Journal of Econometrics 191 (2016) 1–18

Contents lists available at ScienceDirect

Journal of Econometrics

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

Efficient estimation of approximate factor models via penalized maximum likelihood

Jushan Bai[a,*,1], Yuan Liao[b]

[a] Department of Economics, Columbia University, 420 West 118 Street, New York, NY 10027, USA

[b] Department of Mathematics, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 27 January 2014
Received in revised form
19 August 2015
Accepted 10 October 2015
Available online 20 October 2015

Keywords:
High dimensionality
Unknown factors
Principal components
Conditional sparse
Thresholding
Cross-sectional correlation
Penalized maximum likelihood
Adaptive LASSO
SCAD
Heteroskedasticity

ABSTRACT

We study an approximate factor model in the presence of both cross sectional dependence and heteroskedasticity. For efficient estimations it is essential to estimate a large error covariance matrix. We estimate the common factors and factor loadings based on maximizing a Gaussian quasi-likelihood, through penalizing a large covariance sparse matrix. The weighted ℓ_1 penalization is employed. While the principal components (PC) based methods estimate the covariance matrices and individual factors and loadings separately, they require consistent estimation of residual terms. In contrast, the penalized maximum likelihood method (PML) estimates the factor loading parameters and the error covariance matrix jointly. In the numerical studies, we compare PML with the regular PC method, the generalized PC method (Choi 2012) combined with the thresholded covariance matrix estimator (Fan et al. 2013), as well as several related methods, on their estimation and forecast performances. Our numerical studies show that the proposed method performs well in the presence of cross-sectional dependence and heteroskedasticity.

1. Introduction

In many applications of economics, finance, and other scientific fields, researchers often face a large panel dataset in which there are multiple observations for each individual; here individuals can be families, firms, countries, etc. One useful method for summarizing information in a large dataset is the factor model:

$$y_{it} = \alpha_i + \lambda_{0i}' f_t + u_{it}, \quad i \leq N, t \leq T, \tag{1.1}$$

where α_i is an individual effect, λ_{0i} is an $r \times 1$ vector of factor loadings and f_t is an $r \times 1$ vector of common factors; u_{it} denotes the idiosyncratic component of the model. Note that y_{it} is the only observable random variable in this model. If we write $y_t = (y_{1t}, \ldots, y_{Nt})'$, $\Lambda_0 = (\lambda_{01}, \ldots, \lambda_{0N})'$, $\alpha = (\alpha_1, \ldots, \alpha_N)'$ and $u_t = (u_{1t}, \ldots, u_{Nt})'$, then model (1.1) can be equivalently written as

$$y_t = \alpha + \Lambda_0 f_t + u_t.$$

An efficient estimation of the factor loadings and factors should take into account both cross-sectional dependence and heteroskedasticity. This paper uses the penalized maximum (quasi) likelihood estimation under large N, T. The maximum likelihood estimator depends on estimating a high-dimensional covariance matrix $\Sigma_{u0} = \text{cov}(u_t)$, which is a difficult problem when it is non-diagonal and $N/T \to \infty$. Recently, Bai and Li (2012a) studied the maximum likelihood estimation when Σ_{u0} is a diagonal matrix. As was shown by Chamberlain and Rothschild (1983), it is desirable to allow dependence among the error terms $\{u_{it}\}_{i \leq N, t \leq T}$ not only serially but also cross-sectionally. This gives rise to the *approximate factor model*. With approximate factor models, Doz et al. (2012) considered the consistency of MLE for f_t, restricting a diagonal error covariance matrix. Bai and Li (2012b) estimated an approximate factor model for both factors and factor loadings with MLE, also restricting a diagonal error covariance matrix, and derived the limiting distributions of the estimators. These are shrinkage estimators that shrink the off diagonal elements of Σ_{u0} to zero.

In addition to the diagonal elements, this paper also estimates the off-diagonal elements of Σ_{u0}, which has $O(N^2)$ number of parameters. The key assumption we make is that the model is *conditionally sparse*, in the sense that Σ_{u0} is a sparse matrix with

[*] Corresponding author.
E-mail addresses: jb3064@columbia.edu (J. Bai), yuanliao@umd.edu (Y. Liao).

[1] Also at: School of Finance, Nankai University, China.

http://dx.doi.org/10.1016/j.jeconom.2015.10.003
0304-4076/© 2015 Published by Elsevier B.V.

bounded eigenvalues. This assumption requires many off-diagonal elements of Σ_{u0} to be zero or nearly so, but still allows the identities of the sparse positions to be unknown. The conditional sparsity, though slightly stronger than the assumptions in Chamberlain and Rothschild (1983), is meaningful in practice. For example, when the idiosyncratic components represent firms' individual shocks, they are either uncorrelated or weakly correlated among the firms across different industries, because the industry specific components are not necessarily pervasive for the whole economy (Connor and Korajczyk, 1993). Under the sparsity assumption, Fan et al. (2013) proposed a thresholding method to consistently estimate Σ_{u0} when $N > T$. Their method is based on the traditional principal components method, and does not improve the estimation of factors and loadings. This paper proposes a maximum likelihood (ML)-based method that simultaneously estimate the error covariance matrix and loadings, taking into account both cross-sectional correlations and heteroskedasticity.

Let $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$, and $S_y = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})(y_t - \bar{y})'$ be the sample covariance matrix based on the observed data. The quasi-likelihood function is

$$L(\Lambda, \Sigma_u, S_f) = \frac{1}{N}\log|\Lambda S_f \Lambda' + \Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda S_f \Lambda' + \Sigma_u)^{-1}), \tag{1.2}$$

where $S_f = \frac{1}{T}\sum_{t=1}^{T}(f_t - \bar{f})(f_t - \bar{f})'$, with $\bar{f} = \frac{1}{T}\sum_{t=1}^{T} f_t$. In addition, a weighted ℓ_1-penalty is attached to penalize the estimation of off-diagonal entries. So we are solving the following optimization problem:

$$\min_{\Lambda, \Sigma_u, S_f}\left[L(\Lambda, \Sigma_u, S_f) + \sum_{i\neq j}\mu_{N,T} w_{ij}|\Sigma_{u,ij}|\right]$$

where the weight w_{ij} is the entry-dependent weight; $\mu_{N,T}$ is a tuning parameter. We provide data-dependent choices for $\{w_{ij}\}_{i,j\leq N}$ and $\mu_{N,T}$, as well as the corresponding theories.

There has been a large literature on estimating model (1.1). Stock and Watson (1998; 2002) and Bai (2003) considered the principal components analysis (PC), which essentially treats u_{it} to have the same variance across i, and is inefficient. Choi (2012) proposed a generalized PC; also see Breitung and Tenhofen (2011). Additional literature on factor models includes, for example, Tsai and Tsay (2010), Bai and Ng (2002), Wang (2009), Dias et al. (2013), Han (2012), among others. Most of these studies are based on the PC method, which is inefficient under cross-sectional heteroskedasticity with unknown dependence structures. Moreover, this paper studies high-dimensional static factor models although the factors and errors can be serially correlated. For generalized dynamic factor models, the readers are referred to Forni et al. (2000; 2005), Forni and Lippi (2001), Hallin and Liška (2007), among others. Our estimation method is maximum likelihood (ML) based, in which no spectral analysis is involved. The ML-based estimation allows for over-identification restrictions to be imposed on the loadings (in a similar way as Bai and Wang (2015)) and allows for forecasting in the spirit of Giannone et al. (2008).

The theoretical results of our paper are only about the consistency of the estimators, although some convergence rate of the covariance estimator is presented in Lemma B.2 in the Appendix, which is not minimax optimal. We admit that due to the technical difficulty, it is challenging to derive the optimal (or near optimal) rate of convergence, and further research on the optimal rate is needed in the future. This paper aims to propose a novel ML-based method for estimating approximate factor models, and illustrates its appealing features to use in practice. We shall elaborate the advantages of ML-based methods in Section 2.2. In addition, we assume the number of factors r to be known. Both N and T diverge to infinity and r is fixed. In practice, r can be estimated from the data, and there has been a large literature addressing its consistent estimation, for example, Bai and Ng (2002), Kapetanios (2010), Onatski (2010), Alessi et al. (2010), Hallin and Liška (2007), and Lam and Yao (2012), among others.

The recent work by Fan et al. (2013) focuses on the covariance estimation using the regular PC. In contrast, we focus on efficiently estimating the factors, loadings, and the covariance matrices simultaneously using penalized MLE. Hence we focus on different estimation problems. The maximum likelihood method has been one of the fundamental tools for statistical estimation and inference.

Our approach is also closely related to the large covariance estimation literature, which has been rapidly growing in recent years. Our penalization procedure is similar to the method in Lam and Fan (2009), Bien and Tibshirani (2011), etc. However, as we described above, our approach is still quite different from theirs in the sense that the penalized ML method considered in this paper estimates the loadings and error covariance matrix jointly. A major difficulty is that the likelihood function being considered contains a few fast-diverging eigenvalues thanks to $\Lambda_0\Lambda_0'$. One of our main objectives is to show that maximizing the Gaussian likelihood function involving fast-diverging eigenvalues can still achieve consistency. Other works on large covariance estimation include, for example, Cai and Zhou (2012), Bickel and Levina (2008), Fan et al. (2008), Jung and Marron (2009), Witten et al. (2009), Deng and Tsui (2013), Yuan (2010), Ledoit and Wolf (2012), El Karoui (2008), Pati et al. (2012), Rohde and Tsybakov (2011), Zhou et al. (2011) and Ravikumar et al. (2011), etc.

