Journal of Econometrics 212 (2019) 78–96

Contents lists available at ScienceDirect

Journal of Econometrics

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

Rank regularized estimation of approximate factor models[☆]

Jushan Bai[a], Serena Ng[a,b,*]

[a] Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027, United States
[b] NBER

ARTICLE INFO

Article history:
Available online 12 April 2019

JEL classification:
C30
C31

Keywords:
Singular-value thresholding
Robust principal components
Minimum-rank
Low rank decomposition
Nuclear-norm minimization

ABSTRACT

It is known that the common factors in a large panel of data can be consistently estimated by the method of principal components, and principal components can be constructed by iterative least squares regressions. Replacing least squares with ridge regressions turns out to have the effect of removing the contribution of factors associated with small singular values from the common component. The method has been used in the machine learning literature to recover low-rank matrices. We study the procedure from the perspective of estimating an approximate factor model. Under the rank-constraint, the common component is estimated by the space spanned by factors whose singular values exceed a threshold. The desire for minimum rank and parsimony lead to a data-dependent penalty for selecting the number of factors. The new criterion is more conservative than the existing deterministic penalties and is appropriate when the nominal number of factors is inflated by the presence of weak factors or large measurement noise. We provide asymptotic results that can be used to test economic hypotheses.

1. Introduction

A low rank component is characteristic of many economic data. In analysis of international business cycles, the component arises because of global shocks. In portfolio analysis, the component arises because of non-diversifiable risk. One way of modeling this component when given a panel of data X collected for N cross-section units over a span of T periods is to impose a factor structure. If the data have r factors, the r largest population eigenvalues of XX' should increase with N. In a big data (large N large T) setting, it has been shown that the space spanned by the factors can be consistently estimated by the eigenvectors corresponding to the r largest eigenvalues of XX'. But it is not always easy to decisively separate the small from the large eigenvalues from the data. Furthermore, the eigen-space is known to be sensitive to outliers even if they occur infrequently. Sparse spikes, not uncommon in economic and financial data, may inflate the estimated number of factors. It would be useful to recognize such variations in factor estimation.

It is known that eigenvectors, and hence the factor estimates, can be obtained by iterative least squares regressions of X on guesses of the factor scores and of the loadings. This paper considers an estimator of the common component subject to a minimum rank constraint. It can be understood as using iterative ridge regressions to estimate the regularized low rank component. Ridge regressions are known to shrink the parameter estimates of a linear model towards zero. They are

[☆] This paper was presented at the Big Data in Predictive Dynamic Econometric Modeling at UPenn, Columbia University, and at Harvard/MIT. It was previously circulated as 'Principal Components and Regularized Estimation of Factor Models'. We thank the seminar participants and Zongming Ma for helpful comments. This work is supported by the National Science Foundation, United States SES-1658770 (Bai), SES-1558623 (Ng).

[*] Corresponding author at: Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027, United States.
E-mail addresses: jb3064@columbia.edu (J. Bai), serena.ng@columbia.edu (S. Ng).

https://doi.org/10.1016/j.jeconom.2019.04.021
0304-4076/© 2019 Elsevier B.V. All rights reserved.

biased but are less variable. In the present context, iterative ridge regressions shrink and truncate the singular values of the common component. Hastie et al. (2015) show that the algorithm implements *singular value thresholding* (SVT) and delivers robust principal components (RPC) as output.

Our interest in SVT stems from its ability to estimate approximate factor models of minimum rank. Researchers have long been interested in minimum-rank factor analysis, though the effort to find a solution has by and large stopped in the 1980s because of computationally challenges. SVT overcomes the challenge by solving a relaxed surrogate problem and delivers a robust estimate of the common component. But while worse case error bounds for RPC that are uniformly valid over models in that class are available, these algorithmic properties make no reference to the probabilistic structure of the data. Their use in classical statistical inference is limited. We approach the problem from the perspective of a parametric factor model. Since we make explicit assumptions about the data generating process, we can obtain parametric rates of convergence and make precise the effects of singular value thresholding on the factor estimates. Our results are asymptotic, and present an alternative perspective to the algorithmic ones obtained under the assumption of a fixed sample.

Our first contribution is to provide a statistical analysis of RPC that complements the algorithmic results developed from a machine learning perspective.[1] Constrained estimation generally leads to estimates that are less variable, but at the cost of bias. As we will see, rank constrained estimation is no exception. Our second contribution is to incorporate minimum rank consideration into the selection of the number of factors. We propose a new criterion that implicitly adds a data dependent penalty to the deterministic penalty introduced in Bai and Ng (2002). Simulations suggest that the resulting criterion gives a more conservative estimate of the number of factors when there are outliers in the data, and when the contributions of some factors to the common component are small. An appeal of the new procedure is that we do not need to know which assumptions in the factor model are violated.

We adapt results in the machine learning literature to an econometric setting. Connecting the two strands of work requires that we use a different normalization of the factors and the loadings. Thus, we start in Section 2 with new results for unregularized factors estimated under the new normalization. For readers not familiar with matrix completion and low rank decompositions, Section 3 presents a review of relevant work before presenting our main results in Section 4.

The following notation will be used. We use the (T, N) to denote number of rows and columns of the data matrix X used in statistical factor analysis, but (m, n) to denote dimension of a matrix Z when we are considering algorithms. For an arbitrary $m\times n$ matrix Z, the full singular value decomposition (SVD) of Z is $Z = UDV'$ where $U = [u_1, \dots, u_m]$ is $m\times m$ and $V = [v_1, \dots, v_n]$ is $n\times n$, $U'U = I_m$, $V'V = I_n$, and D is a $m\times n$ matrix of zeros except for its $\min(m, n)$ diagonal entries which are taken by the non-zero population singular values $d_1, d_2, \dots, d_{\min(m,n)}$ of Z. The left eigenvectors of ZZ' are the same as the left singular vectors of Z since $ZZ' = UD^2U'$. The nuclear norm $\|Z\|_* = \sum_{i=1}^n d_i(Z)$ is the sum of the singular values of Z. The singular values are ordered such that $d_1(Z)$ is the largest. Let $\|Z\|_1 = \sum_{i,j} |Z_{ij}|$ be the component-wise 1-norm, and let $\|Z\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n |Z_{ij}|^2$ denote the Frobenius (or component-wise-2) norm. Let $U = [U_r; U_{m-r}]$ and $V = [V_r, V_{n-r}]$ where U_r consists of the first r columns, while U_{n-r} consists of the last $(n-r)$ columns of U. A similar partition holds for V. Then $Z = U_rD_rV_r' + U_{m-r}D_{n-r}V_{n-r}'$.

2. Estimation of approximate factor models

We use $i = 1, \dots, N$ to index cross-section units and $t = 1, \dots, T$ to index time series observations. Let $X_i = (X_{i1}, \dots, X_{iT})'$ be a $T \times 1$ vector of random variables. The $T \times N$ data matrix is denoted $X = (X_1, X_2, \dots, X_N)$. The factor representation of the data is

$$X = F^0\Lambda^{0'} + e \tag{1}$$

where F is a $T \times r$ matrix of common factors, Λ is a $N \times r$ matrix of factor loadings and whose true values are F^0 and Λ^0. We observe X, but not F, Λ, or e. The variations in the common component $C = F\Lambda'$ are pervasive, while those in the idiosyncratic errors e are specific. The population covariance structure of $X_t = (X_{1t}, X_{2t}, \dots, X_{Nt})'$ is

$$\Sigma_X = \Sigma_C + \Sigma_e.$$

where $\Sigma_C = \Lambda\Sigma_F\Lambda'$. A strict factor model assumes that Σ_e is diagonal. Under the assumption that T tends to infinity with N fixed, Anderson and Rubin (1956), Joreskog (1967), and Lawley and Maxwell (1971) show that the factor loadings estimated by maximum likelihood or covariance structure methods are $\sqrt{T}$ consistent and asymptotically normal. The thrust of our analysis is to regularize the rank of the common component. As this task cannot be accomplished with the normalization used in the method of asymptotic principal components (APC), we need to first consider unregularized estimation under a new normalization. Section 2.2 shows that the asymptotic properties of the PC estimator based on the new normalization can be obtained from results for the APC presented in Section 2.1.

[1] The literature on PC is vast. See, for example, Jolliffe (2002). Some recent papers on SVT are Udell et al. (2016), Agarwal et al. (2012), Yang et al. (2014), Hastie et al. (2015).

2.1. Asymptotic Principal Components (APC): $(\widetilde{F}, \widetilde{\Lambda})$

The assumption that Σ_e is diagonal is restrictive for many economic applications. The approximate factor model of Chamberlain and Rothschild (1983) relaxes this assumption. A defining characteristic of an approximate factor model with r factors is that the r largest population eigenvalues of Σ_X diverge as N increases, while the $r+1$ and the smaller eigenvalues are bounded. We study estimation of an approximate factor model under the assumptions in Bai and Ng (2002) and Bai (2003).

Assumption A. There exists a constant $M < \infty$ not depending on N, T such that

a. (Factors and Loadings): $E\|F_t^0\|^4 \le M$, $\|\Lambda_i\| \le \overline{\Lambda}$, $\frac{F^{0\,\prime}F^0}{T} \xrightarrow{p} \Sigma_F > 0$, and $\frac{\Lambda^{0\,\prime}\Lambda^0}{N} \xrightarrow{p} \Sigma_\Lambda > 0$.

b. (Idiosyncratic Errors): Time and cross-section dependence

(i) $E(e_{it}) = 0, E|e_{it}|^8 \le M$;

(ii) $E(\frac{1}{N}\sum_{i=1}^N e_{it}e_{is}) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \le M$ for all s and $\frac{1}{T}\sum_{s=1}^T\sum_{t=1}^T |\gamma_N(s,t)| \le M$;

(iii) $E(e_{it}e_{jt}) = \tau_{ij,t}$, $|\tau_{ij,t}| \le |\tau_{ij,t}|$ for some $\tau_{ij,t}$ and for all t, and $\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N |\tau_{ij,t}| \le M$;

(iv) $E(e_{it}e_{js}) = \tau_{ij,st}$ and $\frac{1}{NT}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T |\tau_{ij,ts}| < M$;

(v) $E|N^{-1/2}\sum_{i=1}^N| \sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]^4 \le M$ for every (t,s).

c. (Central Limit Theorems): for each i and t, $\frac{1}{\sqrt{N}}\sum_{i=1}^N \Lambda_i^0 e_{it} \xrightarrow{d} N(0, \Gamma_t)$ as $N \to \infty$, and $\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t^0 e_{it} \xrightarrow{d} N(0, \Phi_i)$ as $T \to \infty$.

Assumption A allows the factors to be dynamic and the errors to be serially and cross-sectionally dependent as well as heteroskedastic. The loadings can be fixed or random. While Σ_e need not be a diagonal matrix, (b) also requires it to be sufficiently sparse (the correlations to be weak). Thus Assumption A imposes a strong factor structure via positive definiteness of Σ_F and Σ_Λ. Parts (a) and (b) imply weak dependence between the factors and the errors: $E(\frac{1}{N}\sum_{i=1}^N \|\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t^0 e_{it}\|^2) \le M$. Bai and Ng (2002) show that r can be consistently estimated. In estimation of F and Λ, the number of factors r is typically treated as known.

For given r, the method of APC solves the following problem:

$$\min_{F,\Lambda} \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (X_{it} - \Lambda_i{}'F_t)^2 = \min_{F,\Lambda}\frac{1}{NT}\left\|X - F\Lambda'\right\|_F^2. \tag{2}$$

If we concentrate out Λ and use the normalization $\frac{F'F}{T} = I_r$, the problem is the same as maximizing $\text{tr}(F(XX')F)$ subject to $F'F/T = I_r$. But the solution is not unique. If F is a solution, then FQ is also a solution for any orthogonal $r \times r$ matrix Q $(QQ' = I_r)$. However, if we put the additional restriction that $\Lambda'\Lambda$ is diagonal, then the solution becomes unique (still up to a column sign change).

The APC estimates, denoted $(\widetilde{F}, \widetilde{\Lambda})$, are defined as[2]

$$\widetilde{F} = \sqrt{T}U_r \tag{3a}$$

$$\widetilde{\Lambda} = X'\widetilde{F}/T \tag{3b}$$

That is, the matrix of factor estimates is $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of $\frac{XX'}{NT}$. It can be verified that $\widetilde{\Lambda}'\widetilde{\Lambda}/N = D_r^2$, a diagonal matrix of the r largest eigenvalues of $\frac{XX'}{NT}$. Bai and Ng (2002) show that as $N, T \to \infty$,

$$\min(N,T)\frac{1}{T}\sum_{t=1}^T \|\widetilde{F}_t - \widetilde{H}_{NT}F_t^0\| = O_p(1).$$

That is to say, $\widetilde{F}_t$ consistently estimates F_t^0 up to a rotation by the matrix $\widetilde{H}_{NT}$, defined as

$$\widetilde{H}_{NT} = \left(\frac{\Lambda^{0\,\prime}\Lambda^0}{N}\right)\left(\frac{F^{0\,\prime}\widetilde{F}}{T}\right)D_r^{-2}. \tag{4}$$

This ‘big data blessing’ has generated a good deal of econometric research in the last two decades.[3] As explained in Bai and Ng (2013), $\widetilde{H}_{NT}$ will not, in general, be an identity matrix implying that the jth factor estimate $\widetilde{F}_j$ will not, in

[2] The non-zero eigenvalues XX' and $X'X$ are the same. An alternative estimator is based on the eigen-decomposition of the $N \times N$ matrix $X'X$ with normalization $\frac{\Lambda'\Lambda}{N} = I_r$.

