PRINCIPAL COMPONENTS AND REGULARIZED ESTIMATION OF FACTOR MODELS

Jushan Bai* Serena Ng†

September 19, 2017

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 shrinking the singular values of the common component and possibly reducing its rank. The method is used in the machine learning literature to recover low-rank matrices. We study the procedure from the perspective of estimating a minimum-rank approximate factor model. We show that the constrained factor estimates are biased but can be more efficient in terms of mean-squared errors. Rank consideration suggests a data-dependent penalty for selecting the number of factors. The new criterion is more conservative in cases when the nominal number of factors is inflated by the presence of weak factors or large measurement noise. The framework is extended to incorporate a priori linear constraints on the loadings. We provide asymptotic results that can be used to test economic hypotheses.

JEL Classsification: C30, C31

Keywords: singular-value thresholding, robust principal components, lowrank decomposition.

*Columbia University, 420 W. 118 St. MC 3308, New York, NY 10027. Email: jb3064@columbia.edu

†Columbia University and NBER, 420 W. 118 St. MC 3308, New York, NY 10027. Email: serena.ng@columbia.edu

The authors would like to thank Zongming Ma for help discussions. This work is supported by the National Science Foundation SES-1658770 (Bai), SES-1558623 (Ng).

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 X 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' or $X'X$. 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 that can be understood as performing iterative ridge instead of least squares regressions. Ridge regressions are known to shrink the parameter estimates of a linear model towards zero. They are biased but are less variable. In the present context, iterative ridge regressions will shrink the eigenvalues of the common component towards zero. Hastie et al. (2015) shows that when combined with a cleanup step that explicitly sets the small eigenvalues to zero, it 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.

This paper makes several contributions. The first is to provide a statistical analysis of RPC that complements the results developed from a machine learning perspective.[1] Constrained esti-

[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).

mation 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 provide a frequentist framework for regularized factor analysis. Economic theory often suggests a priori restrictions in the form of single or cross-equation restrictions. We provide the inferential theory that permits testing of economic hypothesis in the form of general linear restrictions on the loadings, with or without rank constraints.

Our third 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 term due to the desire for a minimal rank common component 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 the assumptions in the factor model are violated.

The follow notation will be used in what follows. We use the (T, N) to denote sample size of X 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 $U = [u_1, \ldots, u_m)$ is $m \times m$ and $V = [v_1, \ldots, v_n]$ 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 $d_1, d_2, \ldots, d_{\min(m,n)}$ of Z. The left singular vectors of ZZ' are the same as those 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. The rank of Z is the largest number of columns of Z that are linearly independent and can be evaluated by the number of non-zero singular values, which can also represented as $\|d(Z)\|_0$. If Z has rank r, we use the partition $U = [U_r; U_{n-r}]$ 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. Since $Z = U_r D_r V_r{}' + U_{n-r} D_{n-r} V_{n-r}{}'$, it suffices to consider the truncated SVD of Z, being $U_r D_r V_r{}'$, when $d_j = 0$ for $j > r$. If the rank of Z is no smaller than r, then $U_r D_r V_r'$ is the optimal rank-r approximation of Z under the Frobenius norm.

An important step in our analysis is to make use of results derived previously for the method asymptotic principal components (APC). In the next section, we show that variants of principal components differ from APC by the normalization used.

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

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

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 $X_t = (X_{1t}, X_{2t}, ..., X_{Nt})'$ 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.

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 diverges as N increases, while the $r+1$ largest eigenvalue is 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): $E||F_t^0||^4 \leq M$, $||\Lambda_i|| \leq \overline{\Lambda}$, $\frac{F^{0'}F^0}{T} \xrightarrow{p} \Sigma_F > 0$, and $\frac{\Lambda^{0'}\Lambda^0}{N} \xrightarrow{p} \Sigma_\Lambda > 0$.

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

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

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

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

(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) $E|N^{-1/2}\sum_{i=1}^N |\sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]^4 \leq M$ 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. Part (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) shows that r can be consistently estimated. In estimation of the F and Λ, the number of factors r is typically treated as known.

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

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

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 $XX'/(NT)$. 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) shows 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

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

[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 $\frac{\Lambda'\Lambda}{N}=I_r$.

This 'big data blessing' has generated a good deal of econometric research in the last two decades.[3] As as explained in Bai and Ng (2013), $\widetilde{H}_{NT}$ will not, in general, be an identity matrix implying that the j-th factor estimate $\widetilde{F}_j$ will not, in general, equal F_j^0 even asymptotically. The exception is when the true F^0 is 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 F is not needed as in 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 plim $_{N,T\to\infty} \frac{\widetilde{F}'F^0}{T} = \mathbb{Q}_r$, plim $_{N,T\to\infty} D_r^2 = \mathbb{D}_r^2$, and

$$\begin{aligned}\sqrt{N}(\widetilde{F}_t - \widetilde{H}_{NT}'F_t^0) &\xrightarrow{d} \mathcal{N}\Big(0, \mathbb{D}_r^{-2}\mathbb{Q}_r\Gamma_t\mathbb{Q}_r'\mathbb{D}_r^{-2}\Big) \equiv \mathcal{N}(0, \textsc{Avar}(\widetilde{F}_t))\\ \sqrt{T}(\widetilde{\Lambda}_i - \widetilde{H}_{NT}^{-1}\Lambda_i^0) &\xrightarrow{d} \mathcal{N}\Big(0, (\mathbb{Q}_r')^{-1}\Phi_i\mathbb{Q}_r^{-1}\Big) \equiv \mathcal{N}(0, \textsc{Avar}(\widetilde{\Lambda}_i))\end{aligned}$$

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.

While $\widetilde{H}_{NT}$ is widely used in asymptotic analysis, it is difficult to interpret. It is useful to consider the following asymptotically equivalent rotation matrices:

Lemma 1 *Let* $\widetilde{H}_{1,NT} = (\Lambda^{0\prime}\Lambda^0)(\widetilde{\Lambda}'\Lambda^0)^{-1}$ *and* $\widetilde{H}_{2,NT} = (F^{0\prime}F^0)^{-1}(F^{0\prime}\widetilde{F})$. *Then (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. 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 eigenvectors 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)

[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) shows consistency of the factor estimates when the principal components are constructed from the population covariance matrix.

define principal components as $U_r D_r$. Nonetheless, both definitions are valid and differ in the normalization used. We now consider yet another definition of principal components for reasons that will soon be obvious.

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. To facilitate asymptotic analysis, 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 eigenvectors of Z are the same as those for X. However, 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 define the PC estimates as:

$$\widehat{F}_z = U_r D_r^{1/2} \tag{6a}$$

$$\widehat{\Lambda}_z = V_r D_r^{1/2}. \tag{6b}$$

Notably, the PC and the 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 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

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

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

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 difficult to impose restrictions on the APC estimates of F, a limitation that is important for the analysis to follow.

[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 eigen on the covariance matrix. This is generally the preferred method for numerical accuracy.

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

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

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 and both N and T go to infinity. Then the PC-estimates $\widehat{F}$ and $\widehat{\Lambda}$ satisfy:*

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

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

Following arguments in Bai and Ng (2006), $\widehat{F}$ can be used in subsequent regressions as though they were F^0 provided $\frac{\sqrt{T}}{N} \to 0$.