The paper is organized as follows. Section 2 defines the simultaneous estimation using penalized MLE, and discusses the advantages of ML-based methods. Section 3 presents theoretical analysis. Section 4 discusses computational issues and implementations. Section 5 numerically compares the proposed methods with competing ones in the literature on both estimation and time series forecasts, using simulated and real data. Finally, Section 6 concludes with further discussions. All proofs are given in the Appendix.

Notation

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of a matrix A respectively. Also let $\|A\|_1$, $\|A\|$ and $\|A\|_F$ denote the ℓ_1, spectral and Frobenius norms of A, respectively. They are defined as $\|A\|_1 = \max_i \sum_j |A_{ij}|$, $\|A\| = \sqrt{\lambda_{\max}(A'A)}$, $\|A\|_F = \sqrt{\mathrm{tr}(A'A)}$. Note that when A is a vector, both $\|A\|$ and $\|A\|_F$ are equal to the Euclidean norm. For two sequences a_T and b_T, we write $a_T \ll b_T$, and equivalently $b_T \gg a_T$, if $a_T = o(b_T)$ as $T \to \infty$. Also, $a_T \asymp b_T$ if $a_T = o(b_T)$ and $b_T = o(a_T)$.

2. Simultaneous estimation based on maximum likelihood

The approximate factor model (1.1) implies the following covariance decomposition:

$$\Sigma_{y0} = \Lambda_0 \,\mathrm{cov}(f_t)\, \Lambda_0' + \Sigma_{u0}, \tag{2.1}$$

assuming f_t to be uncorrelated with u_t, where Σ_{y0} and Σ_{u0} denote the $N \times N$ covariance matrices of y_t and u_t; $\mathrm{cov}(f_t)$ denotes the $r \times r$ covariance of f_t, all assumed to be time-invariant. The approximate factor model typically requires the idiosyncratic covariance Σ_{u0} have bounded eigenvalues and $\Lambda_0'\Lambda_0$ have eigenvalues diverging at rate $O(N)$. One of the key concepts of approximate factor models is that it allows Σ_{u0} to be non-diagonal.

2.1. ℓ_1-penalized maximum likelihood

We jointly estimate (Λ_0, Σ_{u0}), taking into account the cross-sectional dependence and heteroskedasticity simultaneously. Because of the existence of α, the model $y_t = \Lambda_0 f_t + \alpha + u_t$ is observationally equivalent to $y_t = \Lambda_0 f_t^* + \alpha^* + u_t$, where $f_t^* = f_t - \bar{f}$ and $\alpha^* = \alpha + \Lambda_0 \bar{f}$. Therefore without loss of generality, we assume $\bar{f} = 0$. In addition, we focus on a usual restriction for MLE of factor analysis (see e.g., Lawley and Maxwell, 1971) as follows:

$$S_f = I_r, \text{ and } \Lambda'\Sigma_u^{-1}\Lambda \text{ is diagonal,} \tag{2.2}$$

and the diagonal entries of $\Lambda'\Sigma_u^{-1}\Lambda$ are distinct and are arranged in a decreasing order. Motivated by the Gaussian likelihood function, the normalized negative (quasi) log-likelihood is given by

$$L(\Lambda, \Sigma_u) = \frac{1}{N}\log\left|\det\left(\Lambda\Lambda' + \Sigma_u\right)\right| + \frac{1}{N}\mathrm{tr}\left(S_y(\Lambda\Lambda' + \Sigma_u)^{-1}\right). \tag{2.3}$$

We estimate the parameters via the penalized (quasi) MLE:

$$\begin{aligned}(\widehat{\Lambda}, \widehat{\Sigma}_u) &= \arg\min_{(\Lambda,\Sigma_u)\in\Theta_\lambda\times\Gamma} L_1(\Lambda, \Sigma_u) \\ &\equiv \arg\min_{(\Lambda,\Sigma_u)\in\Theta_\lambda\times\Gamma} L(\Lambda, \Sigma_u) + P_T(\Sigma_u)\end{aligned} \tag{2.4}$$

where Θ_λ is a parameter space for the loading matrix, and Γ is the parameter space for Σ_u, to be defined later. Hence minimizing (2.4) estimates the loadings and the error covariance matrix jointly. Here $P_T(\Sigma_u)$ is a penalty function operated on the off-diagonal elements of Σ_u to penalize the inclusion of many off-diagonal elements of $\Sigma_{u,ij}$ in small magnitudes, which therefore produces a sparse estimator $\widehat{\Sigma}_u$. We employ the weighted ℓ_1-penalty

$$P_T(\Sigma_u) = \frac{1}{N}\sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}|.$$

Here $\mu_{N,T}$ is a tuning parameter that converges to zero at a not-too-fast rate; w_{ij} is an entry-dependent weight parameter, which can be either deterministic or stochastic. We suggest three specific choices for w_{ij}, commonly used in the high-dimensional statistical literature:

Lasso The choice $w_{ij} = 1$ for all $i \neq j$ gives the well-known Lasso penalty $N^{-1}\mu_{N,T}\sum_{i\neq j}|\Sigma_{u,ij}|$ studied by Tibshirani (1996). The Lasso penalty puts an equal weight to each element of the idiosyncratic covariance matrix.

Adaptive-Lasso Let $\widehat{\Sigma}^*_{u,ij}$ be a preliminary consistent estimator $\Sigma_{u0,ij}$. Let $w_{ij} = |\widehat{\Sigma}^*_{u,ij}|^{-1}$, then

$$\frac{\mu_{N,T}}{N}\sum_{i\neq j}w_{ij}|\Sigma_{u,ij}| = \frac{\mu_{N,T}}{N}\sum_{i\neq j}|\widehat{\Sigma}^*_{u,ij}|^{-1}|\Sigma_{u,ij}|$$

corresponds to the adaptive-lasso penalty proposed by Zou (2006). Note that the adaptive-lasso puts an entry-adaptive weight on each off-diagonal element of Σ_u, whose reciprocal is proportional to the preliminary estimate. If the true element $\Sigma_{u0,ij}$ is nearly zero, the weight $|\widehat{\Sigma}^*_{u,ij}|^{-1}$ should be quite large, and results in a heavy penalty on that entry. The preliminary estimator $\widehat{\Sigma}^*_{u,ij}$ can be taken, for example, as the PC estimator $\widehat{\Sigma}^{PC}_{u,ij} = T^{-1}\sum_{t=1}^T \widehat{u}^{PC}_{it}\widehat{u}^{PC}_{jt}$, where $\widehat{u}^{PC}_{it}$ is the residual from the PC estimator. It was shown by Bai (2003) that under mild conditions, $\widehat{\Sigma}^{PC}_{u,ij} - \Sigma_{u0,ij} = O_p(N^{-1/2} + (\log N)^{1/2}T^{-1/2})$ uniformly in (i, j).

SCAD Fan and Li (2001) proposed to use, for some $c > 2$ (e.g., $c = 3.7$)

$$\begin{aligned}w_{ij} = {}& \left[I_{(|\widehat{\rho}_{u,ij}|\le\mu_{N,T})} + \frac{(c - |\widehat{\rho}_{u,ij}|/\mu_{N,T})_+}{c-1}I_{(|\widehat{\rho}_{u,ij}|>\mu_{N,T})}\right] \\ & \times(\widehat{\Sigma}^*_{u,ii}\widehat{\Sigma}^*_{u,jj})^{-1/2}.\end{aligned} \tag{2.5}$$

The notation z_+ stands for the positive part of z; z_+ is z if $z > 0$, zero otherwise. Note that $\widehat{\rho}_{ij} = \widehat{\Sigma}^*_{u,ij}/(\widehat{\Sigma}^*_{u,ii}\widehat{\Sigma}^*_{u,jj})^{1/2}$, and $\widehat{\Sigma}^*_{u,ij}$ is still a preliminary consistent estimator, which can be taken as the PC estimator. So essentially this penalty function penalizes the residual correlation matrix, to accommodate the variations of the covariance scales.

After obtaining $\widehat{\Lambda}$ and $\widehat{\Sigma}_u^{-1}$, we estimate f_t via the generalized least squares (GLS):

$$\widehat{f}_t = (\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(y_t - \bar{y}).$$

Note that (2.3) is the exact log-likelihood function when (1) u_t is normal and (2) data are serially independent. We relax both assumptions and particularly allow the data to be serially dependent across t. Hence (2.3) is a quasi-likelihood function. Also, restriction (2.2) guarantees a unique solution to the maximization of the log-likelihood function up to a column sign change for Λ. Therefore we assume the estimator $\widehat{\Lambda}$ and Λ_0 have the same column signs, as part of the identification conditions. Results without assuming that the signs are not correctly estimated can be found in Stock and Watson (2002).