[3] The method of APC is due to Connor and Korajzcyk (1986). Forni et al. (2000) and Stock and Watson (2002a,b) initiated interests in large dimensional factor models. See Bai and Ng (2008) for a review of this work. Fan et al. (2013) show consistency of the factor estimates when the principal components are constructed from the population covariance matrix.

general, equal F_j^0 even asymptotically. The exception is when the true (F^0, Λ^0) are such that $\frac{F^{0\prime}F^0}{T} = I_r$ and $\Lambda^{0\prime}\Lambda^0$ is diagonal. Of course, it would be unusual for F^0 to have second moments that agree with the normalization used to obtain $\widetilde{F}$. Nonetheless, in applications when interpretation of $\widetilde{F}$ is not needed, as in the context of forecasting, the fact that $\widetilde{F}$ consistently estimates the space spanned by F^0 enables $\widetilde{F}$ to be used as though F^0 were observed.

Theorem 1 of Bai (2003) shows that if $\sqrt{N}/T \to 0$ as $N, T \to \infty$, then $\text{plim}_{N,T\to\infty} \frac{\widetilde{F}'F^0}{T} = \mathbb{Q}_r$, $\text{plim}_{N,T\to\infty} D_r^2 = \mathbb{D}_r^2$, and

$$\sqrt{N}(\widetilde{F}_t - \widetilde{H}_{NT}'F_t^0) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{D}_r^{-2}\mathbb{Q}_r\Gamma_t\mathbb{Q}_r{}'\mathbb{D}_r^{-2}\right) \equiv \mathcal{N}(0, \text{Avar}(\widetilde{F}_t))$$

$$\sqrt{T}(\widetilde{\Lambda}_i - \widetilde{H}_{NT}^{-1}\Lambda_i^0) = \xrightarrow{d} \mathcal{N}\left(0, (\mathbb{Q}_r{}')^{-1}\Phi_i\mathbb{Q}_r^{-1}\right) \equiv \mathcal{N}(0, \text{Avar}(\widetilde{\Lambda}_i))$$

where $\mathbb{Q}_r = \mathbb{D}_r\mathbb{V}_r\Sigma_\Lambda^{-1/2}$, and $\mathbb{D}_r^2$ and $\mathbb{V}_r$ are the eigenvalues and eigenvectors of the $r \times r$ matrix $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, respectively. The asymptotic inference of the factors ultimately depends on the eigenvalues and eigenvectors of the true common component. Nonetheless, as shown in Bai and Ng (2006), $\widetilde{F}$ can be used in subsequent regressions as though they were F^0 provided $\frac{\sqrt{T}}{N} \to 0$.

While $\widetilde{H}_{NT}$ is widely used in asymptotic analysis, it is difficult to interpret. A pair of asymptotically equivalent rotation matrices, not previously considered in the literature, is the following:

$$\widetilde{H}_{1,NT} = (\Lambda^{0\prime}\Lambda^0)(\widetilde{\Lambda}'\Lambda^0)^{-1}, \qquad \widetilde{H}_{2,NT} = (F^{0\prime}F^0)^{-1}(F^{0\prime}\widetilde{F}).$$

Lemma 1. *(i): $\widetilde{H}_{NT} = \widetilde{H}_{1,NT} + o_p(1)$ and (ii): $\widetilde{H}_{NT} = \widetilde{H}_{2,NT} + o_p(1)$.*

From Lemma 1, we see that the inverse of $\widetilde{H}_{1,NT}$ is the regression coefficient of $\widetilde{\Lambda}$ on Λ^0, while $\widetilde{H}_{2,NT}$ is the regression coefficient of $\widetilde{F}$ on F^0. The $o_p(1)$ term in the lemma can be shown to be $O_p(1/\min(N, T))$. Theorem 1 of Bai (2003) remains valid when $\widetilde{H}_{NT}$ is replaced by either $\widetilde{H}_{1,NT}$ or $\widetilde{H}_{2,NT}$. In addition to being interpretable, these simpler rotation matrices may simplify proofs in future work, hence of independent interest.

2.2. Principal Components (PC): $(\widehat{F}, \widehat{\Lambda})$

Whereas the eigenvectors of XX' are known in the economics literature as APC, principal components (PC) are sometimes associated with the singular vectors of X. In statistical modeling, APC tends to emerge from a spiked-covariance analysis, while PC tends to follow from a spiked-mean analysis. At a more mechanical level, $\widetilde{F}$ defined above depends only on the eigenvectors but does not depend on D_r. This is a somewhat unusual definition, as textbooks such as Hastie et al. (2001) define principal components as U_rD_r. Nonetheless, both definitions are valid and differ in the normalization used. We now consider a different definition of principal components that will be useful for the analysis in Section 3.

If we write the SVD of X as $X = U\breve{D}V'$, the singular values in $\breve{D}_r$ are of $O_p(\sqrt{NT})$ magnitude. We consider the scaled data

$$Z = \frac{X}{\sqrt{NT}}, \qquad \text{svd}(Z) = UDV', \quad D = \frac{\breve{D}}{\sqrt{NT}}.$$

Note that the left and right singular vectors of Z are the same as those for X. But while the first r singular values of X (i.e. $\breve{D}_r$) diverge and the remaining $N - r$ are bounded as $N, T \to \infty$, the r singular values of Z (i.e. D_r) are bounded and the remaining $N - r$ singular values tend to zero.

The model for the scaled data Z is

$$Z = F^*\Lambda^{*\prime} + e^* \tag{5}$$

where $F^* = \frac{F^0}{\sqrt{T}}$, $\Lambda^* = \frac{\Lambda^0}{\sqrt{N}}$, and $e^* = \frac{e}{\sqrt{NT}}$. Based on the svd of $Z = UDV{}'$, we now introduce the PC estimates, defined as:

$$\widehat{F}_z = U_rD_r^{1/2} \tag{6a}$$
$$\widehat{\Lambda}_z = V_rD_r^{1/2}. \tag{6b}$$

Notably, the PC and APC estimates are equivalent up to a scale transformation. In particular, $\widehat{F}_z = \widetilde{F}\frac{D_r^{1/2}}{\sqrt{T}}$ and $\widehat{\Lambda}_z = \widetilde{\Lambda}\frac{D_r^{-1/2}}{\sqrt{N}}$. One can construct the APC factor estimates directly from a svd of $XX{}'$ and rescale the eigenvalues, or one can construct the PC factor estimates from a svd of the rescaled data Z.[4] While $(\widehat{F}_z, \widehat{\Lambda}_z)$ emerges from the optimization problem

[4] There may be numerical advantages to using PC over APC. The documentation of prcomp in R notes the calculation is done by a singular value decomposition of the (centered and possibly scaled) data matrix, not by using eigenvalues on the covariance matrix. This is generally the preferred method for numerical accuracy.

of $\min_{F,\Lambda}\|Z-F\Lambda'\|_F^2$, $\widehat{F}_z$ is an estimate for $F^*=\frac{F^0}{\sqrt{T}}$, not for F^0. Similarly, $\widehat{\Lambda}_z$ is an estimate for $\Lambda^*=\frac{\Lambda^0}{\sqrt{N}}$. For estimation of F^0 and Λ^0, we define

$$\widehat{F} = \sqrt{T}\,U_rD_r^{1/2} \tag{7a}$$

$$\widehat{\Lambda} = \sqrt{N}\,V_rD_r^{1/2}. \tag{7b}$$

That is to say, $\widehat{F}=\sqrt{T}\widehat{F}_z$ and $\widehat{\Lambda}=\sqrt{N}\widehat{\Lambda}_z$. It follows that

$$\frac{\widehat{F}'\widehat{F}}{T}=\frac{\widehat{\Lambda}'\widehat{\Lambda}}{N}=D_r, \qquad \widehat{F}=\widetilde{F}D_r^{1/2}, \qquad \widehat{\Lambda}=\widetilde{\Lambda}D_r^{-1/2}.$$

In contrast, APC uses the normalization $\frac{F'F}{T}=I_r$, and $\Lambda'\Lambda$ is diagonal. The unit length normalization makes it inconvenient to impose restrictions on the APC estimates of F, a limitation that is important for the analysis to follow.

To establish the large sample properties of the PC estimates, we need to compare the estimates with a rotation of the true factors. Given the relation between $\widetilde{F}$ and $\widehat{F}$, it is natural to define the new rotation matrix as

$$\widehat{H}_{NT}=\widetilde{H}_{NT}D_r^{1/2}.$$

This leads to the identities

$$\sqrt{N}(\widehat{F}_t-\widehat{H}_{NT}'F_t^0) = \sqrt{N}D_r^{1/2}(\widetilde{F}_t-\widetilde{H}_{NT}\,'F_t^0),$$
$$\sqrt{T}(\widehat{\Lambda}_i-\widehat{H}_{NT}^{-1}\Lambda_i^0) = \sqrt{T}D_r^{-1/2}(\widetilde{\Lambda}_i-\widetilde{H}_{NT}^{-1}\Lambda_i^0).$$

From the limiting distributions of $\widetilde{F}_t$ and $\widetilde{\Lambda}_i$, we obtain:

Proposition 1. *Suppose that the data are generated by (1) and Assumption A holds. As N and T both go to infinity, the PC-estimates $\widehat{F}$ and $\widehat{\Lambda}$ have the following properties:*

(i) $\sqrt{N}(\widehat{F}_t-\widehat{H}_{NT}'F_t^0)\xrightarrow{d}N\left(0,\,\mathbb{D}_r^{1/2}\,\textsc{Avar}(\widetilde{F}_t)\,\mathbb{D}_r^{1/2}\right)\equiv N\left(0,\,\textsc{Avar}(\widehat{F}_t)\right)$*;*

(ii) $\sqrt{T}(\widehat{\Lambda}_i-\widehat{H}_{NT}^{-1}\Lambda_i^0)\xrightarrow{d}N\left(0,\,\mathbb{D}_r^{-1/2}\,\textsc{Avar}(\widetilde{\Lambda}_i)\,\mathbb{D}_r^{-1/2}\right)\equiv N\left(0,\,\textsc{Avar}(\widehat{\Lambda}_i)\right)$*;*

(iii) Let $W_{NT}(\widehat{C}_{it})=\frac{1}{T}\widehat{\Lambda}_i'\textsc{Avar}(\widehat{F}_t)\widehat{\Lambda}_i+\frac{1}{N}\widehat{F}_t'\textsc{Avar}(\widehat{\Lambda}_i)\widehat{F}_t$*. Then* $(W_{NT}(\widehat{C}_{it}))^{-1/2}(\widehat{C}_{it}-C_{it}^0)\xrightarrow{d}N(0,1)$.

The PC estimates of F_t and Λ_i are $\sqrt{N}$ and $\sqrt{T}$ consistent respectively, just like the APC estimates. It follows that the estimated common component $\widehat{C}_{it}$ is $\min(\sqrt{N},\sqrt{T})$ consistent. Analogous to Lemma 1, we can define asymptotically equivalent rotation matrices. Let $\widehat{H}_{1,NT}=(\Lambda^{0\prime}\Lambda^0)(\widehat{\Lambda}'\Lambda^0)^{-1}$ and $\widehat{H}_{2,NT}=(F^{0\prime}F^0)^{-1}(F^{0\prime}\widehat{F})$. Proposition 1 holds with $\widehat{H}_{NT}$ replaced by either $\widehat{H}_{1,NT}$ or $\widehat{H}_{2,NT}$.

3. Rank and nuclear-norm minimization

In economic applications of factor models, it is customary to choose the number of factors $r\in[0,\text{rmax}]$ to provide a good fit of the data while taking model complexity into account. Bai and Ng (2002) suggest a class of consistent factor selection criteria, that is, rules that yield $\widehat{r}$ such that $\text{prob}(\widehat{r}=r)\to 1$ as $N,T\to\infty$. While criteria in this class have little difficulty singling out the dominant factors, they tend to be too liberal. Over-estimating the number of factors is possible when validity of Assumption A is questionable. The two leading causes are weak loadings, and idiosyncratic errors with large variances. The first has the implication that the smaller eigenvalues do not increase sufficiently fast with N, an issue emphasized in Onatski (2011) and others. The second has the implication that $r+1$-th eigenvalue (which should not increase with N) is not well separated from the rth eigenvalue (which should increase with N). Such a scenario can be due to outliers, which can increase the variance in an otherwise uninformative direction, and PCA is blind to the source of the variance. The problem is well documented in, for example, Hubert and Rousseeuw (2005), but is not given attention in the econometrics literature. Both weak loadings and outliers can distort the number of factors being estimated. Having criteria that guard against these distortions without pre-specifying the source of over-estimation is desirable.