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 large dimensional factor analysis, it is customary to choose an $r \in [0, \text{rmax}]$ that provides a good fit while taking model complexity into account. Bai and Ng (2002) suggests 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 r-th eigenvalue (which should increase with N). This is because outliers can increase the variance in an otherwise uninformative direction, and PCA is blind to the source of the variance, Hubert and

Rousseeuw (2005). The problem is well known, but is not given attention in the econometrics literature. As will be shown below, the Bai-Ng criteria tend to find more factors than are the truth when there are weak factors and/or outliers. A more conservative criteria that guard against these without pre-specifying the source of over-estimation is desirable. Such a problem has actually been considered under different motivations. We briefly three: minimum rank factor analysis, matrix completion, and low rank matrix decompositions.

Factor analysis has its roots in the study of personality traits. The original goal of factor analysis was to decompose Σ_X into a non-negative definite low rank communality matrix Σ_C, and a diagonal matrix Σ_e that is positive definite so that the Haywood case of negative error variances does not arise. The smallest rank that solves this problem is known as the minimum rank.[5] But research such as by Guttman (1958) suggested that the number of factors in the data is rather large, making the notion of a minimum rank empirically vacuous. Additionally, the rank of a matrix is the number of non-zero singular values, or $||D||_0$. The rank (or cardinality) function is non-convex and non-differentiable. Imposing a minimum rank constraint on the decomposition is a NP hard problem. Though interests in minimum rank factor models have waned, two subsequent developments are of interest. The first is minimum trace factor analysis (MTFA) due to Bentler (1972). MFTA finds a diagonal matrix Σ_e to minimize $\text{TRACE}(\Sigma_X - \Sigma_e) = \sum_{i=1}^n D_{ii}(\Sigma_X - \Sigma_e)$ subject to the constraints that $\Sigma_X = \Sigma_C + \Sigma_e$, $\Sigma_X - \Sigma_e \succeq 0$, $\Sigma_e \succeq 0$ and diagonal.[6] The second is the (approximate) minimum rank factor analysis (MFRA) due to ten Berge and Kiers (1991), and Shapiro and ten Berge (2000). MFRA starts with the premise that though many common factors many exist, only part of Σ_C will be explained. In their setup,

$$\Sigma_X = \Sigma_{C^*} + (\Sigma_C - \Sigma_{C^*}) + \Sigma_e.$$

The quantity Σ_{C^*}, of rank $r^* \leq r$, is an approximation of $\Sigma_X - \Sigma_e$. MFRA then minimizes $\sum_{i=r+1}^n D_{ii}(\Sigma_X - \Sigma_e)$, which is the unexplained common variance, subject to the constraints that $\Sigma_X - \Sigma_e \succeq 0$, $\Sigma_e \succeq 0$.

What makes MFTA and MFRA interesting is that the sum of eigenvalues are convex functions. 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 used in convex optimization. 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 also the key insight of recent work on matrix

[5] ten Berge and Kiers (1991) define the approximate minimum rank of a $n \times n$ matrix Σ_X to be the smallest r s.t. for some tolerance $\delta \geq 0$, the minimum of $f(\Sigma_e) = \sum_{i=r+1}^n D_{ii}(\Sigma_X - \Sigma_e) \leq \delta$, and (i) $\Sigma_X - \Sigma_e \geq 0, (ii)\Sigma_e \geq 0$. Minimum rank is a special case with $\delta = 0$.

[6] .When the constraint $\Sigma_e > 0$ is added, it is known as CMFTA. See Bentler and Woodward (1980); Shapiro (1982); Shapiro and ten Berge (2000).

recovery that we will subsequently use to estimate factor models with minimum rank. We now turn to this work.[7]

3.1 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. To proceed, it is assumed that the matrix M to be recovered has low rank. For the Netflix problem, the low rank assumption amounts to requiring that preferences are defined over a small number of features (such as genres, leading actor/actress). The Netflix problem is then to complete M by finding a Z that is factorizable into two matrices (for preference and features) with smallest rank possible, and such that $Z_{ij} = M_{ij}$ for each (i, j) that is observed. But, as noted earlier, rank minimization is NP-hard. The breakthrough is to replace rank minimization by the nuclear norm minimization. This is important because the nuclear norm, which is also the sum of the eigenvalues, is convex. Candes and Recht (2011) shows M 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.

In exact matrix completion, the problem arises because of incomplete observations, but the 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, a $m \times n$ 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 $\|S\|_0$. The goal is to recover L from data that are sparsely corrupted by noise. While the Eckart-Young theorem states that $U_r D_r V_r'$ provides the best low rank approximation of Z in the spectral norm, eigen-decompositions are 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

[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).

data.[8] 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 pursuit (PCP):

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

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 rank constraint on L and the cardinality constraint on S have been replaced by convex functions. 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.[9] Zhou et al. (2010) allows for the presence of small noise W. With $Z = L + S + W$, it is shown that L and S can still be recovered with high probability upon solving the convex program

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

if, in addition to the incoherence and cardinality conditions, $\|W\| \leq \delta$ 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 $m \times n$ matrix Z of rank r and $\text{SVD}(Z) = U_r D_r V_r' + U_{n-r} D_{n-r} V_{n-r}'$, define

$$D_r^{\gamma} = \Big[D_r - \gamma I_r\Big]_+ \equiv \max(D_r - \gamma I_r, 0). \tag{8}$$

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

[9] 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.

Importantly, the SVT is the proximal operator associated with the nuclear norm.[10] Theorem 2.1 of Cai et al. (2008) shows that

$$U_r D_r^{\gamma} V_r' = \min_L \gamma \|L\|_* + \frac{1}{2} \|Z - L\|_F^2 . \tag{9}$$

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.[11]

3.2 Relation to Factor Models

The previous subsection provides results for recovery of the low rank and spares components, L and S. 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.

$$\min_{A,B} \frac{1}{2} \left\|Z - AB'\right\|_F^2 + \gamma \left\|AB'\right\|_* \tag{10}$$

has solution

$$\overline{A} = U_r (D_r^{\gamma})^{1/2}, \quad \overline{B} = V_r (D_r^{\gamma})^{1/2}, \tag{11}$$

and that $\overline{L} = \overline{A}\ \overline{B}'$ also solves (9). The result that $(\overline{A}, \overline{B})$ solves the problem (10) if and only if $\overline{L} = \overline{A}\ \overline{B}'$ solves (9) 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,

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

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

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

[10] 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).