2.2. Comparison with related methods

Suppose Σ_{u0}^{-1} were known, then the first order condition of $\widehat{\Lambda}$ (see Lemma A.5) implies:

$$\widehat{\Lambda}'\Sigma_{u0}^{-1}S_y = (\widehat{\Lambda}'\Sigma_{u0}^{-1}\widehat{\Lambda} + I_r)\widehat{\Lambda}'.$$

Under the identification condition that $\widehat{\Lambda}'\Sigma_{u0}^{-1}\widehat{\Lambda}$ is diagonal with distinct diagonal entries, the columns of $\Sigma_{u0}^{-1/2}\widehat{\Lambda}$ are the eigenvectors of $\Sigma_{u0}^{-1/2}S_y\Sigma_{u0}^{-1/2}$ corresponding to the first r eigenvalues. Then up to a transformation, $\Sigma_{u0}^{-1/2}\widehat{\Lambda}$ is equivalent to the "generalized principal components" method studied by Choi (2012), who assumed Σ_{u0}^{-1} were known and estimated Λ and $\{f_t\}_{t\le T}$ by solving

$$(\tilde{\Lambda}, \{\tilde{f}_t\}) = \arg\min_{\Lambda,\{f_t\}_{t\le T}}\frac{1}{T}\sum_{t=1}^T(y_t - \bar{y} - \Lambda f_t)'\Sigma_{u0}^{-1}(y_t - \bar{y} - \Lambda f_t). \tag{2.6}$$

Indeed, the solution $\tilde{\Lambda}$ to problem (2.6) is such that the columns of $\Sigma_{u0}^{-1/2}\tilde{\Lambda}(\tilde{\Lambda}'\Sigma_{u0}^{-1}\tilde{\Lambda})^{-1/2}$ are also the eigenvectors of $\Sigma_{u0}^{-1/2}S_y\Sigma_{u0}^{-1/2}$ corresponding to the first r eigenvalues. This motivates a question whether there is any conceptual difference between an ML-based estimator and the generalized PC.

The answer is that ML-based and PC-based methods provide different procedures of estimating Σ_{u0}. To better understand the issue, consider a simpler case where it is known that Σ_{u0} is a diagonal matrix with heteroskedastic diagonal entries σ_{ii}^2. Then the ML-based method solves

$$\begin{aligned}\min_{\mathrm{cov}(f_t),\Lambda,\{\sigma_{ii}^2\}} & \log|\det(\Lambda\mathrm{cov}(f_t)\Lambda' + \Sigma_u)| \\ & + \mathrm{tr}(S_y(\Lambda\mathrm{cov}(f_t)\Lambda' + \Sigma_u)^{-1}),\end{aligned} \tag{2.7}$$

subject to normalization constraints on $\mathrm{cov}(f_t)$ and Λ. The generalized PC, on the other hand, becomes:

$$\min_{\Lambda,\{f_t\}_{t\leq T}} \frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y}-\Lambda f_t)'\Sigma_{u0}^{-1}(y_t-\bar{y}-\Lambda f_t).$$

The ML method estimates all parameters (including Σ_{u0}) simultaneously, while PC requires a separate estimation of $\sigma_{ii}^2=Eu_{it}^2$ in a first step. To estimate Eu_{it}^2, PC-based method also requires consistently estimating Λ and $\{f_t\}_{t\leq T}$ and then the residuals $\{u_{it}\}_{i\leq N,t\leq T}$. Indeed Choi (2012) proposed a two-step estimator, which estimates Σ_{u0}^{-1} in the first step, then (2.6) is solved with Σ_{u0}^{-1} replaced by its consistent estimator. Such a procedure is problematic when N is relatively small. In the case when N is fixed, f_t cannot be consistently estimated. As the estimated residuals $\widehat{u}_{it}$ depend on the estimated factors, $\widehat{u}_{it}$ is not consistent for the true error u_{it}. This means that $\widehat{\sigma}_{ii}^2$ based on $\widehat{u}_{it}$ is not consistent for $\sigma_{ii}^2=Eu_{it}^2$. This issue is related to the "incidental parameters bias" of Neyman and Scott (1948). In this case, the PC-based method treats $F=(f_1,\ldots,f_T)$ as high-dimensional parameters, and with a relatively small N, these parameters cannot be estimated well.

In contrast, the likelihood function (2.7) depends on the factors only through a low-dimensional matrix $\mathrm{cov}(f_t)$; the ML-based method estimates Σ_{u0} directly, avoiding estimating the residuals or factors. Even if N is fixed, the MLEs of Λ_0 and Σ_{u0} are still consistent as $T\to\infty$, because this setting falls in the framework of classical inference; see, for example, Lawley and Maxwell (1971). Our numerical studies demonstrate that N does not need to be very small to reveal the advantages of ML-based methods.

Similarly, when T is small but N is large, the ML method can work with the $T\times T$ data matrix (switching the role of N and T), and produce consistent estimation of the factors and the time series covariance matrix. In contrast, the PC-based method is not consistent when one of the dimensions is small.

Even for non-diagonal Σ_{u0}, generalized-PC relies on residuals to estimate the off-diagonal elements. Indeed, Fan et al. (2013) consistently estimate Σ_{u0}^{-1} as $N,T\to\infty$, based on the PC method. In contrast, the proposed penalized MLE directly treats them as parameters, and does not rely on residuals to estimate them.

The penalized MLE can be generalized in several aspects, while the PC-based method will not be suitable under these settings. (i) Confirmatory factor analysis: the ML method can incorporate additional restrictions. For example, when some components of Λ_0 are known, the ML method will not estimate these components, the maximization is taken with respect to the unknown elements. More general restrictions such as cross-equation restrictions are discussed by Bai and Wang (2015). (ii) Bayesian estimation: the likelihood function is an important component of Bayesian analysis, it is mendable by incorporating prior information. In fact, the penalization itself can be interpreted as a Bayesian estimation under appropriate priors. (iii) Dynamic factor models: the ML method can be extended to dynamic factor models such that $y_t=\Lambda_0 f_t+\Lambda_1 f_{t-1}+u_t$, where the second set of factors f_{t-1} is the lag of f_t. The PC method would treat the model as having $2r$ static factors, while the ML can estimate the model with r factors; the likelihood function is evaluated by the state space method via the Kalman smoother. The ML method can allow f_t itself to be dynamic, for example, Doz et al. (2012). In view of these advantages, it is of interest to study the ML-based method with a possibly non-diagonal error covariance matrix.

Although the penalized likelihood method has been used frequently in the recent literature of large covariance estimation (e.g., Lam and Fan, 2009), the problem being addressed here is technically different and challenging. This is because, besides penalizing Σ_u, the likelihood function is also highly nonlinear in Λ, and $\Lambda\Lambda'$ has r fast-diverging eigenvalues (at rate $O(N)$). In contrast, the literature has only focused on estimating covariances with bounded eigenvalues. Investigating the impact of these "very spiked" eigenvalues on the joint estimation of (Λ,Σ_u) is one of the goals of this paper. In addition, the penalized ML method can also be viewed as an alternative approach to that of Fan et al. (2013) to estimating the error covariance matrix in factor analysis, because it does not rely on the principal components method, and enjoys the advantages of the maximum likelihood method as discussed above. The matrix $\hat{\Lambda}\hat{\Lambda}'+\hat{\Sigma}_u$ is a high dimensional covariance estimator.

3. Theoretical properties

3.1. Sparsity assumptions

First, we define the sparsity condition on Σ_{u0}. The sparsity is characterized through an unknown partition of the off-diagonal elements. Let J_L and J_U denote two disjoint sets of the indices for small and large elements of Σ_{u0} in absolute value, and

$$\{(i,j):i\leq N,j\leq N\}=J_L\cup J_U.$$

Because the diagonal elements represent the individual variances of the idiosyncratic components, we assume $(i,i)\in J_U$ for all $i\leq N$. The sparsity assumes that most of the indices (i,j) belong to J_L when $i\neq j$. The following assumption quantifies the partition $\{(i,j):i\leq N,j\leq N\}=J_L\cup J_U$. The partition need not be unique, and our analysis suffices as long as such a partition exists. One does not need to know which elements belong to J_L or which elements belong to J_U. Define the number of off-diagonal large entries:

$$D_N=\sum_{i\neq j,(i,j)\in J_U}1. \tag{3.1}$$

Assumption 3.1. There exists a partition $\{(i,j):i\leq N,j\leq N\}=J_L\cup J_U$ where J_U and J_L are disjoint, which satisfies:

(i) $(i,i)\in J_U$ for all $i\leq N$, and $D_N=o(\min\{N\sqrt{T/\log N},N^2/\log N\})$,

(ii) $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|=o(N)$.

As there are $O(N^2)$ off-diagonal entries in total, Condition (i) requires that most off-diagonal entries of Σ_{u0} be inside J_L. Condition (ii) quantifies the absolute sum of all the "small" entries. In particular, we do allow $D_N\gg N$ and $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|$ to diverge, which means that Σ_{u0} need not be too sparse. It is likely that J_U only contains the diagonal elements. It then essentially corresponds to the strict factor model where Σ_{u0} is almost a diagonal matrix and error terms are only weakly cross-sectionally correlated, which is a special case of Assumption 3.1. Another special case arises when Σ_{u0} is strictly sparse, in the sense that its elements in small magnitudes (J_L) are exactly zero. For the banded matrix as an example, there is a finite integer k such that

$$\Sigma_{u0,ij}\neq 0\quad \text{if } |i-j|\leq k;\qquad \Sigma_{u0,ij}=0\quad \text{if } |i-j|>k.$$

Then $J_L=\{(i,j):|i-j|>k\}$ and $J_U=\{(i,j):|i-j|\leq k\}$. In this case $D_N=O(N)$, and $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|=0$. Hence all the conditions in the above assumption are satisfied. This assumption is also satisfied by block-diagonal matrices with finite block sizes.