Our approach is based on the notion of minimum rank. A variety of problems have minimum rank as motivation. Recall that the rank of an arbitrary $m\times n$ matrix Z is the largest number of columns of Z that are linearly independent. A related concept is the *spark*, which is the smallest k such that k columns of Z are linearly dependent. If $n\leq m$, the spark equals $n+1$ if and only if Z has full rank, i.e. $\textsc{spark}(Z)=n+1\Leftrightarrow\textsc{rank}(Z)=n$. The spark of a matrix is a lesser known concept because its evaluation is combinatorially hard, see Tillman and Pfetsch (2014). In contrast, the rank of a matrix can be computed as the number of non-zero singular values. Nonetheless, if $\textsc{spark}(Z)\neq n+1$, it holds that $\textsc{spark}(Z)\leq\textsc{rank}(Z)+1$. This implies that the rank of Z, in our case the number of factors, may not be the size of the smallest set of factors that span Z. This smaller set is what we now seek to recover. We motivate with a discussion of three related problems of interest: minimum rank factor analysis, matrix completion, and low rank matrix decompositions. Their resolutions will guide our robust factor analysis in Section 4.

3.1. Minimum rank factor analysis

Factor analysis has its roots in the study of personality traits. The original goal of factor analysis was to find a decomposition for the $N \times N$ matrix Σ_X into $\Sigma_X = \Sigma_C + \Sigma_e$, where Σ_C can be factorized into $\Lambda\Lambda'$, Λ is $N \times r$, and such that

$$\text{(i).}\quad \Sigma_X - \Sigma_e \succeq 0, \qquad \text{(ii)}\quad \Sigma_e \succeq 0, \qquad \text{(iii)}\quad \Sigma_e \quad \text{diagonal.} \tag{8}$$

Constraint (i) requires that the low rank communality matrix $\Sigma_C = \Sigma_X - \Sigma_e$ is positive semi-definite, and constraint (ii) requires that negative error variances do not arise (known in the literature as the Heywood case). The smallest rank that solves this problem has come to be known as the *minimum rank*. But while it is not hard to compute rank(X) from counting the non-zero eigenvalues of X, rank minimization is a NP hard problem because the cardinality function that defines rank is non-convex and non-differentiable. Furthermore, research in the 1950s such as by Guttman (1958) found a large number of factors in the data, casting doubt on the usefulness of the notion of a minimum rank. Interests in the problem subsided.

New attempts were made to tackle the problem in the 1980s. Of interest to us are the two that minimize D^C_{ii}, the eigenvalues of the communality matrix $\Sigma_C = \Sigma_X - \Sigma_e$. Define

$$f(\Sigma_e, ; r_0) = \sum_{i=r_0}^{n} D^C_{ii}.$$

The first attempt is *constrained minimum trace factor analysis* (CMFTA) which finds Σ_e to minimize $f(\Sigma_e; 1)$ subject to (8). Its name arises from the fact that $f(\Sigma_e, 1)$ is the sum of the eigenvalues of $\Sigma_X - \Sigma_e$, which is the trace of $(\Sigma_X - \Sigma_e)$. The second attempt is *approximate minimum rank factor analysis* (MFRA) which distinguishes between explained and unexplained common variance. In the MFRA setup, C is decomposed into $C = C^+ + C^-$, where C^+ is that part of the common component that will be explained, and is spanned by the most significant factors in the data. The communality matrix has analogous decomposition $\Sigma_C = \Sigma_C^+ + \Sigma_C^-$ where Σ_C^+ is the common variance that will be explained, and is associated with the largest eigenvalues of Σ_C. Hence $\Sigma_X = \Sigma_C^+ + (\Sigma_C^- + \Sigma_e)$. For fix r, MFRA then minimizes the unexplained common variance $\sum_{i=r+1}^{n} D^C_{ii} = f(\Sigma_e; r)$ subject to (8). In this context, ten Berge and Kiers (1991) define the *approximate minimum rank* to be the smallest r such that $\min_{\Sigma_e} f(\Sigma_e, r) \le \delta$ with $\delta > 0$. The minimum rank problem that the earlier literature has sought to solve is a special case that sets $\delta = 0$.[5] We will refer to the approximate minimum rank by r^*.

What makes MFTA and MFRA interesting is that the sum of eigenvalues is a convex function. Instead of tackling the original problem that is NP hard, they solve surrogate problems that can take advantage of interior point and semi-definite algorithms developed in the convex optimization literature.[6] Even though the proposed algorithms for solving MTFA and MRFA are no longer efficient given today's know how, convex relaxation of the rank function is an active area of research in recent years. We now turn to this work.[7]

3.2. Singular-Value Thresholding (SVT)

Many problems of interest in the big data era are concerned with recovery of a low rank matrix from limited or noisy information. For example, compressive sensing algorithms seek to reconstruct a signal from a system of underdetermined linear equations. In face recognition analysis, the goal is to recover a background image from frames taken under different conditions. Perhaps the best known example of matrix recovery is the Netflix challenge. Contestants were given a small training sample of data on movie ratings by a subset of Netflix users and were asked to accurately predict ratings of all movies in the evaluation sample. Without any restrictions on the completed matrix, the problem is under-determined since the missing entries can take on any values. But progress can be made if the matrix to be recovered has low rank. In the Netflix context, the low rank assumption amounts to requiring that preferences are defined over a small number of features (such as genres, leading actor/actress). The problem is then to complete Z by finding a matrix L that is factorizable into two matrices (for preference and features) with smallest rank possible, and such that $Z_{ij} = L_{ij}$ for each (i,j) that is observed. But, as noted earlier, rank minimization is NP-hard. The breakthrough is to replace rank minimization by nuclear norm minimization. This is important because the nuclear norm, which is also the sum of the singular values, is convex. Candes and Recht (2011) shows that Z can be recovered with high probability if the number of observed entries satisfies a lower bound, and that the unobserved entries are missing uniformly at random.

[5] CMTFA is due to Bentler (1972). See also Bentler and Woodward (1980), Shapiro (1982), Shapiro and ten Berge (2000). For MFRA, see ten Berge and Kiers (1991), and Shapiro and ten Berge (2000).

[6] A computational decision problem is NP hard if it is at least as hard as the hardest problem in class NP whose solutions can be verified in polynomial time. Intuitively, an NP hard problem is one that admits no general computational solution that is significantly faster than a brute force search.

[7] The connection between factor analysis and low rank matrix decompositions was a focus of Saunderson et al. (2012) and a recent paper by Bertsimas et al. (2016).

In exact matrix completion, the problem arises because of incomplete observations, but the data are perfectly measured whenever they are observed. A different matrix recovery problem arises not because of missing values, but because the data are observed with errors, also referred to as noise corruption. Generically, an $m \times n$ matrix Z can be decomposed as

$$Z = L + S$$

where L is a matrix of reduced rank r, and S is a sparse noise matrix. A measure of sparsity is cardinality, which is the number of non-zero entries, written as $\|S\|_0$. The goal is to recover L from data that are sparsely corrupted by noise. For minimizing $\|Z - L\|_F$ subject to $\text{rank}(L) \le r$, Eckart and Young (1936) obtain a truncated SVD as solution for L.[8] While the SVD solution given by $U_r D_r V_r{}'$ provides the best low rank approximation of Z, singular value decomposition is also known to be sensitive to large errors in practice, even if there are few of them. For example, if the data have fat tails, a small number of extreme values can account for a significant amount of variation in the data.[9] Since PCA is blind to the source of large variations, large noise contamination can corrupt the low rank component being identified, as will be seen in the simulations to follow.

To reduce noise corruption, one may want to penalize S and solve the regulated problem

$$\text{rank}(L) + \overline{\gamma}\|S\|_0, \qquad s.t. \quad Z = L + S.$$

This is challenging because rank and cardinality are both non-convex functions. Wright et al. (2009), Candes et al. (2011) show that under an incoherence on L and a cardinality condition on S, both L and S can be recovered exactly with high probability by solving what is known as the program of principal components pursuit (PCP):

$$\min_{L,S} \|L\|_* + \overline{\gamma}\|S\|_1 \quad \text{subject to} \quad Z = L + S$$

where $\overline{\gamma} = (\max(m, n))^{-1/2}$ is a regularization parameter. The output of the low rank component is referred to as robust principal components (RPC). Compared to the original problem, the non-convex constraints $\text{rank}(L)$ and $\|S\|_0$ have been replaced by convex functions $\|L\|_*$ and $\|S\|_1$, respectively. The incoherence condition on the singular vectors of L prevents the low rank component from being too sparse. The cardinality condition requires the support of S to be selected uniformly at random so that S is not low rank.[10] Zhou et al. (2010) further allow for the presence of small noise W so that $Z = L + S + W$. It is shown that L and S can still be recovered with high probability upon solving the convex program

$$\min_{L,S} \|L\|_* + \overline{\gamma}\,\|S\|_1 \quad \text{s.t.} \quad \|Z - L - S\|_F \le \delta$$

if, in addition to the incoherence and cardinality conditions, $\|W\| \le \delta$ holds for some known δ. This result establishes that RPC is stable in the presence of small entry-wise noise, a setup that seems appropriate for factor analysis. In summary, the main insight from matrix recovery problems is that while rank minimization is NP hard, the surrogate problem of nuclear-norm minimization still permits recovery of the desired low rank matrix with high probability.

In terms of implementation, singular value thresholding algorithm (SVT) plays an important role in the reparameterized problem. For a rank r matrix $Z = U_r D_r V_r{}' + U_{n-r} D_{n-r} V_{n-r}'$, define

$$D_r^\gamma = \Big[D_r - \gamma I_r\Big]_+ \equiv \max(D_r - \gamma I_r, 0). \tag{9}$$

Importantly, the SVT is the proximal operator associated with the nuclear norm.[11] Theorem 2.1 of Cai et al. (2008) shows that

$$U_r D_r^\gamma V_r' = \underset{L}{\text{argmin}}\, \gamma\, \|L\|_* + \frac{1}{2}\,\|Z - L\|_F^2\,. \tag{10}$$

In other words, the optimal approximation of the low rank component of Z under rank constraint is $U_r D_r^\gamma V_r'$ where D_r^γ is a matrix of thresholded singular values. Compared with the unregulated estimate of $U_r D_r V_r'$, the only difference is that the singular values are thresholded. It is possible for D_r^γ to have rank $r^* < r$ because of thresholding. As a consequence, the rank of the regulated estimate of the low rank component can be smaller than the unregulated estimate.[12]

[8] The Eckart–Young result is also known as the matrix approximation lemma, or the Eckart–Young–Minsky Theorem. Proofs are also available when the Frobenius norm is replaced by the spectral norm.

[9] See Delvin et al. (1981), Li and Chen (1985), Ma and Genton (2001), Hubert et al. (2005), among others.

[10] More precisely, Theorem 1.1 of Candes et al. (2011) shows that incoherent low-rank matrix can be recovered from non-vanishing fractions of gross errors in polynomial time. The proof assumes that S is uniformly distributed on the set of matrices with support of a fixed cardinality.

[11] A proximal operator specifies the value that minimizes an approximation to a regularized version of a function. Proximal methods are popular in convex optimizations. See, for example, Boyd et al. (2010).

[12] Algorithms that solve (10) include Augmented Lagrange Multiplier and Accelerated Proximal Gradient (Lin et al. (2013)), ADMM (Boyd et al. (2010)), and CVX (Grant and Boyd (2015)).

3.3. Relation to factor models

The previous subsection provides results for recovery of the low rank and sparse components, L and S, respectively. Since L has rank r, it can be factorized as a product of two rank r matrices, A and B, that is, $L = AB'$. This subsection discusses the recovery of A and B. This is useful since we are eventually interested in the factors and the loadings of a minimum-rank factor model, not just the common component.

The key step that ties a low rank decomposition to factor analysis is to establish that the regularized problem with γ as threshold parameter, i.e.

$$\min_{A,B} \frac{1}{2} \left\| Z - AB\,' \right\|_F^2 + \gamma \left\| AB\,' \right\|_* \tag{11}$$

has solution

$$\overline{A} = U_r (D_r^\gamma)^{1/2}, \quad \overline{B} = V_r (D_r^\gamma)^{1/2}, \tag{12}$$

and that $\overline{L} = \overline{A}\,\overline{B}'$ also solves (10). The result that $(\overline{A}, \overline{B})$ solves the problem (11) if and only if $\overline{L} = \overline{A}\,\overline{B}\,'$ solves (10) was first noted in Rennie and Srebro (2005). Detailed proofs are given in Hastie et al. (2015) and Boyd et al. (2010). A sketch of the idea is as follows.