[11] Algorithms that solve (9) include Augmented Lagrange Multiplier and Accelerated Proximal Gradient (Lin et al. (2013)), ADMM (Boyd et al. (2010)), and CVX (Grant and Boyd (2015)).

with equality when $A = U_r D_r^{1/2}$ and $B = V_r D_r^{1/2}$. Hence (10) is just a reformulation of (9) in terms of A and B. Consider now the first order conditions associated with (10). If $\overline{A}$ and $\overline{B}$ are solutions, it must be that $-(Z - \overline{A}\overline{B}')\overline{B} + \gamma\overline{A} = 0$ and $-(Z - \overline{A}\overline{B}')'\overline{A} + \gamma\overline{B} = 0$. 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

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

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 $\sqrt{d_i - \gamma}$ for positivity. Collecting the thresholded singular-values into D_r^γ, the solution defined in (11) obtains. These are the robust principal components of Z under the assumed normalization that $\overline{A}'\overline{A} = \overline{B}'\overline{B} = \overline{D}_r^\gamma$.[12]

4 Constrained Approximate Factor Models

An important result in 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 unsupervised learning literature is that robust principal components can be obtained via SVT for a given sample of data, as reviewed in Section 3. One might expect RPC to play a role in rank constrained factor analysis. Indeed there is one, and this section makes this precise. Given that RPC is a method developed in the machine learning literature, 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: $X = F\Lambda' + e$. We use Assumption A to restrict the factors, loadings, and the idiosyncratic noise so that the eigenvalues 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

[12] Udell et al. (2016) referred to (10) as *quadratically regularized PC*. See their Appendix A for a complete proof.

no reference to eigenvalues. But when the data are noisy rather missing, one can only hope to recover an approximate (rather than the exact) low rank component.[13] What matters most then is the singular vectors associated with the large singular values. For such problems, which seem more comparable to our setup, Agarwal et al. (2012) obtains an error bound in the Frobenius norm for L and for S under the assumption that $\|L\|_\infty \leq \frac{\alpha}{\sqrt{mn}}$. 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 $\|L\|_\infty \leq \frac{\alpha}{\sqrt{mn}}$ is effectively a restriction on the sum of the eigenvalues of the low rank component L.

What transpires from the above discussion is that machine learning methods restrict the sample eigenvalues of L and obtain finite sample error bounds. In contrast, approximate factor analysis puts restrictions on the population eigenvalues 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 as one 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 Robust Principal Components (RPCA): $(\overline{F}, \overline{\Lambda})$

This subsection considers estimation of at most r factors. In some cases, economic theory may suggest r factors but only $r^* < r$ may be empirically relevant. In other cases, economic theory may suggest r^* factors, as in affine term structure models for interest rates. But financial data tend to have fat tails. Extreme values may lead us to the discovery of more than r^* factors from the data. As principal components are blind as to whether pervasive variations or extreme events underlie the components, it is desirable to have a way to guard against noise corruption.

One way to think about the noise-corruption problem is that the principal component estimates are not efficient when there is significant heterogeneity in the idiosyncratic errors. Boivin and Ng (2006) considers re-weighting each series by the standard-deviation of the idiosyncratic error, obtained from a preliminary (unweighted) estimation of the factor model. A drawback is that these weights are themselves sensitive to outliers. A second approach is the POET estimator considered

[13] 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), .

in Fan et al. (2013), which uses thresholding to enforce sparsity of the idiosyncratic errors of a spiked-covariance matrix.

We propose to to apply thresholding to Σ_C rather than Σ_e. This is appealing because the common component is the object of interest, not the idiosyncratic errors. To do so requires that we map the scaled factor model given in (5) into the problem defined in (10). 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^* by the method of robust principal components (RPC). From the previous subsection, the rank regularized problem is

$$(\overline{F}_z, \overline{\Lambda}_z) = \text{argmin}_{F,\Lambda} \frac{1}{2}\Big(\left\|Z - F\Lambda'\right\|_F^2 + \gamma \left\|F\right\|_F^2 + \gamma \left\|\Lambda\right\|_F^2 \Big), \tag{12}$$

with optimal solution,

$$\overline{F}_z = U_r(D_r^\gamma)^{1/2}, \quad \overline{\Lambda}_z = V_r(D_r^\gamma)^{1/2} \tag{13}$$

where $D_r^\gamma = [D_r - \gamma I_r]_+$. As in the unconstrained case, we work with normalized factors $\overline{F} = \sqrt{T}\overline{F}_z$ and $\overline{\Lambda} = \sqrt{N}\overline{\Lambda}_z$:[14]

$$\overline{F} = \sqrt{T}U_r(D_r^\gamma)^{1/2} \tag{14a}$$

$$\overline{\Lambda} = \sqrt{N}V_r(D_r^\gamma)^{1/2}. \tag{14b}$$

The 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

$$\Delta_{NT}^2 = D_r^\gamma D_r^{-1} = \text{diag}\Big(\frac{(d_1 - \gamma)_+}{d_1}, ..., \frac{(d_r - \gamma)_+}{d_r} \Big). \tag{15}$$

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

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

[14] These are also the optimal solutions from the following optimization problem

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

Though $(\overline{F}Q, \overline{\Lambda}Q)$ is also a solution for any orthonormal matrix Q, $(\overline{F}, \overline{\Lambda})$ as defined in (14a) and (14b) is the only solution (up to a column sign change) that satisfies $\frac{\overline{F}'\overline{F}}{T} = \frac{\overline{\Lambda}'\overline{\Lambda}}{N} = D_r^\gamma$, assuming that the diagonal elements of D_r^γ are distinct.

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

$$\overline{F}_t - \overline{H}'_{NT} F_t^0 = \Delta_{NT}(\widehat{F}_t - \widehat{H}'_{NT} F_t^0). \tag{16}$$

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,

$$\overline{\Lambda}_i - \overline{G}_{NT}\Lambda_i^0 = \Delta_{NT}[\widehat{\Lambda}_i - \widehat{H}_{NT}^{-1}\Lambda_i^0] \tag{17}$$

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, (16), and (17), we obtain the following result.

Proposition 2 *Let $(\overline{F}, \overline{\Lambda})$ be given in (14a) and (14b) with threshold parameter $\gamma > 0$. Suppose that Assumption A holds and $N, T \to \infty$. Then*

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

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

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

Turning to the common component $C^0 = F^0\Lambda^{0\prime}$, the RPC estimate is $\overline{C} = \overline{F}\,\overline{\Lambda}'$ and the PC estimates is $\widehat{C} = \widehat{F}\widehat{\Lambda}'$. Using Δ_{NT}^2 defined in (15), 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, it follows that the elements of $\overline{C}$ are biased towards zero. Recalling that $||\overline{C}||_F^2 = \sum_{i=1}^N \sum_{t=1}^T \overline{C}_{it}^2 = \text{trace}(\overline{C}'\overline{C})$,

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

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

Though $\overline{C}$ is an asymptotically biased estimator for C^0, its asymptotic mean squared errors Amse may 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

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

Because $\widehat{C}_{it}$ is asymptotically unbiased for C^0_{it}, $\textsc{Amse}(\widehat{C}_{it}) = \textsc{Avar}(\widehat{C}_{it})$. The asymptotic bias and variance of $\overline{C}_{it}$ are, respectively,

$$\begin{aligned} \textsc{Abias}(\overline{C}_{it}) &= (\delta_1 - 1)C^0_{it} \\ \textsc{Avar}(\overline{C}_{it}) &= \delta_1^2 \textsc{Amse}(\widehat{C}_{it}). \end{aligned}$$

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

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

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. 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 eigenvalues in the low rank component and noise corruption that motivated RPCA in machine learning. Here, we find that from a statistical perspective, it is also beneficial to use the regularized principal components analysis when the data are noisy and when the pervasive signals are weak.

It is noteworthy that soft-thresholding of the eigenvalues is distinct from regularization of the eigenvectors 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.[15] In constrast, 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 $j = 1, \dots r$.