To compare with the sparsity assumptions of Fan et al. (2013), consider an exact sparse case where $(i,j)\in J_L$ if and only if $\Sigma_{u0,ij}=0$, so many off-diagonal elements are exactly zero. Then our Assumption 3.1 simplifies to:

$$D_N=\sum_{i\neq j}1\{\Sigma_{u0,ij}\neq 0\}=o(\min\{N\sqrt{T/\log N},N^2/\log N\}), \tag{3.2}$$

while the sparsity condition in Fan et al. (2013) based on thresholding estimation in this case becomes:

$$m_N = \max_{i\le N}\sum_{j=1}^{N} 1\{\Sigma_{u0,ij}\neq 0\} = o(\min\{\sqrt{T/\log N}, \sqrt{N}\}). \qquad (3.3)$$

We see that as the covariance estimation approaches are different, the required sparsity conditions are of different types. Our condition controls the overall number of nonzero off-diagonal elements, whereas the condition of the thresholding estimator bounds the maximum number of nonzeros in the rows m_N. Admitted that while Fan et al. (2013) achieves the convergence under the matrix spectral norm, we achieve a weaker consistency (as in Theorem 3.1, the norm $\|.\|_F^2/N$ is weaker than the spectral norm), as a result, we require a weaker sparsity condition than theirs. Indeed, note that $D_N \le Nm_N$, so (3.3) implies (3.2). On the other hand, our weaker condition on the sparsity is more relevant, for instance, when Σ_{u0} has only finitely many very non-sparse rows (or columns), then our condition is still satisfied but (3.3) is not. This case arises in practice when there are only a small portion of firms or units whose individual shocks (u_{it}) are correlated with many other firms.

Our assumption allows all elements of Σ_{u0} to be nonzero, which is verified in the following example.

Example 3.1. Consider a cross-sectional AR(1) model with the covariance matrix: for some $\sigma_u^2 > 0$ and $\rho \in (0, 1)$,

$$\Sigma_{u0} = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{N-1} \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho^{N-1} & \cdots & \rho & 1 \end{bmatrix}.$$

Let $c \in (0, 1)$ be a small number. Let J_U consist of the diagonal and N^c number of rows below and above the diagonal (banded elements), then there are at most $2NN^c = O(N^{1+c})$ number of "big elements". The sum of small elements (all the rest of elements) is (ignoring the constant)

$$2\sum_{j=1}^{N^c} j\rho^{N-j} \le 2\rho^{N-N^c}\sum_{j=1}^{N^c} j \le C\rho^{N-N^c}N^{2c} \to 0$$

for some $C > 0$, as N goes to infinity. So $\Sigma_{ij\in J_L}|\Sigma_{u0,ij}| = o(N)$. On the other hand, as long as T is large, N^{1+c} is smaller than $N\sqrt{T/\log N}$, so the number of "big elements" is not so large. Hence Assumption 3.1 is satisfied. In practice, the cross-sectional variables are arbitrarily ordered. The above argument is valid as long as there exists an permutation of the cross-sections variable such that AR(1) holds. The same argument applies to a much more general dependent structure than AR(1). □

3.2. Assumptions on the data generating process

The following assumption provides the regularity conditions on the data generating process. We allow the serial dependence across t by introducing the strong mixing condition. Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_T^{\infty}$ denote the σ-algebras generated by $\{(f_t, u_t) : -\infty \le t \le 0\}$ and $\{(f_t, u_t) : T \le t \le \infty\}$ respectively. In addition, define the mixing coefficient

$$\alpha(T) = \sup_{A\in\mathcal{F}_{-\infty}^{0}, B\in\mathcal{F}_T^{\infty}} |P(A)P(B) - P(AB)|. \qquad (3.4)$$

Assumption 3.2. (i) $\{u_t, f_t\}_{t\ge 1}$ is strictly stationary. In addition, $Eu_{it} = Eu_{it}f_{jt} = 0$ for all $i \le p, j \le r$ and $t \le T$.

(ii) There exist constants $c_1, c_2 > 0$ such that $c_2 < \lambda_{\min}(\Sigma_{u0}) \le \lambda_{\max}(\Sigma_{u0}) < c_1$, and $\max_{j\le N}\|\lambda_{0j}\| < c_1$.

(iii) There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$, such that for any $s > 0$, $i \le p$ and $j \le r$,

$$P(|u_{it}| > s) \le \exp(-(s/b_1)^{r_1}), \quad P(|f_{jt}| > s) \le \exp(-(s/b_2)^{r_2}).$$

(iv) Strong mixing: There exist $r_3 > 0$ and $C > 0$ satisfying: for all $T \in \mathbb{Z}^+$,

$$\alpha(T) \le \exp(-CT^{r_3}).$$

We allow factors and errors to be serially weakly dependent, satisfying the strong mixing condition. The error covariance is assumed to have bounded eigenvalues, which is common in the factor model literature. Unlike the common factors, the information of idiosyncratic component u_{it} does not grow with the increase of the dimension. For instance, if firms' individual idiosyncratic shocks are correlated within the same industry, but uncorrelated across industries, Σ_u is a block-diagonal matrix with bounded eigenvalues. Condition (iii) also requires exponential-tail bounds, which is a technical condition for high-dimensional weakly dependent data: it allows us to apply large deviation theories to achieve uniform convergences.

The following assumption is standard in the approximate factor models, see e.g., Stock and Watson (1998; 2002). It implies that the first r eigenvalues of $\Lambda_0\Lambda_0'$ are growing rapidly at $O(N)$. Intuitively, it requires the factors be pervasive in the sense that they impact a non-vanishing proportion of time series $\{y_{1t}\}_{t\le T}, \ldots, \{y_{Nt}\}_{t\le T}$. We focus on the case where factors are strong. While our results are possibly extendable to allow for "weaker factors" (e.g., Onatski, 2012; Natalia et al., 2012), it might be technically challenging.

Assumption 3.3. There is a $\delta > 0$ such that for all large N,

$$\delta^{-1} < \lambda_{\min}(N^{-1}\Lambda_0'\Lambda_0) \le \lambda_{\max}(N^{-1}\Lambda_0'\Lambda_0) < \delta.$$

Therefore all the eigenvalues of $N^{-1}\Lambda_0'\Lambda_0$ are bounded away from both zero and infinity as $N \to \infty$.

The following assumption is imposed on the penalty weights. Define the weights ratios

$$\eta_T = \frac{\max_{i\neq j,(i,j)\in J_U} w_{ij}}{\min_{(i,j)\in J_L} w_{ij}}, \qquad \beta_T = \frac{\max_{(i,j)\in J_L} w_{ij}}{\min_{(i,j)\in J_L} w_{ij}}.$$

We assume upper bounds on η_T and β_T respectively (condition (i) in Assumption 3.4). Intuitively, the upper bound on η_T requires the penalty weights on the estimated "large" entries of Σ_u should not be large relative to those on the estimated "small" entries. This eliminates biases from penalizing elements in J_U. On the other hand, the required upper bound on β_T helps control the penalty on "small" entries in a universal scale.

Assumption 3.4. The tuning parameter $\mu_{N,T}$ and the weights $\{w_{ij}\}_{i\le N, j\le N}$ satisfy:

(i)

$$\eta_T = o_P\left[\min\left\{\sqrt{\frac{T}{\log N}\frac{N}{D_N}}, \left(\frac{T}{\log N}\right)^{1/4}\sqrt{\frac{N}{D_N}}, \frac{N}{\sqrt{D_N\log N}}\right\}\right],$$

$$\beta_T\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}| = o_P(N),$$

(ii) $\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}| = o(\min\{N, N^2/D_N, N^2/(D_N\eta_T^2)\})$,

$\mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij} = o(\min\{N/D_N, \sqrt{N/D_N}, N/(D_N\eta_T)\})$,

$\mu_{N,T}\min_{(i,j)\in J_L} w_{ij} \gg \sqrt{\log N/T} + (\log N)/N$.

The above assumption is not as complicated as it looks, and is satisfied by many examples. For instance, the Lasso penalty sets $w_{ij} = 1$ for all $i, j \leq N$. Hence $\eta_T = \beta_T = 1$. Then condition (i) of Assumption 3.4 follows from Assumption 3.1, which is also satisfied if $D_N = O(N)$. Condition (ii) is also straightforward to verify. This immediately implies the following lemma.

Lemma 3.1 (*Lasso*). *Choose $w_{ij} = 1$ for all $i, j \leq N, i \neq j$. Suppose in addition $D_N = O(N)$ and $\log N = o(T)$. Then Assumption 3.4 is satisfied if the tuning parameter $\mu_{N,T} = o(1)$ is such that*

$$\sqrt{\frac{\log N}{T}} + \frac{\log N}{N} = o(\mu_{N,T}).$$

This assumption is also satisfied by SCAD and adaptive lasso. We will also verify this assumption for these two penalties in Section 3.4.

