
Pelger M. 2019. Large-dimensional factor modeling based on high-frequency observations. *J. Econom.* 208(1):23–42
Romano JP, Shaikh AM, Wolf M. 2008. Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST* 17(3):417

Romano JP, Wolf M. 2007. Control of generalized error rates in multiple testing. *Ann. Stat.* 35(4):1378–408
Schott JR. 1994. Determining the dimensionality in sliced inverse regression. *J. Am. Stat. Assoc.* 89(425):141–48
Seo MH, Linton O. 2007. A smoothed least squares estimator for threshold regression models. *J. Econom.* 141(2):704–35
Stock JH, Watson MW. 2002a. Forecasting using principal components from a large number of predictors. *J. Am. Stat. Assoc.* 97:1167–79
Stock JH, Watson MW. 2002b. Macroeconomic forecasting using diffusion indexes. *J. Bus. Econ. Stat.* 20(2):147–62
Stock JH, Watson MW. 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, Vol. 2A, eds. J Taylor, H Uhlig, pp. 415–525. Amsterdam: Elsevier
Storey JD. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. B* 64(3):479–98
Su L, Miao K, Jin S. 2019. *On factor models with random missing: EM estimation, inference, and cross validation.* Work. Pap. 04-2019, Sch. Econ., Singapore Manag. Univ.
Su L, Wang X. 2017. On time-varying factor models: estimation and testing. *J. Econom.* 198(1):84–101
Wang D, Liu X, Chen R. 2019. Factor models for matrix-valued high-dimensional time series. *J. Econom.* 208(1):231–48
Wang S, Yang H, Yao C. 2019. On the penalized maximum likelihood estimation of high-dimensional approximate factor model. *Comput. Stat.* 34(2):819–46
Westerlund J, Urbain JP. 2013. On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Econ. Lett.* 119(3):247–50
Xia D, Yuan M. 2019. Statistical inferences of linear forms for noisy matrix completion. arXiv:1909.00116 [math.ST]
Xiong R, Pelger M. 2019. Large dimensional latent factor modeling with missing observations and applications to causal inference. arXiv:1910.08273 [econ.EM]
Zhu Z, Wang T, Samworth RJ. 2019. High-dimensional principal component analysis with heterogeneous missingness. arXiv:1906.12125 [stat.ME]