4.2 Selection of Factors

The problem defined in (12) 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 large is r, the better is the fit, but variance also increases with the number of factors be estimated. Bai and Ng (2002) suggests to determine the number of factors using criteria that take into account model complexity (hence sampling uncertainty). These take the form

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

The $\widehat{IC_2}$ criterion, often used in empirical work, obtains when

$$g(N,T) = \frac{(N+T)}{NT} \log\left(\frac{NT}{N+T}\right). \tag{18}$$

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

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 (12) yields $\left\|\overline{F}\ \overline{\Lambda}'\right\|_F^2 = \|D_r^\gamma\|_F^2$. 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:

$$\overline{r} = \min_{k=0,\ldots,\text{rmax}} \log\left(1 - \sum_{j=1}^{k} (d_j - \gamma)_+^2\right) + kg(N,T). \tag{19}$$

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 additional factor, hence 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 eigenvalues. 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 j-th 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. While $\gamma = 0.05$ is a small penalty to d_j, $\sum_{j=1}^{k}(2d_j - \gamma)$ makes a non-trivial difference to the determination of r as we will see in simulations as we will see below.

4.3 Practical Considerations

Consider the infeasible linear regression $y_{t+h} = \alpha' F_t + \beta' W_t + \epsilon_{t+h}$ where W_t are observed predictors. The regression is infeasible because F is latent. Bai and Ng (2006) shows 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 $\overline{F}$? 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 eigenvalues so that $\overline{F}$ and $\widehat{F}$ have the same rank. Then $\widetilde{F}$, $\widehat{F}$ and $\overline{F}$ are all spanned by U_r. In other words, they are perfectly correlated. 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 $\overline{F}$ remains identical to the fit provided by the first r^* columns of $\widehat{F}$ by virtual of the fact the omitted factors are orthogonal to the included ones.

The upshot of the above discussion is that use of the rank restricted estimate $\overline{F}$ will make no difference to factor augmented regressions unless an effort is made to shrink the coefficients associated with the factors towards zero. This can be achieved by replacing least squares with ridge regressions. As an example, suppose that W_t is empty and we regress y on $\overline{F}$ only. Let $\overline{\alpha}_{OLS}$ and $\overline{\alpha}_R$ be the OLS and ridge estimator for α respectively. For given κ and $\kappa_T = \frac{\kappa}{T}$,

$$\begin{aligned}
\overline{\alpha}_{ols} &= (\overline{F}'\overline{F})^{-1}\overline{F}y = (D_r^\gamma)^{-1/2}U'y/\sqrt{T} \\
\overline{\alpha}_R &= (\overline{F}'\overline{F} + \kappa I_r)^{-1}\overline{F}y \\
&= (TD_r^\gamma + \kappa I_r)^{-1}(D_r^\gamma)^{1/2}\sqrt{T}U'y = (D_r^\gamma + \kappa_T I_r)^{-1}(D_r^\gamma)^{1/2}U_r y/\sqrt{T} \\
&= (D_r^\gamma + \kappa_T I_r)^{-1}D_r^\gamma\overline{\alpha}_{OLS} = (I_r + \kappa_T(D_r^\gamma))^{-1}\,\overline{\alpha}_{OLS} \\
&\approx (I_r - \kappa_T D_r^\gamma)\overline{\alpha}_{OLS}.
\end{aligned}$$

While SVT implements the rank constraint on $\overline{F}$ via D_r^γ, ridge regression shrinks $\widehat{\alpha}_{OLS}$ towards zero so that the contribution of $\overline{\alpha}_R'\overline{F}_t$ is smaller than $\overline{\alpha}_{OLS}'\overline{F}_t$. Note that $\overline{\alpha}_R$ can be constructed from the least squares estimator, D_r^γ, and κ. Explicit ridge estimation is actually not necessary.

The second practical issue concerns construction of the factors. Whether we are interested in the APC, PC, or RPC estimates, the eigenvectors U_r required. Many numerical methods are available to compute eigenvectors when Z is large in dimension. To compute the first eigenvector, the method of *power iteration* starts from an initial vector that has a non-zero component in the direction of the target eigenvector. It then recursively updates and re-normalizes the vector till convergence. The resulting vector is the largest eigenvector. The idea can be extended to compute the invariant subspace of all singular vectors. As suggested in Hastie et al. (2015), the algorithm can be modified to construct robust principal components. The following algorithm is from Hastie et al. (2015).[16]

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

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

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 the solutions to the nuclear norm minimization problem stated in (12). In theory, this is all that is needed for construction of $\overline{F} = \sqrt{T}F_z$ and $\overline{\Lambda} = \sqrt{N}\Lambda_z$. But improved estimates of the left and right eigenvectors can be obtained using Step B, which also explicitly thresholds the singular values.[17] The final estimates of the factors and loadings that emerge from from step B 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. 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. The ridge regression perspective is nonetheless useful in highlighting the role that regularization plays in RPC.

5 Linear Constraints

The minimization problem in (12) has a unique solution under the normalization $F'F = \Lambda'\Lambda = D_r$. However, the unique solution may or may not have economic interpretations. This section considers m linear restrictions on Λ of the form

$$R\,\text{vec}(\Lambda) = \phi \tag{20}$$

where R is $m \times Nr$, and ϕ is $m \times 1$. Both R and ϕ are assumed known a priori. Economic theory may imply lower triangularity of the top $r \times r$ sub-matrix of Λ when the data are appropriately ordered. By suitable design of R, the causality restriction can be expressed as $R\,\text{vec}(\Lambda) = \phi$ without ordering the data a priori. Cross-equation restrictions are allowed, such as due to homogeneity of the loadings across individuals or a subgroup of individuals suggested by theory. Other restrictions

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

are considered in Stock and Watson (2016). The Appendix provides an example how to implement the restrictions in MATLAB.

The linear restrictions on the loadings considered here contrasts with sparse principal components (SPC) estimation which either imposes LASSO type penalty on the loadings, or shrinks the individual entries to zero in a data dependent way, as our constraints are known a priori.[18] Note that if we regard the diagonality of $F'F$ and $\Lambda'\Lambda$ as identification restrictions (rather than statistical normalizations), the linear constraints on the loadings (20) constitute over-identifying restrictions with which we can use to test economic hypothesis. However, it is now possible to relax some of the diagonality restrictions, so long as they are replaced by a sufficient number of linear restrictions; identification is then still possible. A theory of identification for high dimensional factor models is given in Bai and Wang (2014).

The constrained factor estimates $(\overline{F}_{\gamma,\tau}, \overline{\Lambda}_{\gamma,\tau})$ are defined as solutions to the penalized problem

$$(\overline{F}_{\gamma,\tau}, \overline{\Lambda}_{\gamma,\tau}) = \min_{F,\Lambda} \frac{1}{2} \left\| Z - F\Lambda' \right\|_F^2 + \frac{\gamma}{2}\left(\|F\|_F^2 + \|\Lambda\|_F^2 \right) + \frac{\tau}{2} \|R\,\mathrm{vec}(\Lambda) - \phi\|_2^2 \tag{21}$$

where γ and τ are regularization parameters. The linear constraints can be imposed with or without the rank constraints. Imposing cross-equation restrictions will generally require iteration till the constraints are satisfied.