3.3. Consistency of the joint estimation

We assume the parameter space for Σ_{u0} to be, for some known sufficiently large $M > 0$,

$$\Gamma = \{\Sigma_u : \|\Sigma_u\|_1 < M, \|\Sigma_u^{-1}\|_1 < M\}.$$

Then $\Sigma_{u0} \in \Gamma$ implies that all the eigenvalues of Σ_{u0} are bounded away from both zero and infinity. There are many examples where both the covariance and its inverse have bounded row sums. For example, for each t, when $\{u_{it}\}_{i=1}^N$ follows a cross sectional autoregressive process AR(p) for some fixed p, then the maximum row sum of Σ_{u0} is bounded. The inverse of Σ_{u0} is a banded matrix, whose maximum row sum is also bounded. In view of Assumption 3.3, we define the parameter space:

$$\Theta_\lambda = \{\Lambda : \delta^{-1} < \lambda_{\min}(N^{-1}\Lambda'\Lambda) \leq \lambda_{\max}(N^{-1}\Lambda'\Lambda) < \delta, \\ \Lambda'\Sigma_u^{-1}\Lambda \text{ is diagonal}\}. \tag{3.5}$$

Remark 3.1. The parameter space provides restrictions on the lower and upper bounds of the loadings and the error covariance matrix. These parameter restrictions are needed to prove consistency. On the other hand, when computing the PML in practice, we do not find it necessary to impose these bounds in the calculations. As to be shown in Theorem 4.1 later, starting from a consistent initial value that belongs to the parameter space, the updated solution in each iterative step is also consistent.

Our main theorem is stated as follows. The consistency of $\widehat{\Sigma}_u$ is in terms of the weighted Frobenius norm $\frac{1}{N}\|.\|_F^2$, which is a commonly used assessment for the convergence of sparse covariance estimators (e.g., Rothman et al., 2008; Lam and Fan, 2009). Moreover, it is natural to define the consistency in terms of the averaged estimation errors $\frac{1}{N}\|\widehat{\Lambda} - \Lambda_0\|_F^2 = \frac{1}{N}\sum_{i=1}^N \|\widehat{\lambda}_i - \lambda_{0i}\|^2$. In addition, the theorem presents consistency for the general case where the weights w_{ij} of the ℓ_1-penalizations satisfy Assumption 3.4. By Lemma 3.1, it immediately follows that the LASSO-penalty is included. In addition, we shall show in next subsection that both adaptive LASSO and SCAD also satisfy Assumption 3.4, hence the theorem below also applies to them.

Theorem 3.1. *Suppose $\log N = o(T)$. Under Assumptions 3.1 and 3.4, the penalized ML estimator satisfies: as T and $N \to \infty$,*

$$\frac{1}{N}\|\widehat{\Sigma}_u - \Sigma_{u0}\|_F^2 \to^P 0, \qquad \frac{1}{N}\|\widehat{\Lambda} - \Lambda_0\|_F^2 \to^P 0.$$

For each $t \leq T$,

$$\|\widehat{f}_t - f_t\| \to^P 0.$$

Remark 3.2. In the high-dimensional penalized likelihood literature, to establish the consistency one usually constructs a neighborhood of the true parameters $(\Lambda_0, \Sigma_{u0}) \in U$ (e.g., Rothman et al., 2008; Lam and Fan, 2009), and show that $L_1(\Lambda_0, \Sigma_{u0}) < \inf_{(\Lambda,\Sigma_u)\in\partial U} L_1(\Lambda, \Sigma_u)$ with probability approaching one, where ∂U denotes the boundary of the neighborhood and $L_1(\Lambda, \Sigma_u)$ is the objective function. This strategy however, does not work here due to the technical difficulty in dealing with the term $(\Lambda\Lambda' + \Sigma_u)$ in the likelihood function. This is because its largest r eigenvalues are unbounded and grow at rate $O(N)$ uniformly in the parameter space. In the proof of Theorem 3.1, we use a new strategy to analyze the penalized likelihood function involving diverging eigenvalues.

3.4. Two examples

We present two alternative choices for the weights: one is adaptive lasso, proposed by Zou (2006), and the other is SCAD by Fan and Li (2001). Both weights depend on a preliminary consistent estimate of each element of Σ_{u0}. A simple consistent estimate for each element can be obtained by the principal component (Stock and Watson, 2002).

To simplify the presentation, we will assume that $D_N = O(N)$, which controls the number of off-diagonal large entries of Σ_{u0}. Moreover, we assume the small and large entries of Σ_{u0} are well-separated:

$$\max\{|\Sigma_{u0,ij}| : (i, j) \in J_L\} \ll \omega_T \ll \min\{|\Sigma_{u0,ij}| : \Sigma_{u0,ij} \in J_U\},$$

where throughout the paper, we write

$$\omega_T = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}.$$

Let the initial estimate $\widehat{\Sigma}^*_{u,ij} = R_{ij}$, where R_{ij} is the PC estimator of $\Sigma_{u0,ij}$. The adaptive lasso chooses the weights to be,

$$\text{(Adaptive Lasso)} : \quad w_{ij} = (|\widehat{\Sigma}^*_{u,ij}| + \delta_T)^{-1}, \tag{3.6}$$

where $\delta_T = o(1)$ is a pre-determined nonnegative sequence. The additive δ_T was not included in the original definition of adaptive lasso in Zou (2006), but is used here to prevent w_{ij} getting too large if $|\widehat{\Sigma}^*_{u,ij}|$ is very close to zero. This small adjustment makes the estimator less sensitive to the initial PC estimate. The adaptive lasso has been used extensively in the high dimensional literature, see for example, Huang et al. (2008), van de Geer et al. (2011) and Caner and Fan (2011), etc.

Another important example is SCAD, which is folded concave as defined in (2.5).

The following assumption is needed to ensure the good behavior of w_{ij} for both adaptive lasso and SCAD.

Assumption 3.5. Assume that $E\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \lambda_i u_{it}\|^2 = O(1)$ and $E(\frac{1}{\sqrt{N}}\sum_{i=1}^N(u_{it}u_{is} - Eu_{it}u_{is}))^2 = O(1)$.

We have the following theorem.

Theorem 3.2. *Suppose either the Adaptive Lasso or SCAD is used for the weighted-ℓ_1 penalized objective function. Also, suppose $\log N = o(T)$, $D_N = O(N)$, $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}| = o(N)$ and Assumptions 3.1–3.3 and 3.5 hold. In addition, assume the tuning parameters are such that:*

(i) *for Adaptive Lasso,*

$$\omega_T\left(\frac{\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|}{N}\right) \ll \delta_T \ll \omega_T, \tag{3.7}$$

$$\omega_T^2 \ll \mu_{N,T} \ll \omega_T. \tag{3.8}$$

(ii) for SCAD:

$$\left(\frac{\log N}{T}\right)^{1/4}+\left(\frac{\log N}{N}\right)^{1/2}\ll \mu_{N,T}\ll \min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}|. \tag{3.9}$$

Then Assumption 3.4 is satisfied, and

$$\frac{1}{N}\|\widehat{\Sigma}_u-\Sigma_{u0}\|_F^2\to^P 0,\qquad \frac{1}{N}\|\widehat{\Lambda}-\Lambda_0\|_F^2\to^P 0.$$
$$\|\widehat{f}_t-f_t\|\to^P 0.$$

Like Lemma 3.1, an attractive feature of this theorem is that, if both the upper bound of $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|$ and the lower bound of $\min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}|$ are known, [e.g., in the strictly sparse model, $\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|=0$, and assume $\min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}|$ is bounded away from zero as in MA(1)] then Conditions (3.7)–(3.9) do not depend on any other unknown feature of Σ_{u0}.

4. Implementations

4.1. Majorize–minimize EM algorithm

In this section we discuss the computational issues. Note that even with a known Λ, numerically minimizing the loss function with respect to Σ_u is difficult, because $\log\left|\det\left(\Lambda\Lambda'+\Sigma_u\right)\right|$ is concave in Σ_u, while $\operatorname{tr}\left(S_y(\Lambda\Lambda'+\Sigma_u)^{-1}\right)$ is convex. Therefore, the optimization is a concave + convex problem, and are often solved approximately. One of the commonly used approaches is majorize–minimize, which approximates the concave component by a "majorizing" linear function of Σ_u, using the tangent plane. Then the objective function is approximated by a convex function.

Updating Σ_u

Suppose $(\widehat{\Lambda}_k,\widehat{\Sigma}_{u,k})$ is the covariance matrix in the kth iteration. We now update Σ_u by approximately minimizing $L(\widehat{\Lambda}_k,\Sigma_u)+\frac{1}{N}\sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}|$. The tangent plane of $\log\left|\det\left(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\Sigma_u\right)\right|$ at $\Sigma_u=\widehat{\Sigma}_{u,k}$ is

$$\log\left|\det\left(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k}\right)\right|+\operatorname{tr}((\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}(\Sigma_u-\widehat{\Sigma}_{u,k})).$$

Then instead of minimizing the original problem, we minimize

$$\operatorname{tr}((\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}(\Sigma_u-\widehat{\Sigma}_{u,k}))+\operatorname{tr}\left(S_y(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\Sigma_u)^{-1}\right)$$
$$+\sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}| \tag{4.1}$$

with respect to Σ_u, which is now convex.