Since $AB' = U_r D_r V_r\,'$, and U_r and V_r are orthonormal, then by Cauchy–Schwarz inequality,

$$\begin{aligned} \text{trace}(D_r) &= \text{trace}(U_r\,' AB\,' V_r) \le \|A\|_F \, \|B\|_F \\ &\le \frac{1}{2}\big(\|A\|_F^2 + \|B\|_F^2\big). \end{aligned}$$

But since $L = AB'$ by definition, it follows that $\|L\|_* = \text{trace}(D_r)$ and the above implies

$$\|L\|_* \le \frac{1}{2}\big(\|A\|_F^2 + \|B\|_F^2\big)$$

with equality when $A = U_r D_r^{1/2}$ and $B = V_r D_r^{1/2}$. Hence (11) is just a reformulation of (10) in terms of A and B. Consider now the first order conditions associated with (11). If $\overline{A}$ and $\overline{B}$ are solutions, it must be that $-(Z - \overline{AB}')\overline{B} + \gamma\overline{A} = 0$ and $-(Z - \overline{AB}')'\overline{A} + \gamma\overline{B} = 0$. Left multiplying the first condition by $\overline{A}\,'$ and the second by $\overline{B}\,'$, we see that $\overline{A}\,'\overline{A} = \overline{B}\,'\overline{B}$ when the first order conditions hold. Rearranging terms, we obtain

$$\begin{pmatrix} -\gamma I & Z \\ Z' & -\gamma I \end{pmatrix} \begin{pmatrix} \overline{A} \\ \overline{B} \end{pmatrix} = \begin{pmatrix} \overline{A} \\ \overline{B} \end{pmatrix} \overline{A}\,'\overline{A}. \tag{13}$$

This has the generic structure $\mathbb{Z}\mathbb{V} = \mathbb{V}\mathbb{X}$, which is an eigenvalue problem. In particular, the eigenvalues of $\mathbb{X}$ are those of $\mathbb{Z}$, and $\mathbb{V}$ are the corresponding eigenvectors. In the present context, the eigenvalues of $\overline{A}\,'\overline{A}$ are those of the first matrix on the left, and $(\overline{A}, \overline{B})$ are the corresponding left and right eigenvectors. Though $-\gamma \pm d_i$ are two possible solutions for every i, we only accept those satisfying $\sqrt{d_i - \gamma} > 0$. Collecting the thresholded singular-values into D_r^γ, the solution defined in (12) obtains. The $\overline{A}$ and $\overline{B}$ defined in (12) are the robust principal components of Z under the assumed normalization that $\overline{A}'\overline{A} = \overline{B}'\overline{B} = \overline{D}_r^\gamma$.[13]

It is also of interest to note $\overline{L} = \overline{AB}'$ is the posterior estimate of L under the likelihood $Z_{it} = L_{it} + S_{it}$, $S_{it} \sim N(0, \sigma^2)$ with prior for $L = U_r D_r V_r'$ given by

$$p(L) = p(U_r)p(D_r)p(V_r)$$

where $p(U_r)$ and $p(V_r)$ are (uniform) Haar distributions, and $p(D_r) = p(d_1, \ldots, d_r) = \prod_{i=1}^r \gamma \exp(-\gamma d_k)$. The exponential distribution has a mode of zero and favors sparse solutions. The result, given in Todeschini et al. (2013), implies that the frequentist problem of minimizing the sum of squared residuals subject to a constraint on the nuclear norm of the low rank component is the same as a Bayesian problem with a flat prior on the eigenvectors and an exponential prior on the eigenvalues.

4. Rank constrained approximate factor models

An important result in econometric work on large dimension factor modeling is that the factor space can be consistently estimated by principal components when N and T are large, as reviewed in Section 2. A key finding in the machine learning literature is that the minimum rank component of the data can be obtained by robust principal components, as reviewed in Section 3. This section connects RPC to rank constrained factor analysis. Since we need to adapt results in the machine learning literature to an econometric setup, we start by clarifying some differences between the statistical and the algorithmic approach to low rank modeling.

Notably, the decomposition $Z = L + S$ is consistent with many probabilistic structures. Statistical factor analysis specifies one particular representation: $X = F\Lambda\,' + e$ with $Z = \frac{X}{\sqrt{NT}}$. We use Assumption A to restrict the factors, loadings,

[13] Udell et al. (2016) referred to (11) as *quadratically regularized PC*. See their Appendix for a complete proof.

and the idiosyncratic noise so that the singular values of the common component $C = F\Lambda'$ diverge with N. We find F and Λ to minimize the sum of squared residuals and let e be residually determined. We establish that $\widehat{F}$ is consistent at rate $\sqrt{N}$, $\widehat{\Lambda}$ at rate $\sqrt{T}$, and $\widehat{C}$ at rate $\min(\sqrt{N}, \sqrt{T})$ assuming that the factor model is correctly specified.

In contrast, machine learning analysis is 'distribution free' and the data generating process is left unspecified. Often, S is explicitly chosen together with L, not residually determined. For Netflix type problems when a low rank matrix is to be recovered from data with missing values, the probability of exact recovery is typically obtained under an incoherence condition that makes no reference to singular values. But when the data are noisy rather than missing, one can only hope to recover an approximate (rather than the exact) low rank component.[14] In such cases, more important are the singular vectors associated with the large singular values. For such problems, which seem more comparable to our setup, Agarwal et al. (2012) obtain an error bound in the Frobenius norm for L and for S under the assumption that $\|L\|_\infty \le \frac{\alpha}{\sqrt{mn}}$. But recall that $\|L\|_\infty = \max_{i,t} |L_{it}|$ and $\|L\|_F^2 = \sum_{i=1}^m \sum_{t=1}^n |L_{it}|^2 = \sum_{i=1}^r d_{L,i}^2$. The condition $\|L\|_\infty \le \frac{\alpha}{\sqrt{mn}}$ is effectively a restriction on the sum of the singular values of the low rank component L.

What transpires from the above discussion is that machine learning methods restrict the sample singular values of L and obtain finite sample error bounds. In contrast, approximate factor analysis puts restrictions on the population singular values of the common component C through moment conditions collected into Assumption A, which also enable precise parametric convergence rate to be obtained. Interestingly, Corollary 2 of Agarwal et al. (2012) suggests that for the spike mean model, the error of the low rank component is of order $\frac{N+T}{NT} \approx \min(N, T)^{-1}$ with high probability. This agrees with the asymptotic convergence rate obtained in our previous work on the unconstrained case.

A machine learning analysis closest in spirit to ours is Bertsimas et al. (2016). This paper reformulates estimation of minimum rank factor model for iid data from the perspective of smooth optimization under convex compact constraints. Via lower bounds, the solutions are shown to be certifiably optimal in many cases, requiring only that $\Sigma_X = \Sigma_C + \Sigma_e$, where Σ_e is diagonal. Our data need not be iid, and Σ_e need not be diagonal, but we invoke more assumptions to provide parametric convergence rates. The results provide different perspectives to a related problem.

4.1. The Rank Constrained Factor Estimates (RPCA): $(\overline{F}, \overline{\Lambda})$

Data driven by r common factors should have r singular values much larger than the remaining ones. But some factors may contribute little to the common component and their associated singular values may not be noticeably large. Furthermore, some singular values could be large because of extreme outliers in the idiosyncratic errors, and nothing to do with pervasive variations. Boivin and Ng (2006) suggest to account for noise in the data by re-weighing each series with the standard-deviation of the idiosyncratic error. A drawback is that these weights may themselves be sensitive to outliers. The POET estimator of Fan et al. (2013) suggests to use thresholding to enforce a sparse matrix of idiosyncratic errors.

We propose to apply thresholding to Σ_C rather than Σ_e. This is appealing because the common component is typically the object of interest, not the idiosyncratic errors. To estimate a factor model that explicitly recognizes minimum rank as one of objectives requires that we map the scaled factor model given in (5) into the problem defined in (11). Consider the goal of recovering C^* in the decomposition

$$Z = C + e = C^* + C^- + e \qquad C^* = F^* \Lambda^{*'}.$$

As in ten Berge and Kiers (1991), the common-component is decomposed into $C = C^* + C^-$. While C has rank r, it is well approximated by C^* whose rank is r^*, and we want to estimate C^*.[15] From the previous subsection, the rank regularized problem is

$$(\overline{F}_Z, \overline{\Lambda}_Z) = \underset{F,\Lambda}{\operatorname{argmin}} \frac{1}{2}\Big(\big\|Z - F\Lambda'\big\|_F^2 + \gamma\, \|F\|_F^2 + \gamma\, \|\Lambda\|_F^2\Big). \tag{14}$$

The optimal solution defines the robust principal components:

$$\overline{F}_Z = U_r(D_r^\gamma)^{1/2}, \quad \overline{\Lambda}_Z = V_r(D_r^\gamma)^{1/2} \tag{15}$$

where $D_r^\gamma = [D_r - \gamma I_r]_+$. As in the unconstrained case, we work with the normalized estimates[16]:

$$\overline{F} = \sqrt{T} U_r (D_r^\gamma)^{1/2} = \sqrt{T}\overline{F}_Z \tag{16a}$$

$$\overline{\Lambda} = \sqrt{N} V_r (D_r^\gamma)^{1/2} = \sqrt{N}\overline{\Lambda}_Z. \tag{16b}$$

[14] For application of the incoherence condition in matrix completion, see Candes et al. (2011). For general matrix recovery problems, see Agarwal et al. (2012) and Negahban and Wainwright (2012).

[15] Motivated from an asset pricing perspective, Lettau and Pelger (2017) suggest to apply PCA to a covariance matrix of overweighted mean, or in our notation, $Z'Z + \gamma\overline{Z}\overline{Z}'$, where $\overline{Z}$ is the mean of Z. Their risk-premium PCA also uses regularization to control for weak factors.

[16] These are also the optimal solutions from the following optimization problem using X instead of Z

$$(\overline{F}, \overline{\Lambda}) = \underset{F,\Lambda}{\operatorname{argmin}}\Big(\frac{1}{NT}\, \big\|X - F\Lambda'\big\|_F^2 + \frac{\gamma}{T}\, \|F\|_F^2 + \frac{\gamma}{N}\, \|\Lambda\|_F^2\Big).$$

Though $(\overline{F}Q, \overline{\Lambda}Q)$ is also a solution for any orthonormal matrix Q, $(\overline{F}, \overline{\Lambda})$ defined in (16a) and (16b) is the only solution (up to a column sign change) that satisfies $\frac{\overline{F}'\overline{F}}{T} = \frac{\overline{\Lambda}'\overline{\Lambda}}{N} = D_r^\gamma$ when the diagonal elements of D_r^γ are distinct.

The rank restricted and the unrestricted estimates are related by

$$\begin{aligned}\overline{F} &= \widehat{F}\,\Delta_{NT}\\ \overline{\Lambda} &= \widehat{\Lambda}\,\Delta_{NT}\end{aligned}$$

where

$$\Delta_{NT}^2 = D_r^{\gamma} D_r^{-1} = \mathrm{diag}\left(\frac{(d_1-\gamma)_+}{d_1}, \ldots, \frac{(d_r-\gamma)_+}{d_r}\right). \tag{17}$$

Hence while $\frac{\widehat{F}'\widehat{F}}{T} = \frac{\widehat{\Lambda}'\widehat{\Lambda}}{N} = D_r$, now $\frac{\overline{F}'\overline{F}}{T} = \frac{\overline{\Lambda}'\overline{\Lambda}}{N} = D_r^{\gamma}$. Define the rotation matrix for $\overline{F}$ by

$$\overline{H}_{NT} = \widehat{H}_{NT}\Delta_{NT}$$

From the relationship between $\overline{F}$ and $\widehat{F}$,

$$\overline{F}_t - \overline{H}_{NT}' F_t^0 = \Delta_{NT}(\widehat{F}_t - \widehat{H}_{NT}' F_t^0). \tag{18}$$

Hence $\overline{F}_t$ estimates a rotation of F_t^0. But the inverse of $\overline{H}_{NT}$ is not the rotation matrix for $\overline{\Lambda}$. As shown in Appendix,

$$\overline{\Lambda}_i - \overline{G}_{NT}\Lambda_i^0 = \Delta_{NT}[\widehat{\Lambda}_i - \widehat{H}_{NT}^{-1}\Lambda_i^0] \tag{19}$$

where the rotation matrix for $\overline{\Lambda}$ is

$$\overline{G}_{NT} = \Delta_{NT}^2(\overline{H}_{NT})^{-1} = \Delta_{NT}\widehat{H}_{NT}^{-1}$$

Denote the probability limit of Δ_{NT} as $\Delta_\infty = (\mathbb{D}_r^{\gamma}\mathbb{D}_r^{-1})^{1/2}$, where $\mathbb{D}_r^2$ is the diagonal matrix consisting of the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and $\mathbb{D}_r^{\gamma} = (\mathbb{D}_r - \gamma I_r)_+$. Using Proposition 1, (18) and (19), we obtain the following result.

Proposition 2. *Let* $(\overline{F}, \overline{\Lambda})$ *be given in* (16a) and (16b) *with threshold parameter* $\gamma > 0$. *Suppose that* Assumption A *holds and* $N, T \to \infty$. *Then*

(i) $\sqrt{N}(\overline{F}_t - \overline{H}_{NT}' F_t^0) \xrightarrow{d} N\left(0, \Delta_\infty \textsc{Avar}(\widehat{F}_t)\Delta_\infty\right) \equiv N(0, \textsc{Avar}(\overline{F}_t))$

(ii) $\sqrt{T}(\overline{\Lambda}_i - \overline{G}_{NT}\Lambda_i^0) \xrightarrow{d} N\left(0, \Delta_\infty \textsc{Avar}(\widehat{\Lambda}_i)\Delta_\infty\right) \equiv N(0, \textsc{Avar}(\overline{\Lambda}_i))$.

Since the diagonal elements of Δ_∞ are less than 1, Proposition 2 implies that $\textsc{Avar}(\overline{F}_t) \le \textsc{Avar}(\widehat{F}_t)$, and $\textsc{Avar}(\overline{\Lambda}_i) \le \textsc{Avar}(\widehat{\Lambda}_i)$.