The first order condition with respect to F for a given Λ is unaffected by the introduction of the linear constraints on Λ. Hence, the solution

$$\overline{F}_{\gamma,\tau} = Z\Lambda(\Lambda'\Lambda + \gamma I_r)^{-1}, \quad \forall \tau \geq 0 \tag{22}$$

can be obtained from a ridge regression of Z of Λ. To derive the first order condition with respect to Λ, we rewrite the problem in vectorized form:

$$\|Z - F\Lambda'\|_F^2 = \|\mathrm{vec}(Z') - (F \otimes I_N)\mathrm{vec}(\Lambda)\|_2^2, \quad \|\Lambda\|_F^2 = \|\mathrm{vec}(\Lambda)\|_2^2.$$

The first order condition with respect to $\mathrm{vec}(\Lambda)$ is

$$\begin{aligned} 0 &= -(F' \otimes I_N)\Big[\mathrm{vec}(Z') - (F \otimes I_N)\mathrm{vec}(\Lambda)\Big] + \gamma\,\mathrm{vec}(\Lambda) + \tau R'[R\,\mathrm{vec}(\Lambda) - \phi] \\ &= -\mathrm{vec}(Z'F) - \tau R'\phi + (F'F \otimes I_N)\,\mathrm{vec}(\Lambda) + \gamma\,\mathrm{vec}(\Lambda) + \tau R'R\,\mathrm{vec}(\Lambda). \end{aligned}$$

Solving for $\mathrm{vec}(\Lambda)$ and and denoting the solution by $\mathrm{vec}(\overline{\Lambda}_{\gamma,\tau})$, we obtain

$$\begin{aligned} \overline{\Lambda}_{\gamma,\tau} &= \Big((F'F \otimes I_N) + \gamma I_{Nr} + \tau R'R\Big)^{-1}\Big[\mathrm{vec}(Z'F) + \tau R'\phi\Big] \\ &= \Big((F'F + \gamma I_r) \otimes I_N + \tau R'R\Big)^{-1}\Big[\mathrm{vec}(Z'F + \tau R'\phi)\Big] \end{aligned} \tag{23}$$

[18] For SPC, see Jolliffee et al. (2003), Ma (2013), Shen and Huang (2008), and Zou et al. (2006). The SPC is in turn different from the POET estimator of Fan et al. (2013) which constructs the principal components from a matrix that shrinks the small eigenvalues towards zero.

where the last line follows from the fact that $(F'F \otimes I_N) + \gamma I_{Nr} = (F'F + \gamma I_r) \otimes I_N$. Equations (22) and (23) completely characterize the solution under rank and linear restrictions. In general, the solution will need to be solved by iterating the two equations until convergence. A reasonable starting value is $(\overline{F}, \overline{\Lambda})$, the solution satisfying the rank constraint and before the linear restrictions are imposed. However, while $\overline{F}'\overline{F} = \overline{\Lambda}'\overline{\Lambda} = D_r^\gamma$ and D_r^γ is diagonal, $\overline{F}'_{\gamma,\tau}\overline{F}_{\gamma,\tau}$ and $\overline{\Lambda}'_{\gamma,\tau}\overline{\Lambda}_{\gamma,\tau}$ will not, in general, be diagonal when linear restrictions are present.

These constraint will not bind unless $\tau = \infty$, and we denote by $\Lambda_{\gamma,\infty}$ the binding solution. Observe that in the absence of linear constraints (i.e. $\tau = 0$),

$$\text{vec}(\overline{\Lambda}_{\gamma,0}) = \Big((F'F + \gamma I_r) \otimes I_N\Big)^{-1} \text{vec}(Z'F) \tag{24}$$

which is a ridge estimator. Furthermore, (22) and (24) are nothing but the RPCA estimates. The following is an estimator that satisfies both the rank constraint and $R\,\text{vec}(\Lambda) = \phi$.

Proposition 3 *For given F, let $\overline{\Lambda}_{\gamma,\infty}$ be the solution to (21) with $\tau = \infty$. Also let $\overline{\Lambda}_{\gamma,0}$ be the solution with $\tau = 0$. Then two solutions are related as follows:*

$$\text{vec}(\overline{\Lambda}_{\gamma,\infty}) = \text{vec}(\overline{\Lambda}_{\gamma,0}) - [(F'F+\gamma I_r)^{-1} \otimes I_N]R' \cdot \Big[R[(F'F+\gamma I_r)^{-1} \otimes I_N]R'\Big]^{-1} \Big(R\,\text{vec}(\overline{\Lambda}_{\gamma,0}) - \phi\Big) \tag{25}$$

Proposition 3 says that for given F, a restricted estimate of Λ that satisfies both the rank and linear restrictions can be obtained by imposing the linear restrictions on RPCA solution of Λ that only imposes rank restrictions. It is easy to verify $\overline{\Lambda}_{\gamma,\infty}$ satisfies restriction (20). Once the restricted estimates are obtained, F needs to be re-estimated based on (22). The final solution is obtained by iterating (22) and (25). We note again that $\overline{F}'_{\gamma,\infty}\overline{F}_{\gamma,\infty}$ and $\overline{\Lambda}'_{\gamma,\infty}\overline{\Lambda}_{\gamma,\infty}$ will not, in general, be diagonal matrices in the presence of linear restrictions.

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

$$(\overline{F}_{\gamma_1,\gamma_2,\tau}, \overline{\Lambda}_{\gamma_1,\gamma_2,\tau}) = \text{argmin}_{F,\Lambda}\bigg(\frac{1}{2}||Z - F\Lambda'||_F^2 + \frac{\gamma_1}{2}||F||_F^2 + \frac{\gamma_2}{2}||\Lambda||_F^2 + \frac{\tau}{2}||R\text{vec}(\Lambda) - \phi)||_2^2\bigg).$$

Let

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

Relaxing the constraint that $\gamma_1 = \gamma_2$ with $\tau = 0$ yields the general solution

$$\overline{F}_{\gamma_1,\gamma_2,\tau=0} = \Big(\frac{\gamma_2}{\gamma_1}\Big)^{1/4} U_r(\overline{D}_r^\gamma)^{1/2} \tag{26a}$$

$$\overline{\Lambda}_{\gamma_1,\gamma_2,\tau=0} = \Big(\frac{\gamma_1}{\gamma_2}\Big)^{1/4} (\overline{D}_r^\gamma)^{1/2}. \tag{26b}$$

The corresponding common component is

$$\overline{C}_{\gamma_1,\gamma_2,0} = U_r\overline{D}_r^\gamma V_r'$$

When $\gamma_1 = \gamma_2 = \gamma$, the solution coincides with (13), which can be computed by Algorithm RPC. Once the rank constrained solutions are obtained, Proposition 3 can be used to impose linear constraints ($\tau = \infty$). For other values of τ, optimal solutions can be obtained via iterated ridge regressions, for which γ is replaced by γ_1 in equation (22), and by γ_2 in equation (23).

6 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 simulatios, 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 r^* to be the number of factors that contributes at least a fraction $c(s) = \frac{\text{var}(S)}{\text{var}(X)}$ of the variance of the common component of X. Note that r^* is also the rank of D_r^γ. With $c(S) = 0.05$, DGP 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 r-th (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 the smallest factor estimated on all r^* eigenvectors of the true common component. If the factor space is precisely estimated, the R^2 of this regression should be close to one. These are denoted $\widehat{R}^2$ and $\overline{R}^2$ for the PC and RPC estimates of F, respectively.

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 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, $\overline{r}$ is more robust and correctly selects five factors in many cases.