For the convex problem (4.1), many algorithms in the recent literature on covariance estimations solve the problem column-by-column, e.g., Friedman et al. (2008) and Rothman (2012). However, we find that column-by-column iterating is slow when the dimension of Σ_u is relatively large. Alternatively, we employ the *projected gradient* algorithm recently proposed by Bien and Tibshirani (2011), which further approximates $\operatorname{tr}((\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}(\Sigma_u-\widehat{\Sigma}_{u,k}))+\operatorname{tr}\left(S_y(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\Sigma_u)^{-1}\right)$ by

$$\tilde{L}(\Sigma_u)=\frac{1}{2t}\|\Sigma_u-\widehat{\Sigma}_{u,k}+t[(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}$$
$$-(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}S_y(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}]\|_F^2$$

where t is the depth of projection (see Bien and Tibshirani, 2011). Solving

$$\Sigma_{u,k+1}=\arg\min_{\Sigma_u}\frac{1}{N}\tilde{L}(\Sigma_u)+\frac{1}{N}\sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}|$$

yields an analytical solution: for $B=\widehat{\Sigma}_{u,k}-t[(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}-(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}S_y(\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k})^{-1}]$,

$$(\Sigma_{u,k+1})_{i,j}=\begin{cases}\operatorname{sign}(B_{ij})(|B_{ij}|-\mu_{N,T}w_{ij}t)_+ & \text{if } i\neq j\\ B_{ij} & \text{if } i=j.\end{cases}$$

Here $(x)_+=\max\{0,x\}$. Thus we gain a much faster iterating algorithm than the column-by-column procedure. We note that the trade-off between computational efficiency and the use of approximations often exists in statistical computings of high-dimensional problems. While we cannot expect solving such an approximated problem yields a global minimum of our nonconvex problem, existing research on marjorize–minimize algorithms (e.g. An and Tao, 2005) shows that limiting points of such an algorithm are critical points of the original penalized ML problem.

Updating Λ

Given the current $(\widehat{\Lambda}_k,\widehat{\Sigma}_{u,k})$, Λ is updated using the EM algorithm (e.g., Bai and Li, 2012a,b). The EM algorithm updates the estimator according to: for $\widehat{\Sigma}_{y,k}=\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k}$,

$$\widehat{\Lambda}_{k+1}=S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_kM^{-1},$$
$$M=\widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k+I_r-\widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k.$$

The algorithm is summarized as follows. Step 2 is the E-step and step 3 is the majorize–minimize step. Note that the majorize–minimize method in Bien and Tibshirani (2011) uses two loops to update the covariance estimator until convergence. In contrast, step 3 in the following algorithm only updates $\widehat{\Sigma}_u$ by one step, which speeds up the convergence.

1. Set $k=0$. Initialize $\widehat{\Lambda}_0$ and $\widehat{\Sigma}_{u,0}$.
2. At step $k+1$, $\widehat{\Lambda}_{k+1}=AM^{-1}$, where $M=\widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k+I_r-\widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k$,
$$A=S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k,\quad \widehat{\Sigma}_{y,k}=\widehat{\Lambda}_k\widehat{\Lambda}_k'+\widehat{\Sigma}_{u,k}.$$
3. Still at step $k+1$, for some small value $t>0$ (fixed as 0.1 in all our numerical studies), let
$$B=\widehat{\Sigma}_{u,k}-t(\widehat{\Sigma}_{y,k}^{-1}-\widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}),$$
$$\widehat{\Sigma}_{u,k+1}=((\widehat{\Sigma}_{u,k+1})_{i,j})_{N\times N},$$
where $(\widehat{\Sigma}_{u,k+1})_{i,j}$ denotes the (i,j)th element of $\widehat{\Sigma}_{u,k+1}$, given by
$$(\widehat{\Sigma}_{u,k+1})_{i,j}=\begin{cases}\mathit{sign}(B_{ij})(|B_{ij}|-\mu_{N,T}w_{ij}t)_+ & \text{if } i\neq j\\ B_{ij} & \text{if } i=j.\end{cases}$$
4. Repeat 2–3 until convergence.

Although the algorithm does not impose the restrictions as those of the parameter space, it can be shown that as long as a consistent initial value is used, and belongs to the parameter space, the updated solution in each iterative step is also consistent. We formally state this result in the following theorem. For a matrix $A=(a_{ij})$, let $\|A\|_\infty=\max_{ij}|a_{ij}|$. Recall that $m_N=\max_{i\leq N}\sum_{j=1}^N 1\{\Sigma_{u0,ij}\neq 0\}$. For the asymptotic analysis of the algorithm, we shall assume $\mu_{N,T}m_N\to 0$ and $\sqrt{\frac{\log N}{T}}=o(\mu_{N,T})$. For technical simplicity, in the following theorem, we focus on the SCAD weights.

Theorem 4.1. *Suppose at step k, $(\widehat{\Lambda}_k,\widehat{\Sigma}_{u,k})$ satisfy:*

(i) $\|\widehat{\Lambda}_k-\Lambda_0\|_\infty=o_P(\mu_{N,T})$,

(ii) $\|\widehat{\Sigma}_{u,k}-\Sigma_{u0}\|_\infty=o_P(\mu_{N,T})$,

(iii) $\|\widehat{\Sigma}_{u,k}^{-1}\|_1<M$, $\|\widehat{\Sigma}_{u,k}\|_1<M$.

Then at step $k+1$,

$$\|\widehat{\Lambda}_{k+1}-\Lambda_0\|_\infty=o_P(\mu_{N,T}),\quad \|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_\infty=o_P(\mu_{N,T}),$$
$$\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1=o_P(\mu_{N,T}m_N),$$

and $\|\widehat{\Sigma}_{u,k+1}^{-1}\|_1\le\|\Sigma_{u0}^{-1}\|_1+o_P(\mu_{N,T}m_N)$, $\quad\|\widehat{\Sigma}_{u,k+1}\|_1\le\|\Sigma_{u0}\|_1+o_P(\mu_{N,T}m_N)$,

$$\lambda_{\min}\left(\frac{\Lambda_0'\Lambda_0}{N}\right)-o_P(\mu_{N,T})<\lambda_{\min}\left(\frac{\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1}}{N}\right)$$
$$\le\lambda_{\max}\left(\frac{\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1}}{N}\right)<\lambda_{\max}\left(\frac{\Lambda_0'\Lambda_0}{N}\right)+o_P(\mu_{N,T}).$$

Note that when $N,T\to\infty$, many existing loading estimators satisfy condition (i): e.g., PC and generalized PC ($\|\widehat{\Lambda}-\Lambda_0\|_\infty=O_P(\sqrt{\frac{\log N}{T}}+\frac{1}{\sqrt{N}})$ for PC-based estimators, see Fan et al. (2013), Choi (2012)). Also, the thresholded covariance estimator of Fan et al. (2013) satisfies conditions (ii)(iii) (for any $\mu_{N,T}\gg\sqrt{\frac{\log N}{T}}+\frac{1}{\sqrt{N}}$). Therefore, they can be used as the initial estimators at step $k=0$. Moreover, given the nonlinear objective function, it is beneficial to adopt various starting values from pre-estimators to find the global optimizer.

It is important to show that the algorithm being employed indeed converges to the defined estimator. This involves showing that: (i) the algorithm converges; (ii) the limiting point of the algorithm is also a stationary point of the penalized-ML problem, and (iii) the stationary point that the algorithm converges to also satisfies the restrictions in (3.5). In fact, (iii) is partially achieved by Theorem 4.1. The remaining task is to show the convergence point also satisfies the restriction. In addition, Lemma D.2 in the Appendix shows that (ii) is also achieved. Finally, we acknowledge that proving (i) is theoretically important and yet challenging. We shall leave it for the future research.

4.2. Choosing the tuning parameter by cross-validations

We suggest choosing $\mu_{N,T}$ based on K-fold cross validations, which is also a common practice in the literature on estimating large covariances using penalized methods (e.g., Bien and Tibshirani, 2011 and Xue et al., 2012). For a given index set of validation data $\mathcal{A}\subset\{1,\ldots,T\}$, let $S_{y,\mathcal{A}}=|\mathcal{A}|_0^{-1}\sum_{t\in\mathcal{A}}y_ty_t'$, which is the sample covariance calculated using the validation data. Here $|\mathcal{A}|_0$ denotes the cardinality of $\mathcal{A}$. Let $\widehat{\Lambda}(\mathcal{A}^c,\mu)$, $\widehat{\Sigma}_u(\mathcal{A}^c,\mu)$ denote the estimated loading and covariance matrices using the training data in $\mathcal{A}^c$, based on a tuning parameter $\mu_{N,T}=\mu$. Partitioning $\{1,\ldots,T\}$ into K subsets $\mathcal{A}_1,\ldots,\mathcal{A}_K$, we would like to choose a value $\mu_{N,T}=\mu$ that minimizes

$$\frac{1}{K}\sum_{k=1}^K L(\widehat{\Lambda}(\mathcal{A}_k^c,\mu),\widehat{\Sigma}_u(\mathcal{A}_k^c,\mu),S_{y,\mathcal{A}_k})$$

where

$$L(\widehat{\Lambda},\widehat{\Sigma}_u,S_y)=\frac{1}{N}\log\left|\det\left(\widehat{\Lambda}\widehat{\Lambda}'+\widehat{\Sigma}_u\right)\right|+\frac{1}{N}\mathrm{tr}\left(S_y(\widehat{\Lambda}\widehat{\Lambda}'+\widehat{\Sigma}_u)^{-1}\right).$$

5. Numerical illustrations

5.1. Simulation result

We present a numerical experiment to illustrate the performance of the proposed method. The data was generated as following: $\{e_{it}\}_{i\le N,t\le T}$ are both serially and cross-sectionally independent as $\mathcal{N}(0,1)$. Let

$$u_{1t}=e_{1t},\qquad u_{2t}=e_{2t}+a_1e_{1t},\qquad u_{3t}=e_{3t}+a_2e_{2t}+b_1e_{1t},$$
$$u_{i+1,t}=e_{i+1,t}+a_ie_{it}+b_{i-1}e_{i-1,t}+c_{i-2}e_{i-2,t},$$

where $\{a_i,b_i,c_i\}_{i=1}^N$ are independently from $0.7\mathcal{N}(0,1)$. Let the two factors $\{f_{1t},f_{2t}\}$ be i.i.d. $\mathcal{N}(0,1)$, and $\{\lambda_{i,1},\lambda_{i,2}\}_{i\le N}$ be uniform on $[0,1]$. Then Σ_{u0} is a banded matrix.