Turning to the common component, the RPC estimate is $\overline{C} = \overline{F}\,\overline{\Lambda}'$ and the PC estimate is $\widehat{C} = \widehat{F}\widehat{\Lambda}'$. Using Δ_{NT}^2 defined in (17), we see that

$$\overline{C} = \overline{F}\,\overline{\Lambda}' = \widehat{F}\Delta_{NT}^2\widehat{\Lambda}' \neq \widehat{F}\widehat{\Lambda}' = \widehat{C}.$$

Since $\widehat{C}$ is an asymptotically unbiased estimate for the corresponding element of $C^0 = F^0\Lambda^{0\prime}$, it follows that the elements of $\overline{C}$ are biased towards zero. From $\|\overline{C}\|_F^2 = \sum_{i=1}^N \sum_{t=1}^T \overline{C}_{it}^2 = \mathrm{trace}(\overline{C}'\overline{C})$, we have that

$$\frac{\mathrm{trace}(\overline{C}\overline{C}')}{\mathrm{trace}(XX')} = \frac{\mathrm{trace}((D_r^{\gamma})^2)}{\mathrm{trace}(D^2)} < \frac{\mathrm{trace}(D_r^2)}{\mathrm{trace}(D^2)} = \frac{\mathrm{trace}(\widehat{C}'\widehat{C})}{\mathrm{trace}(XX')}.$$

Thus $\overline{C}$ accounts for a smaller fraction of the variation in X.

We can also derive the limiting distribution of the estimated common components. Let $\overline{C}_{it} = \overline{F}_t'\overline{\Lambda}_i$ and $C_{it}^0 = F^{0\prime}\Lambda_i^0$. As shown in the Appendix:

$$(W_{NT}(\overline{C}_{it}))^{-1/2}(\overline{C}_{it} - C_{it}^0 - \mathrm{bias}) \xrightarrow{d} N(0, 1) \tag{20}$$

where $\mathrm{bias} = \gamma F_t^{0\prime}\overline{H}D_r^{-1}\overline{H}^{-1}\Lambda_i^0$ and

$$W_{NT}(\overline{C}_{it}) = \frac{1}{T}\overline{\Lambda}_i'\textsc{Avar}(\overline{F}_t)\overline{\Lambda}_i + \frac{1}{N}\overline{F}_t'\textsc{Avar}(\overline{\Lambda}_i)\overline{F}_t.$$

The convergence rate for $\overline{C}_{it}$ is thus $m_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. Though we can remove the bias by setting $\gamma \to 0$, γ cannot be too small if we were to avoid over-estimating the dimension of the common component. Specifically, we need $\gamma \ge O(m_{NT}^{-1})$ so that the estimated number of factors would not be contaminated by the idiosyncratic errors. This is needed because the largest singular value of $\frac{e}{\sqrt{NT}}$ (or the square root of the largest eigenvalue of $\frac{e'e}{NT}$) is no smaller than $O_p(m_{NT}^{-1})$.

Though $\overline{C}$ is an asymptotically biased estimator for C^0, $\overline{C}$ may not yield smaller mean-squared error than $\widehat{C}$ because $\overline{C}$ is based on the objective of finding a good fit subject to a minimum rank constraint. Indeed, its asymptotic mean squared error AMSE may or may not be smaller than that of the unbiased estimator $\widehat{C}$. To see why, consider (i, t)-th element: $\overline{C}_{it} = \overline{F}_t' \overline{\Lambda}_i = \widehat{F}_t' \Delta_{NT}^2 \widehat{\Lambda}_i$. Suppose that there is only a single factor $r = 1$, then

$$\overline{C}_{it} = \left(\frac{(d_1 - \gamma)_+}{d_1} \right) \widehat{C}_{it} \equiv \delta_1 \widehat{C}_{it}.$$

The asymptotic bias and variance of $\overline{C}_{it}$ are, respectively,

$$\begin{aligned} \text{ABIAS}(\overline{C}_{it}) &= (\delta_1 - 1) C_{it}^0 \\ \text{AVAR}(\overline{C}_{it}) &= \delta_1^2 \text{AMSE}(\widehat{C}_{it}). \end{aligned}$$

Thus the MSE of $\overline{C}_{it}$ is $\text{AMSE}(\overline{C}_{it}) = (\delta_1 - 1)^2 (C_{it}^0)^2 + \delta_1^2 \text{AMSE}(\widehat{C}_{it})$ so that

$$\frac{\text{AMSE}(\overline{C}_{it})}{\text{AMSE}(\widehat{C}_{it})} = (\delta_1 - 1)^2 \frac{(C_{it}^0)^2}{\text{AMSE}(\widehat{C}_{it})} + \delta_1^2.$$

As shown in Bai (2003), the asymptotic MSE of $\widehat{C}_{it}$ depends on Σ_Λ and Σ_F, and on the variance of idiosyncratic errors. But $\widehat{C}_{it}$ is asymptotically unbiased for C_{it}^0, and $\text{AMSE}(\widehat{C}_{it}) = \text{AVAR}(\widehat{C}_{it})$. Hence the relative AMSE can be less than one when the signal of the common component is weak, which can be due to a small Σ_Λ and/or a small Σ_F, or when the idiosyncratic error variance is large. These correspond to cases of small singular values in the low rank component and noise corruption that motivated RPCA in machine learning. Here, in addition to algorithmic advantages, regularized principal components analysis can also reduce MSE when the data are noisy and when the pervasive signals are weak.

It is noteworthy that soft-thresholding of the singular values is distinct from regularization of the singular vectors for a given rank of the low rank component, referred to in the literature as sparse principal components (SPC). The thresholding operation in SPC analysis does not change the rank of the factor estimates. It only performs shrinkage.[17] In contrast, SVT constrains the rank of the low rank component to be no larger than r as any factor corresponding to $d_i \leq \gamma$ will effectively be dismissed. It has efficiency implications since $D_r - D_r^\gamma \geq 0$ by construction. As a consequence, the asymptotic variance $\overline{F}_{jt}$ cannot exceed that of the unrestricted estimates $\widehat{F}_{jt}$ for all $j = 1, \ldots, r$.

4.2. Selection of factors

The problem defined in (14) penalizes components with small contributions to the low rank component. Optimality is defined from an algorithmic perspective that does not take into account that $(\overline{F}, \overline{\Lambda})$ are estimates. The larger is r, the better is the fit, but variance also increases with the number of factors be estimated. Bai and Ng (2002) suggest to determine the number of factors using criteria that take into account model complexity (hence sampling uncertainty). These take the form

$$\widehat{r} = \min_{k=0,\ldots,\text{rmax}} \widehat{IC}(k), \qquad \widehat{IC}(k) = \log(\text{SSR}_k) + k g(N, T).$$

The $\widehat{IC}_2$ criterion, often used in empirical work, obtains when

$$g(N, T) = \frac{(N + T)}{NT} \log\left(\frac{NT}{N + T} \right). \tag{21}$$

The SSR term in the criterion function is the sum of squared residuals from fitting a model with k factors. Suppose that each series in standardized, then $\|Z\|_F^2 = 1$; together with $\|\widehat{F} \widehat{\Lambda}\,'\| = \|D_r\|$, SSR_k can be written as

$$\text{SSR}_k = 1 - \sum_{j=1}^{k} d_j^2 = \left\| Z - \widehat{F}_k \widehat{\Lambda}_k' \right\|_F^2.$$

In the constrained problem (14) yields $\left\| \overline{F}\, \overline{\Lambda}\,' \right\|_F^2 = \left\| D_r^\gamma \right\|_F^2$. For given k and $\gamma > 0$,

$$\text{SSR}_k(\gamma) = 1 - \sum_{j=1}^{k} (d_j - \gamma)_+^2 = \left\| Z - \overline{F}_k \overline{\Lambda}_k' \right\|_F^2.$$

This suggests to define a class of criteria that takes into account both the rank of the common component and sampling uncertainty as follows:

$$\overline{r} = \min_{k=0,\ldots,\text{rmax}} \log\left(1 - \sum_{j=1}^{k} (d_j - \gamma)_+^2 \right) + k g(N, T). \tag{22}$$

[17] Sparse PCA is often motivated by easy interpretation of the factors. See, for example, Shen and Huang (2008).

In other words, SSR_k is evaluated at the rank restricted estimates $(\overline{F}, \overline{\Lambda})$. Taking the approximation $\log(1+x) \approx x$, we see that

$$\overline{IC}(k) = \widehat{IC}(k) + \gamma \sum_{j=1}^{k} \frac{(2d_j - \gamma)}{\widehat{\text{SSR}}_k}.$$

Since $d_j \geq d_j - \gamma \geq 0$, the penalty is heavier in $\overline{IC}(k)$ than $\widehat{IC}(k)$. The rank constraint adds a data dependent term to each factor to deliver a more conservative estimate of r. An appeal of the criterion is that the data-dependent adjustment does not require the researcher to be precise about the source of the small singular values. They can be due to genuine weak factors, noise corruption, omitted lagged and non-linear interaction of the factors that are of lesser importance.

The larger is γ, the stronger is the penalty. A natural question to ask is how to determine γ. Since X is standardized and $Z = \frac{X}{\sqrt{NT}}$, we have $\|Z\|_F^2 = 1$ by construction. Thus $|d_j|^2$ is the total variation of Z that the jth unconstrained factor will explain. Regularization reduces its contribution in a non-linear way. In applications, we set γ to 0.05. If $d_1 = 0.4$, the contribution of F_1 will fall from $.4^2 = 0.16$ to $(.4 - .05)^2 = .1225$. If $d_2 = 0.2$, the contribution of F_2 will fall from 0.04 to 0.025, a larger percentage drop than F_1. Any $d_j < .05$ will be truncated to zero. As we will see in simulations below, adding $\sum_{j=1}^{k}(2d_j - \gamma)$ to the penalty makes a non-trivial difference to the determination of r as we will see in simulations.

4.3. Practical considerations

Consider the infeasible linear regression $y_{t+h} = \alpha' F_t + \beta' W_t + \epsilon_{t+h}$ where $\underset{\sim}{W_t}$ are observed predictors. The regression is infeasible because F is latent. Bai and Ng (2006) show that the APC estimate $\widetilde{F}$ can be used in place of the latent F, and inference can proceed as though F was known provided $\frac{\sqrt{T}}{N} \to 0$. Section 2 shows that it is equally valid to replace F with the PC estimate $\widehat{F}$.

What happens if we replace F by the RPC estimate $\overline{F}$ in an augmented regression? It may be tempting to think that $\overline{F}$ will reduce the goodness of fit because $\text{var}(\overline{F}) < \text{var}(\widehat{F})$. This, however, is not true. Least squares estimation will give identical fit whether we use $\widetilde{F}$, $\widehat{F}$, or $\overline{F}$ as regressors. To appreciate this point, there are two cases to consider. First, suppose that SVT shrinks but does not threshold the singular values so that $\overline{F}$ and $\widehat{F}$ have the same rank. Then $\widetilde{F}$, $\widehat{F}$ and $\overline{F}$ will be perfectly correlated because they are all spanned by U_r. The estimates of α will simply adjust to compensate for the difference in scale. In the second case when $r^* = \dim(\overline{F}) < \dim(\widehat{F}) = r$ because of thresholding, the fit provided by $\overline{F}$ will remain identical to the fit provided by the first r^* columns of $\widehat{F}$ by virtue of the fact the omitted factors are orthogonal to the included ones. The upshot of the above discussion is that even though $\overline{F}$ has been regularized, it will give the same prediction as a factor augmented regression using $\widehat{F}$ as predictor unless an effort is made to shrink the coefficients associated with $\overline{F}$ towards zero. This can be achieved by replacing least squares with ridge regressions in estimation of the factor augmented models.

The second practical issue concerns construction of the factors. Whether we are interested in the APC, PC, or RPC estimates, the singular vectors U_r are required. This can be easily computed when the sample size is small. But for data of high dimension, the memory required in the computation is not trivial. A practical method to compute the first singular vector is *power iteration*. One starts with an initial vector that has a non-zero component in the direction of the target singular vector. It then recursively updates and re-normalizes the vector till convergence. The resulting vector is the largest singular vector. The idea has been extended to compute the invariant subspace of all singular vectors. Of interest is that this extension can be modified to construct robust principal components. The following algorithm is from Hastie et al. (2015).[18]

Algorithm RPC (Iterative Ridge): Given a $m \times n$ matrix Z, initialize a $m \times r$ matrix $F = \mathbb{U}\mathbb{D}$ where $\mathbb{U}$ is orthonormal and $\mathbb{D} = I_r$.

A. Repeat till convergence

   i. (solve Λ given F): $\widetilde{\Lambda} = Z\,' F(F\,' F + \gamma I_r)^{-1}$.
   ii (orthogonalize): Do $\text{SVD}(\widetilde{\Lambda}) = \widetilde{\mathbb{U}}_\Lambda \widetilde{\mathbb{D}}_\Lambda \widetilde{\mathbb{V}}_\Lambda\,'$ and let $\Lambda = \widetilde{\mathbb{U}}_\Lambda \widetilde{\mathbb{D}}_\Lambda$ and $\mathbb{D} = \widetilde{\mathbb{D}}_\Lambda$.
   iii (solve F given Λ): $\widetilde{F} = Z\Lambda(\Lambda\,'\Lambda + \gamma I_r)^{-1}$.
   iv (orthogonalize): Do $\text{SVD}(\widetilde{F}) = \widetilde{\mathbb{U}}_F \widetilde{\mathbb{D}}_F \widetilde{\mathbb{V}}_F\,'$ and let $F = \widetilde{\mathbb{U}}_F \widetilde{\mathbb{D}}_F$ and $\mathbb{D} = \widetilde{\mathbb{D}}_F$.