Table 2 shows results for 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, $\overline{r}$ 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 estimate find an additional factor compared to the top panel. In contrast, $\overline{r}$ 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 and parsimony, it is up to the user whether to use $\widehat{r}$ or $\overline{r}$. In applications, researchers tend to focus on the dominant ones. Our $\overline{IC}_2$ criterion provides a way to determine the minimum number of factors that should be used.

We also estimate the factors using 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 unbalance panel has $N = 128$ variables with $T = 676$ observations. We also consider a balanced panel with $N = 92$ series.[19]

The squared-eigenvalue values can be interpreted as the percent contribution of the factor to the variation of Z. As noted above, regularization shrinks the eigenvalues and hence the length of the factors and the loadings towards zero. Hence the unregularized eigenvalues $\overline{d}_i^2$ are always smaller than the unregularized ones by roughly $\gamma^2 = (0.05)^2$. The original $\widehat{IC}_2$ finds eight factors for the balanced panel. After regularization, $\overline{IC}_2$ finds three factors. In Gorodnichenko and Ng (2017), this difference between $\widehat{r}$ and $\overline{r}$ 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.

Eigenvalues of FredMD data

[19] One idea is to use matrix completion algorithms to fill the missing values. But the data are not 'missing at random', and we leave it for future work.

F	Balanced Panel		Non-Balanced Panel	
	$\widehat{d}_1^2$	$\overline{d}_1^2$	$\widehat{d}_1^2$	$\overline{d}_1^2$
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

7 Conclusion

This paper considers estimation of approximate factor models by regularized principal components with focus on two problems. The first problem is rank regularization with RPCA as output. This is useful when the idiosyncratic errors have large eigenvalues such as due to extreme outliers, or when some factors have small eigenvalues, 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. The second problem is linear restrictions such as motivated by economic theory. We show that the solution is a transformation of the unrestricted estimates. Our analysis provides a statistical view of matrix recovery algorithms and complements results in the machine learning literature.

Table 1: DGP 1

parameters				signal		noise	mean		prob. $= r$		prob. $= r^*$		spanning	
N	T	r^*	ω	C^r	C_r	$c(S)$	$\widehat{r}$	$\overline{r}$	$\widehat{r}$	$\overline{r}$	$\widehat{r}$	$\overline{r}$	$\widehat{R}^2$	$\overline{R}^2$
100	100	5	5.00	0.83	0.12	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	100	5	10.00	0.83	0.12	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	100	5	20.00	0.83	0.12	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	200	5	5.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	200	5	10.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	200	5	20.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	400	5	5.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	400	5	10.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
100	400	5	20.00	0.83	0.13	0.00	5.00	5.00	1.00	1.00	1.00	1.00	0.98	0.98
50	100	5	5.00	0.83	0.10	0.00	5.00	4.95	1.00	0.95	0.00	0.05	0.95	0.95
50	100	5	10.00	0.83	0.10	0.00	5.00	4.95	1.00	0.95	0.00	0.05	0.95	0.95
50	100	5	20.00	0.83	0.10	0.00	5.00	4.95	1.00	0.95	0.00	0.05	0.95	0.95
50	200	5	5.00	0.83	0.11	0.00	5.02	5.00	0.98	1.00	0.00	0.00	0.93	0.96
50	200	5	10.00	0.83	0.11	0.00	5.02	5.00	0.98	1.00	0.00	0.00	0.93	0.96
50	200	5	20.00	0.83	0.11	0.00	5.02	5.00	0.98	1.00	0.00	0.00	0.93	0.96
50	400	5	5.00	0.83	0.11	0.00	5.05	5.00	0.95	1.00	0.95	1.00	0.91	0.96
50	400	5	10.00	0.83	0.11	0.00	5.05	5.00	0.95	1.00	0.95	1.00	0.91	0.96
50	400	5	20.00	0.83	0.11	0.00	5.05	5.00	0.95	1.00	0.95	1.00	0.91	0.96
100	100	5	5.00	0.81	0.12	0.02	5.36	5.00	0.64	1.00	0.64	1.00	0.63	0.98
100	100	5	10.00	0.78	0.12	0.06	5.79	5.00	0.28	1.00	0.28	1.00	0.28	0.98
100	100	5	20.00	0.69	0.12	0.17	6.81	5.00	0.00	1.00	0.00	1.00	0.01	0.97
100	200	5	5.00	0.81	0.13	0.02	5.67	5.00	0.33	1.00	0.33	1.00	0.32	0.98
100	200	5	10.00	0.78	0.13	0.06	5.91	5.00	0.19	1.00	0.19	1.00	0.19	0.98
100	200	5	20.00	0.69	0.13	0.17	7.13	5.00	0.00	1.00	0.00	1.00	0.00	0.98
100	400	5	5.00	0.81	0.13	0.02	5.88	5.00	0.12	1.00	0.12	1.00	0.12	0.98
100	400	5	10.00	0.78	0.13	0.06	5.90	5.00	0.17	1.00	0.17	1.00	0.16	0.98
100	400	5	20.00	0.69	0.13	0.18	7.15	5.00	0.00	1.00	0.00	1.00	0.00	0.98
50	100	5	5.00	0.81	0.10	0.02	5.32	4.92	0.68	0.92	0.00	0.08	0.65	0.95
50	100	5	10.00	0.78	0.10	0.06	5.69	4.89	0.36	0.90	0.00	0.10	0.35	0.95
50	100	5	20.00	0.69	0.10	0.17	6.39	4.83	0.02	0.83	0.00	0.17	0.04	0.94
50	200	5	5.00	0.81	0.11	0.02	5.42	4.99	0.59	0.99	0.00	0.01	0.57	0.95
50	200	5	10.00	0.78	0.11	0.06	5.71	4.99	0.36	0.99	0.00	0.01	0.35	0.95
50	200	5	20.00	0.69	0.11	0.17	6.58	4.98	0.03	0.98	0.00	0.02	0.04	0.94
50	400	5	5.00	0.81	0.11	0.02	5.54	5.00	0.49	1.00	0.00	0.00	0.47	0.95
50	400	5	10.00	0.78	0.11	0.06	5.71	5.00	0.36	1.00	0.00	0.00	0.35	0.95
50	400	5	20.00	0.69	0.11	0.17	6.66	5.00	0.03	1.00	0.00	0.00	0.04	0.94

Notes: $X_{it} = F_t^{0\prime}\Lambda_i^0 + e_{it} + s_{it}$, $e_{it} \sim (0,1)$, $s_{it} \sim (0,\omega^2)$. F_t 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 $\gamma = 0.05$, $C^r = \frac{\text{var}(C^0)}{\text{var}(X)}$, $C_r = \frac{\text{var}(F_r\Lambda_r')}{\text{var}(X)}$, $c(S) = \frac{\text{var}(S)}{\text{var}(X)}$. The column labeled 'spanning' is the R^2 from a regression of the smallest factor on U_r.