We apply the SCAD penalty for our joint estimation, with various choices of the tuning parameter $\mu_{N,T}$. The estimator is compared with three other competing methods: (1) the PC-estimator, (2) the estimator of (unpenalized) heteroskedastic ML, denoted by HML, and (3) the feasible efficient PC (denoted by EPC). The EPC uses the generalized PC method combined with PC-based covariance matrix estimator of Fan et al. (2013), which was formally studied recently by Bai and Liao (2013). Note that HML estimates Σ_u to be diagonal, which solves:

$$\min_{\Sigma_{u,ij}=0\text{ for }i\neq j}\min_{\Lambda}\frac{1}{N}\log|\Lambda\Lambda'+\Sigma_u|+\frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda'+\Sigma_u)^{-1}).\quad(5.1)$$

In our simulation setup, Σ_{u0} is non-diagonal, so HML does not take the idiosyncratic cross-sectional dependence into account. Moreover, the EPC combines the generalized PC method of Choi (2012) with the thresholded estimator of Σ_u. Specifically, it estimates the factors by the principal components of the $T\times T$ matrix $Y'\widetilde{\Sigma}_u^{-1}Y$, where the covariance estimator $\widetilde{\Sigma}_u^{-1}$ is recently developed by Fan et al. (2013). In our simulation, we tried four thresholding parameters for estimating Σ_u^{-1} using Fan et al. (2013)'s approach: $C=0.05, 0.1, 0.5, 0.7$, and find that $C=0.5$ and 0.7 yield better performance, which are reported here.

For each estimator, the smallest canonical correlation (the higher the better) between the estimator and the parameter is used as a measurement to assess the accuracy of each estimator. We employed two pre-estimators as starting values, PC and HML, to compute the proposed PML. Both starting values yield the same numerical result. Table 1 lists the results based on one thousand replications. When N is relatively small, it is clear that the ML-based methods (proposed PML and HML) perform better than the PC-based methods (PC and efficient PC). Taking into account cross-sectional dependence, the PML further outperforms HML; the latter treats Σ_{u0} to be diagonal, without penalizations. When N is relatively large, it is hard to see whether PML or EPC dominates the other; both methods estimate Σ_{u0} consistently. This further demonstrates that the ML-based method is desirable especially for relatively small N, and is also very competitive for large N. We point out that HML coincides with Doz et al. (2011) for static factor models.

5.2. Forecast based on simulated data

This section numerically illustrates the improvement of time series forecast based on efficient estimations of the factor model. As in Stock and Watson (2002), we aim at forecasting a time series model with a single factor:

$$x_{t+1}=\beta f_t+\epsilon_t,\quad f_t=\rho f_{t-1}+v_t,$$

where the unknown factors can be learned from a factor model: $y_t=\Lambda f_t+u_t$.

We set $\beta=2$, and $\rho=0.5$. The data generating process for Λ and u_t are the same as those in Section 5.1. We conduct one-step-ahead out-of-sample forecast m times using a moving window of a fixed size T. Here T is also the sample size for estimations, and in our numerical study $T=50, 100$ are used. We simulate $m+T$ observations in total. For each $t=0,\ldots,m-1$, we use the data $\{(y_{t+1},x_{t+1}),\ldots,(y_{t+T},x_{t+T})\}$ to conduct one-step-ahead forecast of x_{t+T+1}. Specifically, we estimate the factors using

Table 1
Canonical correlations: the larger the better. Simulation of Section 5.1.

			PML			EPC		HML	PC
	T	N	$\mu_{N,T} = 0.08$	$\mu_{N,T} = 0.2$	$\mu_{N,T}$ =CV	$C = 0.7$	$C = 0.5$		
Loadings	50	50	0.395	0.398	0.394	0.309	0.298	0.369	0.277
		100	0.605	0.600	0.596	0.483	0.455	0.552	0.424
		150	0.656	0.655	0.654	0.638	0.600	0.609	0.468
	100	50	0.424	0.447	0.461	0.292	0.283	0.368	0.263
		100	0.756	0.748	0.743	0.664	0.625	0.664	0.554
		150	0.845	0.847	0.846	0.834	0.821	0.822	0.781
Factors	50	50	0.394	0.400	0.395	0.319	0.307	0.354	0.276
		100	0.708	0.705	0.698	0.559	0.609	0.629	0.464
		150	0.838	0.838	0.836	0.788	0.731	0.763	0.573
	100	50	0.489	0.503	0.513	0.379	0.373	0.439	0.358
		100	0.823	0.816	0.798	0.715	0.666	0.674	0.547
		150	0.938	0.944	0.938	0.927	0.905	0.887	0.817

Averaged results based on one thousand replications are presented. PML represents the proposed method. $\mu_{N,T} = 0.08, 0.2$ represent the tuning used for PML and $\mu_{N,T}$ = CV is the tuning chosen by 5-fold cross-validation ; $C = 0.7$ and $C = 0.5$ are the thresholds used to estimate $\widehat{\Sigma}_u$ (Fan et al., 2013).

Table 2
Methods.

Name	Short description
PML	Proposed penalized ML
PC	Principal components
HML	ML with diagonal Σ_{u0}, solving (5.1)
DOZ	Kalman smoother, proposed in Doz et al. (2011)
EPC	Weighted principal components with Fan et al. (2013)'s estimated Σ_{u0}^{-1} as the weight matrix

$\{y_{t+1}, \ldots, y_{t+T}\}$, and obtain $\{\widehat{f}_{t+1}, \ldots, \widehat{f}_{t+T}\}$. The coefficient β in the forecasting regression is then estimated by the OLS in the regression of $\{x_{t+2}, \ldots, x_{t+T}\}$ onto $\{\widehat{f}_{t+1}, \ldots, \widehat{f}_{t+T-1}\}$, denoted by $\widehat{\beta}_{t+T}$. We then forecast x_{t+T+1} by $\widehat{x}_{t+T+1|t+T} = \widehat{\beta}_{t+T}\widehat{f}_{t+T}$. The forecasting error is then $(x_{t+T+1} - \widehat{x}_{t+T+1|t+T})^2$. Such a procedure continues for $t = 0, \ldots, m-1$.

Five methods are compared: PML, PC, HML, DOZ, and EPC. In particular, the DOZ method, proposed by Doz et al. (2011), applies the Kalman smoother for estimating the state space f_t. The Kalman smoother takes into account the dynamics in f_t. In terms of estimating the state variable f_t, it can be shown that HML and DOZ estimators are asymptotically equivalent under large N, despite that the latter takes into account the dynamics in f_t.

For each method M, we calculate the mean squared out-of-sample forecasting error:

$$MSE(\mathrm{M}) = \frac{1}{m}\sum_{t=0}^{m-1}(x_{t+T+1} - \widehat{x}_{t+T+1|t+T})^2,$$

and report the relative MSE to the PC method:

$$\frac{MSE(\mathrm{M})}{MSE(\mathrm{PC})}, \quad \mathrm{M} = \mathrm{PML, EPC, HML, DOZ}.$$

The results are reported in Table 3 for $m = 50$. Because the cross-sectional correlations are taken into account, PML and EPC perform significantly better than PC. This demonstrates that more efficient estimations of the factors/loadings also result in better forecasts in this model. The improvement of PML is more significant for a relatively small N. We note that PML is also competitive when the tuning parameter is chosen by the 5-fold cross-validation ($\mu_{N,T}$ = CV). When $N = 150$, it is hard to see whether PML or EPC dominates the other. This result is in contrast with Luciani (2014), who uses a different method to incorporate the cross-sectional correlations in the errors u_{it}. More specifically, Luciani (2014) uses PC or unpenalized ML to estimate the factors and the residuals $\widehat{u}_{it}$. The forecasts are constructed using both f_t and all residuals $\widehat{u}_{it}$ as predictors, but with LASSO penalty on the regression coefficients for the residuals.

5.3. *Diffusion index forecast based on real data*

This section compares the impact of how factors are estimated on real-data forecasts. We present the forecast results of the industrial production based on real-time macroeconomic time series of the United States. The dataset consists of 131 series of monthly data spanning the period from 1959 to 2007 (with a total of $\mathcal{T} = 528$ sampling periods), and was previously studied by Ludvigson and Ng (2011). Using the information criterion, Ludvigson and Ng (2011) finds eight factors. We adopt the diffusion index framework as in Stock and Watson (2002) to model the multi-step ahead variable:

$$x_{t+h}^h = \alpha_h + \beta_h f_t + \gamma_{1h}x_t + \cdots + \gamma_{lh}x_{t-l} + \epsilon_{t+h}^h, \quad y_t = \Lambda f_t + u_t,$$

where $x_{t+h}^h = \frac{1}{h}\sum_{i=1}^h x_{t+i}$ is the h-step-ahead variable to be forecast, defined in Stock and Watson (2002), and is specified to be the industrial production.