B. (Cleanup) From $\text{SVD}(Z\widetilde{\mathbb{U}}_\Lambda) = U_r D_r \mathbb{V}_r\,'$, let $V_r' = \mathbb{V}_r' \widetilde{\mathbb{U}}_\Lambda$ and $D_r^\gamma = (D_r - \gamma I_r)_+$.

Algorithm RPC uses iterative ridge regressions to construct the factors and the loadings. The two SVD steps ensure that the factors and loadings are mutually orthogonal. The converged result of Step A gives $(\overline{F}_z, \overline{\Lambda}_z)$, which is also the solution to the nuclear norm minimization problem stated in (14). In theory, this is all that is needed for construction of $\overline{F} = \sqrt{T}F_z$ and $\overline{\Lambda} = \sqrt{N}\Lambda_z$. Iterating till the eigenvalues are small by computer precision may be difficult. Explicit thresholding

[18] Our algorithm differs only in that we do SVD of F and Λ in Steps (ii) and (iv) instead of FD and ΛD.

of the singular values overcomes this numerical problem. The output from Step B is an approximation to the low rank component of minimum rank.[19]

The ridge regression perspective is useful in highlighting the role that regularization plays in RPC. It should, however, be noted that while a ridge regression alone performs shrinkage, the SVT solution that emerges from iterative ridge regressions satisfies (13). Hence it not only shrinks, but also thresholds the singular values. The final estimates of the factors and loadings that emerge are $\overline{F}_z = U_r(D_r^\lambda)^{1/2}$ and $\overline{\Lambda}_z = V_r(D_r^\lambda)^{1/2}$. The entire procedure only involves SVD for matrices of dimension $m \times r$ and $n \times r$, not dimension of $m \times n$. This is important when both m and n are large. Of course, when Z is not huge in dimension, $(\overline{F}_z, \overline{\Lambda}_z)$ can be directly computed from an SVD of Z, and the algorithm is not necessary.

Finally, a more general regularized problem we can consider is:

$$(\overline{F}_{\gamma_1,\gamma_2,\tau}, \overline{\Lambda}_{\gamma_1,\gamma_2,\tau}) = \underset{F,\Lambda}{\text{argmin}}\left(\frac{1}{2}\|Z - F\Lambda'\|_F^2 + \frac{\gamma_1}{2}\|F\|_F^2 + \frac{\gamma_2}{2}\|\Lambda\|_F^2\right)$$

where the weights γ_1 and γ_2 may not be equal. Let

$$\overline{D}_r^\gamma = (D_r - \sqrt{\gamma_1\gamma_2}\, I_r)_+.$$

The optimal solution is given by (see appendix for a proof)

$$\overline{F}_{\gamma_1,\gamma_2} = \left(\frac{\gamma_2}{\gamma_1}\right)^{1/4} U_r(\overline{D}_r^\gamma)^{1/2} \tag{23a}$$

$$\overline{\Lambda}_{\gamma_1,\gamma_2} = \left(\frac{\gamma_1}{\gamma_2}\right)^{1/4} V_r(\overline{D}_r^\gamma)^{1/2}. \tag{23b}$$

The corresponding common component is

$$\overline{C}_{\gamma_1,\gamma_2} = U_r\overline{D}_r^\gamma V_r'$$

When $\gamma_1 = \gamma_2 = \gamma$, the solution coincides with (15), which can be computed by Algorithm RPC.

5. Simulations and application

A small simulation exercise is used to highlight the issues. Data are generated according to

$$X_{it} = F_t^{0\prime}\Lambda_i^0 + e_{it} + s_{it}, \qquad e_{it} \sim (0,1)$$

where the sparse error $s_{it} \sim N(\mu, \omega^2)$ if $(i,t) \in \Omega$ and zero otherwise, Ω is an index set containing (i,t) positions with non-zero values of s_{it}. It is assumed a fraction κ_N of cross-section units have outliers in a fraction κ_T of the sample. In the simulations, we let $(\kappa_N, \kappa_T) = (0.1, 0.03)$, $\mu = 5$. Two data generating processes both with $r = 5$ are considered.

- DGP1: $F_t^0 \sim N(0, I_r)$, $\Lambda_i^0 \sim N(0, I_r)$, with $\omega \in (5, 10, 20)$;
- DGP2: $F^0 = U_r D_r^{1/2}$, $\Lambda^0 = V_r D_r^{1/2}$, with $\text{diag}(D_r) = [1, 0.8, 0.5, 0.3, 0.2\theta]$, and $\omega = 5$. Three values of θ are considered: $(1, 0.75, 0.5)$.

The first DGP is designed to study the effect of outliers, which is expected to lead to an over-estimation of r. The second DGP varies the contribution of the smallest factor by the parameter θ. The minimum rank is expected to decrease with θ. We define the minimum rank r^* to be the number of factors that contribute to at least 5% to the variance of the common component. For DGP 1 has $r^* = 5$ and DGP 2 has $r^* = 3$.

The properties of the factor selection rules depend on the strength of the factors as well as the extent of noise contamination. We summarize these features using three statistics. The first is C^r, which denotes the fraction of population variance X due to all r factors. The second is C_r, which denotes the fraction of variance due to rth (i.e. smallest) factor. These two indicate the relative importance of the common component and the smallest factor in the data, respectively. The third is $c(S)$, which denotes the fraction of variance of Z due to the outliers in S. We report the mean of $\widehat{r}$ and $\overline{r}$, the probability that $\widehat{r} = r$ and $\overline{r} = r$ in 5000 replications with *rmax* set to 8. To evaluate how well the estimated factors approximate the space spanned by the true factors, we regress $\widehat{F}_{\widehat{r}}$, the smallest factor as determined by $\widehat{IC}$, on all r^* singular vectors of the true common component. If the factor space is precisely estimated, the $R^2_{\widehat{r}}$ of this regression should be close to one. Results based on $\overline{F}$ are similarly defined. These are denoted $\widehat{R}^2_{\widehat{r}}$ and $\overline{R}^2_{\overline{r}}$ for the PC and RPC estimates of F, respectively. We also compute the average of the canonical correlations between $\widehat{F}_j$ and the space spanned by the true factors. This is denoted $\widehat{CC}$. The average canonical correlations between $\overline{F}_j$ and the true factors are denoted $\overline{CC}$.

Table 1 shows that in the absence of outliers, i.e. $c(S) = 0$, the IC performs well and always correctly selects $r = 5$ factors when N and T are both reasonably large. Rank regularization does not affect the number of factors selected in this

[19] In principle, the converged $\widetilde{\mathbb{U}}_\Lambda$ should be V_r. Step B essentially computes an improved estimate by performing a SVD of $Z\widetilde{\mathbb{U}}_\Lambda = U_r D_r \mathbb{V}_r{}'$, and then recovers V_r from the right eigenvector of $Z\widetilde{\mathbb{U}}_\Lambda$.

Table 1
DGP 1: $r^* = 5$.

Parameters			Signal		Noise	Mean		Spanning			
N	T	ω	C^r	C_r	$c(S)$	$\widehat{r}$	$\bar{r}$	$\widehat{R}^2_{\widehat{r}}$	$\bar{R}^2_{\bar{r}}$	$\widehat{CC}$	$\overline{CC}$
100	100	5.00	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	100	10.00	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	100	20.00	0.83	0.12	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	200	5.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	200	10.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	200	20.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	400	5.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	400	10.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
100	400	20.00	0.83	0.13	0.00	5.00	5.00	0.98	0.98	0.98	0.98
50	100	5.00	0.83	0.10	0.00	5.00	4.95	0.95	0.95	0.96	0.96
50	100	10.00	0.83	0.10	0.00	5.00	4.95	0.95	0.95	0.96	0.96
50	100	20.00	0.83	0.10	0.00	5.00	4.95	0.95	0.95	0.96	0.96
50	200	5.00	0.83	0.11	0.00	5.02	5.00	0.93	0.96	0.96	0.97
50	200	10.00	0.83	0.11	0.00	5.02	5.00	0.93	0.96	0.96	0.97
50	200	20.00	0.83	0.11	0.00	5.02	5.00	0.93	0.96	0.96	0.97
50	400	5.00	0.83	0.11	0.00	5.05	5.00	0.91	0.96	0.96	0.97
50	400	10.00	0.83	0.11	0.00	5.05	5.00	0.91	0.96	0.96	0.97
50	400	20.00	0.83	0.11	0.00	5.05	5.00	0.91	0.96	0.96	0.97
100	100	5.00	0.81	0.12	0.02	5.36	5.00	0.63	0.98	0.92	0.98
100	100	10.00	0.78	0.12	0.06	5.79	5.00	0.28	0.98	0.85	0.97
100	100	20.00	0.69	0.12	0.17	6.81	5.00	0.01	0.97	0.72	0.97
100	200	5.00	0.81	0.13	0.02	5.67	5.00	0.32	0.98	0.87	0.98
100	200	10.00	0.78	0.13	0.06	5.91	5.00	0.19	0.98	0.84	0.98
100	200	20.00	0.69	0.13	0.17	7.13	5.00	0.00	0.98	0.69	0.98
100	400	5.00	0.81	0.13	0.02	5.88	5.00	0.12	0.98	0.84	0.98
100	400	10.00	0.78	0.13	0.06	5.90	5.00	0.16	0.98	0.84	0.98
100	400	20.00	0.69	0.13	0.18	7.15	5.00	0.00	0.98	0.69	0.98
50	100	5.00	0.81	0.10	0.02	5.32	4.92	0.65	0.95	0.91	0.96
50	100	10.00	0.78	0.10	0.06	5.69	4.89	0.35	0.95	0.85	0.96
50	100	20.00	0.69	0.10	0.17	6.39	4.83	0.04	0.94	0.76	0.96
50	200	5.00	0.81	0.11	0.02	5.42	4.99	0.57	0.95	0.90	0.96
50	200	10.00	0.78	0.11	0.06	5.71	4.99	0.35	0.95	0.85	0.96
50	200	20.00	0.69	0.11	0.17	6.58	4.98	0.04	0.94	0.74	0.96
50	400	5.00	0.81	0.11	0.02	5.54	5.00	0.47	0.95	0.88	0.97
50	400	10.00	0.78	0.11	0.06	5.71	5.00	0.35	0.95	0.86	0.97
50	400	20.00	0.69	0.11	0.17	6.66	5.00	0.04	0.94	0.73	0.96

Notes: $X_{it} = F_t^{0'}\Lambda_i^0 + e_{it} + s_{it}$, $e_{it} \sim (0, 1)$, $s_{it} \sim (0, \omega^2)$. F_t^0 is $r \times 1$, $r = 5$. Let $C^0 = F^0\Lambda^{0\prime} = U_r D_r V_r'$. Then $r^* = \sum_{j=1}^r 1(\frac{d_i^2}{\sum_{k=1}^r d_i^2} > \gamma)$ with $\gamma = 0.05$. Let F_r^0 be the rth column of F^0, $C^r = \frac{\text{var}(C^0)}{\text{var}(X)}$ $C_r = \frac{\text{var}(F_r^0\Lambda_r^{0'})}{\text{var}(X)}$, $c(S) = \frac{\text{var}(S)}{\text{var}(X)}$. The column labeled $R^2_{\widehat{r}}$ is the R^2 from a regression of the smallest factor on U_r. The column $\widehat{CC}$ is the average canonical correlation between $\widehat{r}$ factors. Results for $\bar{r}$ are similarly defined.

setting. However, when noise corruption is present and $c(S) > 0, \widehat{r}$ tends to exceed r and has a mean of over 6. The higher is $c(S)$, the larger is the contribution of the outliers, and the larger is $\widehat{r}$. However, $\bar{r}$ is more robust and correctly selects five factors in many cases. When too many factors are estimated, the smallest unregulated factor is not well correlated with the true factors, and the average canonical correlation between the true and estimated factors is well below one. In contrast, the $\bar{F}_{\bar{r}}$ is well spanned by the true factors, and the average canonical correlation between $\bar{F}$ and the true factors is close to one.

Table 2 shows results for DGP 2 which has $r = 5$ but $r^* = 3$. This means that the smallest two factors contribute less than $c(S) = 0.05$ of the variation in C. Even in the absence of measurement noise, $\widehat{r}$ has a mean of four, implying that it tends to accept at least one of the small factors as valid. In contrast, $\bar{r}$ which has a mean of three tends to disregard both small factors. When measurement errors are allowed as in the bottom panel, $\widehat{r}$ tends to find yet an additional factor compared to the top panel. In contrast, $\bar{r}$ is unaffected by noise contamination. Of course, for this DGP, the true factor is r and one can argue that $\widehat{r}$ yields the correct estimate. As there is a tension between consistent estimation of r on the one hand, and minimum rank/parsimony on the other, it is up to the user whether to use $\widehat{r}$ or $\bar{r}$ in this case.

The first empirical application uses data from FRED-MD (McCracken and Ng (2016)), a macroeconomic database consisting of a panel of 134 series over the sample 1960M1–2016M08. Consistent with previous studies, some series are transformed by taking logs and first differencing before the factors are estimated. The panel is not balanced, we use the EM algorithm suggested in Stock and Watson (2002a) which imputes the missing values from the factor model and iterate till convergence. The nonbalanced panel has $N = 128$ variables with $T = 676$ observations. We also consider a balanced panel with $N = 92$ series.