Table 2: DGP 2

parameters				signal		noise	mean		prob. $= r$		prob. $= r^*$		spanning	
N	T	r^*	ω	C^r	C_r	$c(S)$	$\widehat{r}$	$\overline{r}$	$\widehat{r}$	$\overline{r}$	$\widehat{r}$	$\overline{r}$	$\widehat{R}^2$	$\overline{R}^2$
100	100	3	1.00	0.67	0.02	0.00	3.94	3.00	0.00	0.00	0.06	1.00	0.07	0.95
100	100	3	0.75	0.67	0.01	0.00	3.95	3.00	0.00	0.00	0.05	1.00	0.05	0.95
100	100	3	0.50	0.67	0.01	0.00	3.97	3.00	0.00	0.00	0.03	1.00	0.04	0.95
100	200	3	1.00	0.67	0.02	0.00	4.01	3.00	0.01	0.00	0.00	1.00	0.00	0.95
100	200	3	0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.00	1.00	0.00	0.95
100	200	3	0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.00	1.00	0.00	0.95
100	400	3	1.00	0.67	0.02	0.00	4.26	3.00	0.26	0.00	0.00	1.00	0.00	0.95
100	400	3	0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.00	1.00	0.00	0.95
100	400	3	0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.00	1.00	0.00	0.95
50	100	3	1.00	0.67	0.02	0.00	3.55	2.57	0.00	0.00	0.45	0.57	0.41	0.93
50	100	3	0.75	0.67	0.01	0.00	3.60	2.62	0.00	0.00	0.40	0.62	0.37	0.93
50	100	3	0.50	0.67	0.01	0.00	3.64	2.66	0.00	0.00	0.36	0.66	0.33	0.93
50	200	3	1.00	0.67	0.02	0.00	3.95	2.97	0.00	0.00	0.06	0.97	0.06	0.91
50	200	3	0.75	0.67	0.01	0.00	3.96	2.98	0.00	0.00	0.04	0.98	0.04	0.91
50	200	3	0.50	0.67	0.01	0.00	3.97	2.98	0.00	0.00	0.03	0.98	0.04	0.91
50	400	3	1.00	0.67	0.02	0.00	4.00	3.00	0.00	0.00	0.01	1.00	0.01	0.91
50	400	3	0.75	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.01	1.00	0.01	0.91
50	400	3	0.50	0.67	0.01	0.00	4.00	3.00	0.00	0.00	0.00	1.00	0.01	0.91
100	100	3	1.00	0.60	0.02	0.11	4.81	2.93	0.81	0.00	0.00	0.93	0.01	0.93
100	100	3	0.75	0.59	0.01	0.11	4.84	2.95	0.84	0.00	0.00	0.95	0.01	0.93
100	100	3	0.50	0.59	0.01	0.11	4.86	2.96	0.86	0.00	0.00	0.96	0.01	0.93
100	200	3	1.00	0.60	0.02	0.11	5.01	3.00	0.99	0.00	0.00	1.00	0.01	0.93
100	200	3	0.75	0.59	0.01	0.11	5.00	3.01	1.00	0.00	0.00	0.99	0.01	0.93
100	200	3	0.50	0.59	0.01	0.11	5.00	3.01	1.00	0.00	0.00	0.99	0.01	0.93
100	400	3	1.00	0.60	0.02	0.11	5.21	3.10	0.80	0.00	0.00	0.90	0.00	0.84
100	400	3	0.75	0.59	0.01	0.11	5.00	3.12	1.00	0.00	0.00	0.88	0.00	0.83
100	400	3	0.50	0.59	0.01	0.11	5.00	3.13	1.00	0.00	0.00	0.87	0.00	0.82
50	100	3	1.00	0.60	0.02	0.11	4.18	2.27	0.34	0.00	0.16	0.27	0.16	0.94
50	100	3	0.75	0.59	0.01	0.11	4.23	2.31	0.38	0.00	0.15	0.31	0.15	0.93
50	100	3	0.50	0.59	0.01	0.11	4.28	2.34	0.41	0.00	0.13	0.34	0.14	0.93
50	200	3	1.00	0.60	0.02	0.11	4.82	2.83	0.82	0.00	0.01	0.83	0.02	0.89
50	200	3	0.75	0.59	0.01	0.11	4.84	2.86	0.84	0.00	0.01	0.86	0.02	0.89
50	200	3	0.50	0.59	0.01	0.11	4.86	2.87	0.86	0.00	0.01	0.87	0.02	0.89
50	400	3	1.00	0.60	0.02	0.11	4.97	2.97	0.96	0.00	0.00	0.97	0.01	0.89
50	400	3	0.75	0.59	0.01	0.11	4.96	2.98	0.97	0.00	0.00	0.98	0.01	0.89
50	400	3	0.50	0.59	0.01	0.11	4.97	2.98	0.97	0.00	0.00	0.98	0.01	0.89

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 $\gamma = 0.05$, $C^r = \frac{\text{var}(C^0)}{\text{var}(X)}$, $C_r = \frac{\text{var}(F_r \Lambda_r')}{\text{var}(X)}$, $c(S) = \frac{\text{var}(S)}{\text{var}(X)}$. The column labeled 'spanning' is the R^2 from a regression of the smallest factor on U_r.

Appendix A

Linear Restrictions in matlab

The factor model is $X_{it} = F_{1t}\Lambda_{i1} + F_{2t}\Lambda_{i2} + F_{3t}\Lambda_{i3} + e_{it}$. Consider the restrictions (i) (i) $\Lambda_{12} = 0$, (ii)) $\Lambda_{13} = 0$, (iii) (v) $\Lambda_{21} = \Lambda_{31}$. Given unrestricted estimates of the factor loadngs LHATU, the following returns the restricted loadings LHATRV.

```
R.cmat{1}=zeros(N,r); R.cmat{1}(1,2)=1.0; R.phi{1}=0.0;
R.cmat{2}=zeros(N,r); R.cmat{2}(1,3)=1.0; R.phi{2}=0.0;
R.cmat{3}=zeros(N,r); R.cmat{3}(2,1)=1.0; R.cmat{3}(3,1)=-1;R.phi{3}=0.0;

Rvec=[];
phivec=[];
for j=1:length(R.cmat);
Rvec=[Rvec vec(R.cmat{j})];
phivec=[phivec; R.phi{j}];
end;
Rvec=Rvec';
dum1=kron(inv(Fhat'*Fhat),eye(N));
dum2= Rvec*dum1*Rvec';
adj=dum1*Rvec'*inv(dum2)*(Rvec*vec(Lhatu)-phivec);
Lhatrv=vec(Lhatu)-adj;

% one more iteration
L_u=Lhatu;
F_u=Fhat;
it=1; maxit=100; done=0;
while done==0 & it<=maxit;
    dum1=kron(inv(F_u'*F_u),eye(N));
    dum2= Rvec*dum1*Rvec';
    adj=dum1*Rvec'*inv(dum2)*(Rvec*vec(L_u)-phivec);
    Lamrv=vec(L_u)-adj;
    L_r=reshape(Lamrv,N,r);
    F_r=X*L_r*inv(L_r'*L_r);
    err1=norm(F_r'*F_r-F_u'*F_u,'fro');
    err2=norm(L_r'*L_r-L_u'*L_u,'fro');
    if err1+err2> 1e-8;
        F_u=F_r;
        L_u=L_r;
        disp(sprintf('%d %f %f ',it,err1,err2));
        it=it+1;
    else;
       disp(sprintf('Converged: %d %f %f ',it,err1,err2));
        done=1;
    end;
end;
disp('Converged ');
disp('estimates: unrestricted            restricted');
disp([Lhatu Lhatr]);
disp(['Restricted LL      FF']);
disp([L_r'*L_r F_r'*F_r]);
for i=1:r;
    c1=corr(Fhat(:,i),F_r(:,i));
    c2=corr(Lhatu(:,i),L_r(:,i));
    mymprint([c1 c2  ]);
end;
```

$$\begin{pmatrix} 4.70 & -1.13 & 0.89 \\ 1.21 & 0.77 & 2.41 \\ -3.67 & 0.05 & 1.73 \\ -1.27 & -3.71 & 0.45 \\ 0.16 & -0.81 & -0.93 \end{pmatrix} \quad \begin{pmatrix} 4.70 & 0.00 & 0.00 \\ -1.23 & 0.77 & 2.41 \\ -1.23 & 0.05 & 1.73 \\ -1.27 & -3.71 & 0.45 \\ 0.16 & -0.81 & -0.93 \end{pmatrix}$$

Appendix B

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

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

Take transpose and inverse, we have

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

Substitute 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 multiply the equation $X = F^0\Lambda^{0\prime} + e$ by $F^{0\prime}$ and right multiply it by $\widetilde{\Lambda}$, divide 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

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

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 left 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 for the second claim is similar by using $\widehat{F} = \widetilde{F}D_r^{1/2}$.