Table 2
DGP 2: $r^* = 3$.

Parameters			Signal		Noise	Mean		Spanning			
N	T	ω	C^r	C_r	$c(S)$	$\hat{r}$	$\bar{r}$	$\hat{R}^2_{\hat{r}}$	$\bar{R}^2_{\bar{r}}$	$\widehat{CC}$	$\overline{CC}$
100	100	1.00	0.67	0.02	0.00	3.94	3.00	0.07	0.95	0.74	0.96
100	100	0.75	0.67	0.01	0.00	3.95	3.00	0.05	0.95	0.73	0.96
100	100	0.50	0.67	0.01	0.00	3.97	3.00	0.04	0.95	0.73	0.96
100	200	1.00	0.67	0.02	0.00	4.01	3.00	0.00	0.95	0.73	0.97
100	200	0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
100	200	0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
100	400	1.00	0.67	0.02	0.00	4.26	3.00	0.00	0.95	0.69	0.97
100	400	0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
100	400	0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.95	0.73	0.97
50	100	1.00	0.67	0.02	0.00	3.55	2.57	0.41	0.93	0.81	0.95
50	100	0.75	0.67	0.01	0.00	3.60	2.62	0.37	0.93	0.80	0.95
50	100	0.50	0.67	0.01	0.00	3.64	2.66	0.33	0.93	0.79	0.95
50	200	1.00	0.67	0.02	0.00	3.95	2.97	0.06	0.91	0.72	0.94
50	200	0.75	0.67	0.01	0.00	3.96	2.98	0.04	0.91	0.72	0.94
50	200	0.50	0.67	0.01	0.00	3.97	2.98	0.04	0.91	0.72	0.94
50	400	1.00	0.67	0.02	0.00	4.00	3.00	0.01	0.91	0.71	0.95
50	400	0.75	0.67	0.01	0.00	4.00	3.00	0.01	0.91	0.71	0.95
50	400	0.50	0.67	0.01	0.00	4.00	3.00	0.01	0.91	0.71	0.95
100	100	1.00	0.60	0.02	0.11	4.81	2.93	0.01	0.93	0.61	0.96
100	100	0.75	0.59	0.01	0.11	4.84	2.95	0.01	0.93	0.60	0.96
100	100	0.50	0.59	0.01	0.11	4.86	2.96	0.01	0.93	0.60	0.96
100	200	1.00	0.60	0.02	0.11	5.01	3.00	0.01	0.93	0.58	0.96
100	200	0.75	0.59	0.01	0.11	5.00	3.01	0.01	0.93	0.58	0.96
100	200	0.50	0.59	0.01	0.11	5.00	3.01	0.01	0.93	0.58	0.96
100	400	1.00	0.60	0.02	0.11	5.21	3.10	0.00	0.84	0.56	0.94
100	400	0.75	0.59	0.01	0.11	5.00	3.12	0.00	0.83	0.58	0.94
100	400	0.50	0.59	0.01	0.11	5.00	3.13	0.00	0.82	0.58	0.93
50	100	1.00	0.60	0.02	0.11	4.18	2.27	0.16	0.94	0.69	0.95
50	100	0.75	0.59	0.01	0.11	4.23	2.31	0.15	0.93	0.68	0.95
50	100	0.50	0.59	0.01	0.11	4.28	2.34	0.14	0.93	0.67	0.95
50	200	1.00	0.60	0.02	0.11	4.82	2.83	0.02	0.89	0.59	0.94
50	200	0.75	0.59	0.01	0.11	4.84	2.86	0.02	0.89	0.59	0.94
50	200	0.50	0.59	0.01	0.11	4.86	2.87	0.02	0.89	0.59	0.94
50	400	1.00	0.60	0.02	0.11	4.97	2.97	0.01	0.89	0.57	0.94
50	400	0.75	0.59	0.01	0.11	4.96	2.98	0.01	0.89	0.57	0.94
50	400	0.50	0.59	0.01	0.11	4.97	2.98	0.01	0.89	0.57	0.94

Notes: $X = C^0 + e + s$, $e_{it} \sim (0, 1)$, $s_{it} \sim (0, \omega^2)$, $C^0 = U_r D_r V_r'$, $D_r = [1, .8, .5, .3, .2]$. Then $r^* = \sum_{j=1}^{r} 1(\frac{d_i^2}{\sum_{k=1}^{r} d_i^2} > \gamma)$ with $\gamma = 0.05$. Let F_r^0 be the rth column of F^0, $C^r = \frac{\text{var}(C^0)}{\text{var}(X)}$ $C_r = \frac{\text{var}(F_r^0 \Lambda_r^{0'})}{\text{var}(X)}$, $c(S) = \frac{\text{var}(S)}{\text{var}(X)}$. The column labeled $R^2_{\hat{r}}$ is the R^2 from a regression of the smallest factor on U_r. The column $\widehat{CC}$ is the average canonical correlation between $\hat{r}$ factors. Results for $\bar{r}$ are similarly defined.

The results are reported in the top panel of Table 3. The squared-singular values can be interpreted as the percent contribution of the factor to the variation of Z. Due to shrinkage, the regularized singular values $\bar{d}_i^2$ are always smaller than the unregularized ones by roughly $\gamma^2 = (0.05)^2$. The original $\widehat{IC}_2$ finds eight factors in the balanced panel. After regularization, $\overline{IC}_2$ finds three factors. In Gorodnichenko and Ng (2017), this difference between $\hat{r}$ and $\bar{r}$ is attributed to interactions of the level factors disguising as separate factors. Instability in the loadings, along with outliers may also contribute to the difference. We then use eight factors to impute missing values in the non-balanced panel. The $\overline{IC}_2$ criterion continues to find three factors in the resulting balanced panel. In this data, the first factor loads heavily on real activity variables, the second on interest rate spreads, and the third on prices.

The second data set consists of 148 monthly financial indicators used in Ludvigson et al. (2017). There is only one series with missing data, hence the balanced and non-balanced panels are almost identical. The bottom panel of Table 3 shows that the first unconstrained factor explains close to 70% of the variations in the data, and is significantly more important than the second factor, which explains less than 5%. Thresholding at $\gamma = 0.05$ yields only three factors because all factors with $\hat{d}_i^2$ less than $(0.05)^2 = 0.025$ are eliminated.

Even though both the macro and the finance datasets both find three factors, it is worth nothing that the conclusions are arrived in different ways. In the macro data, all eight factors have singular values that exceed $\gamma = 0.05$. It is the additional penalty for model complexity that results in three factors. In the financial dataset, there are only three factors with singular values that exceed $\gamma = 0.05$. The concern for model complexity does not change the conclusion on the number of factors in this case. As with any regularized estimation, the results will depend on γ and $g(N, T)$, and our analysis is not immune to this caveat. Nonetheless, our data are normalized such that $\|Z\|_F^2 = 1$. A factor that contributes less than γ^2 to the variance of the common component will be thresholded. The user can choose γ as is deemed appropriate. The $g(N, T)$s

Table 3
Singular values.

F_j	FRED-MD			
	Balanced: N=128		Non-Balanced: N=134	
j	$\widehat{d}_1^2$	$\overline{d}_1^2$	$\widehat{d}_1^2$	$\overline{d}_1^2$
1	0.1828	0.1426	0.1493	0.1131
2	0.0921	0.0643	0.0709	0.0468
3	0.0716	0.0473	0.0682	0.0446
4	0.0604	0.0384	0.0561	0.0349
5	0.0453	0.0265	0.0426	0.0245
6	0.0416	0.0237	0.0341	0.0182
7	0.0301	0.0152	0.0317	0.0164
8	0.0287	0.0143	0.0268	0.0129
r^*	8	3	8	3
	Financial data			
	Balanced: N=146		Non-Balanced: N=145	
	$\widehat{d}_1^2$	$\overline{d}_1^2$	$\widehat{d}_1^2$	$\overline{d}_1^2$
1	0.6896	0.6090	0.6800	0.6001
2	0.0464	0.0274	0.0447	0.0261
3	0.0341	0.0181	0.0337	0.0178
4	0.0138	0.0045	0.0141	0.0047
5	0.0114	0.0032	0.0133	0.0043
6	0.0092	0.0021	0.0109	0.0030
7	0.0072	0.0012	0.0090	0.0020
8	0.0066	0.0010	0.0075	0.0013
r^*	8	3	8	3

considered in Bai and Ng (2002) assume N and T are of comparable magnitudes. Gagliardini et al. (2018) suggest different $g(N,T)$ for $N \gg T$.

6. Conclusion

This paper considers estimation of approximate factor models by regularized principal component. This is useful when the idiosyncratic errors have large singular values such as due to extreme outliers, or when some factors have small singular values, such as when the loadings are small. A new class of factor selection criteria is proposed that will give more conservative estimates when the strong factor assumption is questionable. Our analysis provides a statistical view of matrix recovery algorithms and complements results in the machine learning literature. Incomplete data is a prevalent problem in empirical work. A promising idea is to extend the RPC framework to allow for missing values not necessarily missing at random. This is studied in Bai and Ng (2019).

Appendix

Proof of Lemma 1. Proof of part (i).

$$\widetilde{H}_{NT} = (\Lambda^{0\prime}\Lambda^0/N)(F^{0\prime}\widetilde{F}/T)D_r^{-2}$$

By the fact that D_r^2 is the matrix of eigenvalues of $\frac{XX'}{NT}$ associated with the eigenvectors $\widetilde{F}$ (and noting the normalization $\widetilde{F}'\widetilde{F} = TI_r$, we have $\widetilde{F}'(\frac{XX'}{NT})\widetilde{F} = TD_r^2$. Substituting $X = F^0\Lambda^{0\prime} + e$ into the above, we have

$$D_r^2 = (\widetilde{F}'F^0/T)(\Lambda^{0\prime}\Lambda^0/N)(F^{0\prime}\widetilde{F}/T) + \frac{1}{T}\widetilde{F}'ee'\widetilde{F}/(NT).$$

Since the second term is $o_p(1)$, we can substitute $D_r^{-2} = (F^{0\prime}\widetilde{F}/T)^{-1}(\Lambda^{0\prime}\Lambda^0/N)^{-1}(\widetilde{F}'F^0/T)^{-1} + o_p(1)$ into $\widetilde{H}_{NT}$ to give

$$\widetilde{H}_{NT} = (\widetilde{F}'F^0/T)^{-1} + o_p(1).$$

Denote

$$\widetilde{H}_{1,NT} = (\Lambda^{0\prime}\Lambda^0/N)(\widetilde{\Lambda}'\Lambda^0/N)^{-1}$$

Left and right multiplying $X = F^0\Lambda^{0\prime} + e$ by $\widetilde{F}'$ and Λ^0 respectively, dividing by NT, and using $\widetilde{\Lambda} = \widetilde{F}'X/T$, we obtain

$$\frac{\widetilde{\Lambda}'\Lambda^0}{N} = \frac{\widetilde{F}'F^0}{T}\frac{\Lambda^{0\prime}\Lambda^0}{N} + o_p(1).$$

Substituting $\left(\frac{\widetilde{\Lambda}\,'\Lambda^0}{N}\right)^{-1}=\left(\frac{\Lambda^{0\,\prime}\Lambda^0}{N}\right)^{-1}\left(\frac{\widetilde{F}\,'F^0}{T}\right)^{-1}+o_p(1)$ into $\widetilde{H}_{1,NT}$, we obtain

$$\widetilde{H}_{1,NT}=\left(\frac{\widetilde{F}\,'F^0}{T}\right)^{-1}+o_p(1).$$

Thus $\widetilde{H}_{NT}$ and $\widetilde{H}_{1,NT}$ have the same asymptotic expression. This proves part (i).

Proof of part (ii). The proof of part (i) shows that

$$(\Lambda^{0\,\prime}\Lambda^0/N)(\widetilde{\Lambda}\,'\Lambda^0/N)^{-1}=(\widetilde{F}'F^0/T)^{-1}+o_p(1)$$

Taking transpose and inverse, we have

$$F^{0\prime}\widetilde{F}/T=(\Lambda^{0\,\prime}\Lambda^0/N)^{-1}(\Lambda^{0\,\prime}\widetilde{\Lambda}/N)+o_p(1)$$

Substituting this expression into the original definition of $\widetilde{H}_{NT}$, we have

$$\widetilde{H}_{NT}=(\Lambda^{0\,\prime}\widetilde{\Lambda}/N)D_r^{-2}+o_p(1)$$

Now left multiplying the equation $X=F^0\Lambda^{0\,\prime}+e$ by $F^{0\,\prime}$ and right multiplying it by $\widetilde{\Lambda}$, dividing by NT, we obtain

$$F^{0\,\prime}X\widetilde{\Lambda}/(NT)=(F^{0\,\prime}F^0/T)(\Lambda^{0\,\prime}\widetilde{\Lambda}/N)+F^{0\,\prime}e\widetilde{\Lambda}/(NT)$$

But

$$X\widetilde{\Lambda}=X\widetilde{\Lambda}(\widetilde{\Lambda}'\widetilde{\Lambda})^{-1}(\widetilde{\Lambda}'\widetilde{\Lambda})=\widetilde{F}(\widetilde{\Lambda}'\widetilde{\Lambda})=\widetilde{F}D_r^2N$$

Thus we have

$$(F^{0\,\prime}\widetilde{F}/T)D_r^2=(F^{0\,\prime}F^0/T)(\Lambda^{0\,\prime}\widetilde{\Lambda}/N)+o_p(1)$$

Equivalently,

$$(F^{0\,\prime}F^0/T)^{-1}(F^{0\,\prime}\widetilde{F}/T)=(\Lambda^{0\,\prime}\widetilde{\Lambda}/N)D_r^{-2}+o_p(1)$$

But the right hand side is equal to $\widetilde{H}_{NT}+o_p(1)$. This completes the proof of (ii).