Proof of (17).

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

Moving the second term to the left hand side, we obtain (17).

Proof of Proposition 3. The estimator by direction calculations is given by

$$\mathrm{vec}(\overline{\Lambda}_{\gamma,\infty}) = \mathrm{vec}(\overline{\Lambda}_{\gamma,0}) - \Delta(\gamma,\phi)$$

where $\Delta(\gamma,\phi) = ((F^+\otimes I_N)'(F^+\otimes I_N))^{-1}R'\Big[R\Big((F^+\otimes I_N)'(F^+\otimes I_N)\Big)^{-1}R'\Big]^{-1}\Big(R\,\mathrm{vec}(\overline{\Lambda}_{\gamma,0}) - \phi\Big)$.
But

$$\begin{aligned}
\Big((F^+\otimes I_N)'(F^+\otimes I_N)\Big)^{-1}R' &= \Big((F'F+\gamma I_r)^{-1}\otimes I_N\Big)R' \\
R\Big[(F^+\otimes I_N)'(F^+\otimes I_N)\Big]^{-1}R' &= R\Big[(F'F+\gamma I_r)^{-1}\otimes I_N\Big]R'.
\end{aligned}$$

Hence

$$\Delta(\gamma,\phi) = [(F'F+\gamma I_r)^{-1}\otimes I_N]R'\cdot\Big[R[(F'F+\gamma I_r)^{-1}\otimes I_N]R'\Big]^{-1}\Big(R\,\mathrm{vec}(\overline{\Lambda}_{\gamma,0}) - \phi\Big).$$

Proof of (26a) and (26b). With $\tau = 0$, the objective function becomes

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

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 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, we know 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 (26a) and (26b).

References

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

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

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

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models, *Econometrica* **70:1**: 191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference with factor-augmented regressions, *Econometrica* **74:4**: 1133–1150.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis, *Foundations and Trends in Econometrics* **3:2**: 89–163.

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

Bai, J. and Wang, P. (2014). Identification theory for high dimensional static and dynamic factor models, *Journal of Econometrics* **178**(2): 794–804.

Bentler, P. (1972). A lower bound method for the dimension-free measurement of internal consistency, *Social Science Research* **1**: 343–357.

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

Bertsimas, D., Copenhaver, M. and Mazumder, R. (2016). Certifiably optimal low rank factor analysis. arXiv:1604.06837v1.

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis, *Journal of Econometrics* **132**: 169–194.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine Learning* **3**(1): 1–122.

Cai, J., Candes, E. and Shen, Z. (2008). A singular value thresholding algorithm for matrix completion, *Siam Journal on Optimization* **40**(4): 1956–1982.

Candes, E. J. and Recht, B. (2011). Exact matrix completion via convex optimization, *Journal of the ACM* **58**: 11–37.

Candes, E., Li, X., Ma, Y. and Wright, J. (2011). Robust principal compoennt analysis, *Journal of the ACM* **58**(3): Article 11.

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

Connor, G. and Korajzcyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis, *Journal of Financial Economics* **15**: 373–394.

Delvin, S., Gnanadesikan, R. and Kettering, J. (1981). Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association* **354-362**.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements, *Journal of Royal Statistical Society, Series B* **75**(4): 603–680.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation, *Review of Economics and Statistics* **82:4**: 540–554.

Gorodnichenko, Y. and Ng, S. (2017). Level and volatility factors in macroeconomic data, *Journal of Monetary Economics, forthcoming* .

Grant, M. and Boyd, S. (2015). The cvx users' guide.
URL: *http://cvxr.com/cvx/doc/CVX.pdf*

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

Hastie, T., Mazumder, R., Lee, J. and Zadeh, R. (2015). Matrix completion and low rank svd via fast alternating least squares, *Journal of Machine Learning Research* **16**: 3367–3402.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

Hubert, M. and Rousseeuw, P. (2005). Robpca: A new approach to robust principal analysis, *Journal of American Statistical Association* **47**(64-79).

Hubert, M., Rousseeuw, P. and 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 edn, Springer Series in Statistics, New York.

Jolliffee, I., Trendafilov, T. and Uddin, M. (2003). A modified principal component technique based on the lasso, *Journal of Computational and Graphical Statistics* **12**(531-547).

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

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

Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal componets: Primary theory and monte carlo, *Journal of American Statistical Association* **80**(391): 759–766.

Lin, Z., Chen, M. and Ma, Y. (2013). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055v3.

Ma, Y. and Genton, M. (2001). Highly robust esetimation of dispersion matrices, *Journal of Multivariate Analysis* **78**: 11–36.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding, *Annals of Statistics* **41**(2): 772–801.

McCracken, M. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research, *Journal of Business and Economic Statistcs* **Forthcoming**.

Negahban, S. and Wainwright, M. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise, *Journal of Machine Learning Research* **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. and Srebro, J. (2005). Fast maximum margin matrix factorization for collaborative prediction, *Proceedings of the 22nd International Conference on Machine Learning* .

Saunderson, J., Chandrasekaran, V., Parrilo, P. and Willsky, S. (2012). Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting, *Siam Journal of Matrix Analysis and Applications* **38**(4): 1395–1415. arXiv: 1204.1220v1.

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

Shapiro, A. and 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. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximations, *Journal of Multivariate Analysis* **99**: 1015–1034.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association* **97**: 1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics* **20:2**: 147–162.

Stock, J. and Watson, M. (2016). Factor models and structural vector autoregressions in macroeconomics, *in* J. B. Taylor and H. Uhlig (eds), *Handbook of Macroeconomics*, Vol. 2A, pp. 415–526.

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

Udell, M., Horn, C., Zadeh, R. and Boyd, S. (2016). Generalized low rank models, *Foundations and Trends in Machine Learning* **9**(1): 1–118. arXiv:1410.0342v4.

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

Yang, D., Ma, Z. and Buja, A. (2014). A sparse singular value decomposition method for high dimensional data, *Journal of Computational and Graphical Statistics* **23**(4): 923–942.

Zhou, Z., Li, X., Wright, J., Candes, E. and Ma, Y. (2010). Stable principal component pursuit, *International Symposium on Information Theory (ISIT)*, pp. 1518–1522.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal components, *Journal of Computational and Graphical Statistics* **15**(2): 265–286.