Analogously, $\widehat{H}_{1,NT}=\widehat{H}_{NT}+o_p(1)$ and $\widehat{H}_{2,NT}=\widehat{H}_{NT}+o_p(1)$. Consider the first claim. From $\widehat{\Lambda}=\widetilde{\Lambda}D_r^{-1/2}$, we have

$$(\Lambda^{0\prime}\Lambda^0)(\widehat{\Lambda}'\Lambda^0)^{-1}=(\Lambda^{0\prime}\Lambda^0)(\widetilde{\Lambda}'\Lambda^0)^{-1}D_r^{1/2}=\widetilde{H}_{NT}D_r^{1/2}+o_p(1)=\widehat{H}_{NT}+o_p(1)$$

the second equality uses Lemma 1(i), and the last equality uses the definition of $\widehat{H}_{NT}$. The proof of the second claim is similar by using $\widehat{F}=\widetilde{F}D_r^{1/2}$.

Proof of (19).

$$\begin{aligned}\overline{\Lambda}_i &= \Delta_{NT}\widehat{\Lambda}_i\\ &= \Delta_{NT}(\widehat{\Lambda}_i-\widehat{H}_{NT}^{-1}\Lambda_i^0+\widehat{H}_{NT}^{-1}\Lambda_i^0)\\ &= \Delta_{NT}(\widehat{\Lambda}_i-\widehat{H}_{NT}^{-1}\Lambda_i^0)+\Delta_{NT}\widehat{H}_{NT}^{-1}\Lambda_i^0\\ &= \Delta_{NT}(\widehat{\Lambda}_i-\widehat{H}_{NT}^{-1}\Lambda_i^0)+\Delta_{NT}^2(\overline{H}_{NT})^{-1}\Lambda_i^0\end{aligned}$$

Moving the second term to the left hand side, we obtain (19).

Proof of (20). For notational simplicity, write $\overline{H}$ for $\overline{H}_{NT}$, and $\overline{G}$ for $\overline{G}_{NT}$.

$$\begin{aligned}\overline{C}_{it}-F_t^{0\,\prime}\overline{H}\,\overline{G}\Lambda_i^0 =&\\ &=(\overline{F}_t-\overline{H}'F_t^0)'\overline{\Lambda}_i+F_t^{0\,\prime}\overline{H}(\overline{\Lambda}_i-\overline{G}\Lambda_i^0)\\ &=\sqrt{T}(\overline{F}_t-\overline{H}'F_t^0)'(\overline{\Lambda}_i/T^{1/2})+(F_t^{0\,\prime}\overline{H}/N^{1/2})\sqrt{N}(\overline{\Lambda}_i-\overline{G}\Lambda_i^0)\end{aligned}$$

From Proposition 2,

$$\sqrt{T}(\overline{F}_t-\overline{H}'F_t^0)\xrightarrow{d}N(0,Avar(\overline{F}_t)),\quad \sqrt{N}(\overline{\Lambda}_i-\overline{G}\Lambda_i^0)\xrightarrow{d}N(0,Avar(\overline{\Lambda}_i))$$

The two asymptotic distributions are independent because the first involves sum of random variables over the cross sections for period t, and the second distribution involves random variables over all time periods for individual i. Let

$$\overline{A}_{NT}=\frac{1}{T}\overline{\Lambda}_i'Avar(\overline{F}_t)\overline{\Lambda}_i+\frac{1}{N}\overline{F}_t'Avar(\overline{\Lambda}_i)\overline{F}_t$$

If we replace $Avar(\overline{F}_t)$ and $Avar(\overline{\Lambda}_i)$ by their estimated versions, then $\overline{A}_{NT}$ is the estimated variance of $\overline{C}_{it}-F_t^{0\,\prime}\overline{H}\,\overline{G}\Lambda_i^0$. Using argument similar to Bai (2003), we have

$$(\overline{A}_{NT})^{-1/2}(\overline{C}_{it}-F_t^{0\,\prime}\overline{HG}\Lambda_i^0)\xrightarrow{d}N(0,1)$$

Since

$$\overline{H}\,\overline{G} = \overline{H}\Delta_{NT}^2\overline{H}^{-1} = \overline{H}(I_r - \gamma D_r^{-1})\overline{H}^{-1} = I_r - \gamma\overline{H}D_r^{-1}\overline{H}^{-1}$$

This gives

$$F_t^{0\,\prime}\overline{H}\,\overline{G}\Lambda_i^0 = F_t^{0\,\prime}\Lambda_i^0 - \gamma F_t^{0\,\prime}\overline{H}D_r^{-1}\overline{H}^{-1}\Lambda_i^0 = C_{it}^0 - \gamma F_t^{0\,\prime}\overline{H}D_r^{-1}\overline{H}^{-1}\Lambda_i^0$$

Thus

$$(\overline{A}_{NT})^{-1/2}(\overline{C}_{it} - C_{it}^0 - bias) \xrightarrow{d} N(0, 1)$$

where $bias = \gamma F_t^{0\,\prime}\overline{H}D_r^{-1}\overline{H}^{-1}\Lambda_i^0$.

Proof of (23a) and (23b). Consider a change of variables $F = (\gamma_2/\gamma_1)^{1/4}\ddot{F}$ and $\Lambda = (\gamma_1/\gamma_2)^{1/4}\ddot{\Lambda}$. Then the objective function

$$\|Z - F\Lambda'\|_F^2 + \gamma_1\|F\|_F^2 + \gamma_2\|\Lambda\|_F^2$$

can be rewritten as

$$\|Z - \ddot{F}\ddot{\Lambda}'\|_F^2 + \sqrt{\gamma_1\gamma_2}\|\ddot{F}\|_F^2 + \sqrt{\gamma_1\gamma_2}\|\ddot{\Lambda}\|_F^2$$

This is an objective function with equal weights. The optimal solution is

$$\ddot{F} = U_r[(D_r - \sqrt{\gamma_1\gamma_2}I_r)_+]^{1/2}, \quad \ddot{\Lambda} = V_r[(D_r - \sqrt{\gamma_1\gamma_2}I_r)_+]^{1/2}$$

In terms of the original variables, the optimal solution is

$$F = (\gamma_2/\gamma_1)^{1/4}U_r[(D_r - \sqrt{\gamma_1\gamma_2}I_r)_+]^{1/2}, \quad \Lambda = (\gamma_2/\gamma_1)^{1/4}V_r[(D_r - \sqrt{\gamma_1\gamma_2}I_r)_+]^{1/2}$$

This gives (23a) and (23b).

References

Agarwal, A., Negahban, S., Wainwright, M., 2012. Noisy matrix decompositions via convex relation: Optimal rates in high dimensions. Ann. Statist. 40 (2), 1171–1197.

Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Neyman, J. (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. V. University of California Press, Berkeley, pp. 114–150.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71 (1), 135–172.

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 with factor-augmented regressions. Econometrica 74 (4), 1133–1150.

Bai, J., Ng, S., 2008. Large dimensional factor analysis. Found. Trends Econom. 3 (2), 89–163.

Bai, J., Ng, S., 2013. Principal components estimation and identification of the factors. J. Econometrics 176, 18–29.

Bai, J., Ng, S., 2019. Low Rank Decomposition and Factor Analysis with Missing Data. mimeo, Columbia University.

Bentler, ., 1972. A lower bound method for the dimension-free measurement of internal consistency. Soc. Sci. Res. 1, 343–357.

Bentler, P., Woodward, J., 1980. Inequalities among lower bounds to reliability: With applications to test construction of factor analysis. Psychometrika 45, 249–267.

ten Berge, J., Kiers, H., 1991. A numerical approach to the exact and the approximate minimum rank of a covariance matrix. Psychometrika 56, 309–315.

Bertsimas, D., Copenhaver, M., Mazumder, R., 2016. Certifiably optimal low rank factor analysis, arXiv:160406837v1.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis. J. Econometrics 132, 169–194.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2010. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1), 1–122.

Cai, J., Candes, E., Shen, Z., 2008. A singular value thresholding algorithm for matrix completion. SIAM J. Opt. 40 (4), 1956–1982.

Candes, E., Li, X., Ma, Y., Wright, J., 2011. Robust principal compoennt analysis. J. ACM 58 (3), 11.

Candes, E.J., Recht, B., 2011. Exact matrix completion via convex optimization. J. ACM 58, 11–37.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean–variance analysis in large asset markets. Econometrica 51, 1281–2304.

Connor, G., Korajzcyk, R., 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. J. Financ. Econ. 15, 373–394.

Delvin, S., Gnanadesikan, R., Kettering, J., 1981. Robust estimation of dispersion matrices and principal components. J. Amer. Statist. Assoc. 354–362.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 (211–8).

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc., Ser. B 75 (4), 603–680.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic factor model: Identification and estimation. Rev. Econom. Stat. 82 (4), 540–554.

Gagliardini, P., Ossola, E., Scaillet, O., 2018. A diagnostic criterion for approximate factor structure.

Gorodnichenko, Y., Ng, S., 2017. Level and volatility factors in macroeconomic data. J. Monetary Econ. forthcoming.

Grant, M., Boyd, S., 2015. The cvx users' guide. . http://cvxr.com/cvx/doc/cvxpdf.

Guttman, L., 1958. To what extent can communalities reduce rank. Psyhhometrika 23 (4), 297–308.

Hastie, T., Mazumder, R., Lee, J., Zadeh, R., 2015. Matrix completion and low rank svd via fast alternating least squares. J. Mach. Learn. Res. 16, 3367–3402.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

Hubert, M., Rousseeuw, ., 2005. Robpca: A new approach to robust principal analysis. J. Amer. Statist. Assoc. 47 (64–79).

Hubert, M., Rousseeuw, P., Branden, K.V., 2005. Robpca: A new approach to robust principal component analysis. Technometrics 47 (1), 64–79.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer Series in Statistics, New York.

Joreskog, K., 1967. Some contribution to maximum likelihood factor analysis. Psychometrika 32 (443–482).

Lawley, D.N., Maxwell, A.E., 1971. Factor Analysis in a Statistical Method. Butterworth, London.

Lettau, M., Pelger, M., 2017. Estimating Latent Asset-Pricing Factors. Stanford University, mimeo.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal componets: Primary theory and monte carlo. J. Amer. Statist. Assoc. 80 (391), 759–766.

Lin, Z., Chen, M., Ma, Y., 2013. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:10095055v3.

Ludvigson, S., Ma, S., Ng, S., 2017. Uncertainty and business cycles: Exogenous impulse or endogenous response. NBER Working Paper 21803.

Ma, Y., Genton, M., 2001. Highly robust esetimation of dispersion matrices. J. Multivariate Anal. 78, 11–36.

McCracken, M., Ng, S., 2016. Fred-md: A monthly database for macroeconomic research. J. Bus. Econom. Stat. 36 (4), 574–589.

Negahban, S., Wainwright, M., 2012. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Mach. Learn. Res. 13, 1665–1697.

Onatski, A., 2011. Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak. manuscript under revision.

Rennie, J., Srebro, J., 2005. Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the 22nd International Conference on Machine Learning.

Saunderson, J., Chandrasekaran, V., Parrilo, P., Willsky, S., 2012. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. SIAM J. Matrix Anal. Appl. 38 (4), 1395–1415, arXiv:12041220v1.

Shapiro, A., 1982. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. Psychometrika 47 (2), 187–199.

Shapiro, A., ten Berge, J.M., 2000. The asymptotic bias of minimum trace factor analysis with applications to the gregest lower bound to reliability. Psychometrika 65 (3), 413–425.

Shen, H., Huang, J., 2008. Sparse principal component analysis via regularized low rank matrix approximations. J. Multivariate Anal. 99, 1015–1034.

Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97 (460), 1167–1179.

Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20 (2), 147–162.

Tillman, A., Pfetsch, M., 2014. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Trans. Inform. Theory 60 (2), 1248–1259.

Todeschini, A., Caron, F., Chavent, M., 2013. Probabilistic Low Rank Matrix Completion with Adaptive Spectral Regularization Algorithms, Vol. 26. Adances in Neural Information Processing Systems (NIPS),

Udell, M., Horn, C., Zadeh, R., Boyd, S., 2016. Generalized low rank models. Found. Trends Mach. Lear. 9 (1), 1–118, arXiv:14100342v4.

Wright, J., Ma, W., Ganesh, A., Rao, S., 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimzation. In: Proceedings of Neural Information Processing Systems 1–0.

Yang, D., Ma, Z., Buja, A., 2014. A sparse singular value decomposition method for high dimensional data. J. Comput. Graph. Statist. 23 (4), 923–942.

Zhou, Z., Li, X., Wright, J., Candes, E., Ma, Y., 2010. Stable principal component pursuit. Int. Symp. Inform. Theory (ISIT) 1518–1522.