Similar to Section 5.2, forecasts of x_{t+h}^h are constructed based on a moving window with a fixed length ($T = 422 = 0.8\mathcal{T}$). For each fixed window, the sample data of y_t are first used to estimate a time series of factors, using one of the five methods in Table 2. We then forecast x_{T+h}^h by

$$\widehat{x}_{T+h|T}^h = \widehat{\alpha}_h + \widehat{\beta}_h\widehat{f}_T + \widehat{\gamma}_{1h}x_T + \cdots + \widehat{\gamma}_{lh}x_{T-l},$$

where the coefficients are estimated by regressing x_{t+h}^h onto a constant, $\widehat{f}_t$ and x_t (and l lags). For the proposed PML, the tuning parameters are selected using the 5-fold cross-validations.

As noted in Boivin and Ng (2005), how the factors are estimated can affect the mean-squared forecast error. The out-of-sample relative forecast MSE for each method (relative to the PC) is reported in Table 4. We report results for $h = 1, 12, 24$, $r = 7, 8$ factors, and $l = 1, 3$ lags. It is observed that in 1-period-ahead out-of-sample forecast, the differences across methods are not so strong as to immediately favor a particular method, which is consistent with the findings in Boivin and Ng (2005) and Luciani (2014). In particular, Doz et al. (2011)'s method is based on a Kalman filter, which models the dynamic factors using a VAR(1) model. Boivin and Ng (2005) also noted that dynamic factor models do not forecast better, and our finding is consistent with theirs. We also observe that in 12 and 24 period-ahead forecasts, using eight factors is significantly better than using 7 factors for the PML method. On the other hand, forecasts of long horizons may be less reliable due to the potential loss of the stationarity.

Table 3
$MSE(\mathrm{M})/MSE(\mathrm{PC})$.

T	N	PML			EPC		HML	DOZ
		$\mu_{N,T} = 0.08$	$\mu_{N,T} = 0.2$	$\mu_{N,T}$ =CV	$C = 0.7$	$C = 0.1$		
50	50	0.657	0.647	0.642	0.865	0.873	0.655	0.669
	100	0.801	0.732	0.771	0.852	0.833	0.852	0.856
	150	0.914	0.879	0.901	0.943	0.940	1.008	0.992
100	50	0.616	0.642	0.609	0.756	0.751	0.799	0.837
	100	0.666	0.686	0.616	0.761	0.768	0.794	0.814
	150	0.817	0.863	0.887	0.853	0.917	0.925	0.911

Table 4
Relative MSE for out-of-sample forecast: The benchmark method is PC.

		PML	EPC	HML	DOZ
		1-period-ahead forecast			
One lag	7 factors	0.959	0.989	1.068	1.068
	8 factors	0.957	1.013	1.070	1.035
Three lags	7 factors	0.947	0.992	1.066	1.060
	8 factors	0.944	1.012	1.068	1.027
		12-period-ahead forecast			
One lag	7 factors	0.999	0.979	1.129	1.049
	8 factors	0.642	1.009	1.052	1.029
Three lags	7 factors	1.036	0.948	1.154	1.051
	8 factors	0.631	0.993	1.053	1.050
		24-period-ahead forecast			
One lag	7 factors	0.813	0.853	0.881	0.951
	8 factors	0.640	0.948	0.875	1.002
Three lags	7 factors	0.822	0.946	0.893	0.953
	8 factors	0.625	0.942	0.875	1.021

The tuning parameter $\mu_{N,T}$ for PML is chosen by 5-fold cross-validations.

6. Conclusion

We study the estimation of a high dimensional approximate factor models in the presence of cross sectional dependence and heteroskedasticity. The classical PC method does not efficiently estimate the factor loadings or common factors because it essentially treats the idiosyncratic error to be homoskedastic and cross sectionally uncorrelated. For the efficient estimation it is essential to estimate a large error covariance matrix.

We assume the model to be conditionally sparse in the sense that after the common factors are taken out, the idiosyncratic components have a sparse covariance matrix. This enables us to combine the merits of both sparsity and high dimensional factor analysis. The method is based on the penalized maximum-likelihood, both involves regularizing a large covariance sparse matrix. Our method allows data-dependent adaptive penalties, such as adaptive Lasso and SCAD. We establish the consistency of these estimators.

Because the first order condition of the likelihood function is highly nonlinear in (Λ, Σ), and there are r fast-diverging eigenvalues (at order $O(N)$) in the involved covariance matrix $\Lambda_0\Lambda_0' + \Sigma_0$ due to the pervasiveness condition, it is challenging to obtain the optimal rate of convergence and limiting distributions of the proposed estimators directly. We shall leave it as a future research direction. Finally, it is important to show that the algorithm being employed converges to the defined estimator. We shall leave it for the future research.

Acknowledgments

We are grateful to three anonymous referees, an Associate Editor and the Co-Editor for their insightful comments, which helped generate a substantially improved paper.

Appendix A. Technical lemmas

Define

$$Q_1(\Sigma_u) = \frac{1}{N}\log|\Sigma_u| + \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{\mu_{N,T}}{N}\sum_{i\neq j} w_{ij}|\Sigma_{u,ij}|$$

$$-\frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\mathrm{tr}(S_u\Sigma_{u0}^{-1}) - \frac{\mu_{N,T}}{N}\sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}|, \qquad (\mathrm{A.1})$$

$$Q_2(\Lambda, \Sigma_u) = \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0$$

$$-\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0), \qquad (\mathrm{A.2})$$

$$Q_3(\Lambda, \Sigma_u) = \frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1})$$

$$-\frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) - \frac{1}{N}\log|\Sigma_u| - Q_2(\Lambda, \Sigma_u). \qquad (\mathrm{A.3})$$

Define the set,

$$\Xi_\delta = \{(\Lambda, \Sigma_u) : \ \delta^{-1} < \lambda_{\min}(N^{-1}\Lambda'\Lambda) \leq \lambda_{\max}(N^{-1}\Lambda'\Lambda) < \delta,$$
$$\delta^{-1} < \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) < \delta\}.$$

We first present a lemma that will be needed throughout the proof.

Lemma A.1. (i) $\max_{i,j\leq r} |\frac{1}{T}\sum_{t=1}^T f_{it}f_{jt} - Ef_{it}f_{jt}| = O_P(\sqrt{1/T})$.

(ii) $\max_{i,j\leq N} |\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - Eu_{it}u_{jt}| = O_P(\sqrt{(\log N)/T})$.

(iii) $\max_{i\leq r, j\leq N} |\frac{1}{T}\sum_{t=1}^T f_{it}u_{jt}| = O_P(\sqrt{(\log N)/T})$.

Proof. See Lemmas A.3 and B.1 in Fan et al. (2011). □

Lemma A.2. *Under Assumptions 3.2 and 3.3, for any* $\delta > 0$,

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} |Q_3(\Lambda, \Sigma_u)| = O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right).$$

Therefore we can write, with the $O(\cdot)$ *uniformly over* Ξ_δ,

$$\frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1})$$

$$= \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u| + Q_2(\Lambda, \Sigma_u)$$

$$+ O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right). \qquad (\mathrm{A.4})$$

Proof. First of all, note that $|\Lambda\Lambda' + \Sigma_u| = |\Sigma_u| \times |I_r + \Lambda'\Sigma_u^{-1}\Lambda|$, and $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \frac{1}{N}\log|I_r + \Lambda'\Sigma_u^{-1}\Lambda| = O\left(\frac{\log N}{N}\right)$, hence we have

$$\frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| = \frac{1}{N}\log|\Sigma_u| + O\left(\frac{\log N}{N}\right), \qquad (\mathrm{A.5})$$

where $O(\cdot)$ is uniform in Ξ_δ. Eq. (A.5) will be used later in the proof.

We now consider the term $N^{-1}\mathrm{tr}(S_y(\Lambda\Lambda'+\Sigma_u)^{-1})$. With the identification condition $\frac{1}{T}\sum_{t=1}^{T} f_tf_t' = I_r, \bar{f} = 0$, and $S_u = \frac{1}{T}\sum_{t=1}^{T} u_tu_t'$,

$$S_y = \frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y})(y_t-\bar{y})' = \Lambda_0\Lambda_0' + S_u + \Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_tu_t' + \left(\Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_tu_t'\right)' - \bar{u}\bar{u}'.$$

By the matrix inversion formula $(\Lambda\Lambda'+\Sigma_u)^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\Lambda(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}$,

$$\begin{aligned}\frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda'+\Sigma_u)^{-1}) &= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0) + \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1})\\ &\quad - A_1 + A_2 + A_3 - A_4 - A_5, \qquad (A.6)\end{aligned}$$

where $A_1 = N^{-1}\mathrm{tr}(\Lambda_0\Lambda_0'\Sigma_u^{-1}\Lambda(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1})$, $A_2 = \frac{1}{N}\mathrm{tr}(\frac{1}{T}\sum_{t=1}^{T}\Lambda_0 f_tu_t'(\Lambda\Lambda'+\Sigma_u)^{-1})$, $A_3 = \frac{1}{N}\mathrm{tr}(\frac{1}{T}\sum_{t=1}^{T}u_tf_t'\Lambda_0'(\Lambda\Lambda'+\Sigma_u)^{-1})$, and $A_4 = \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1})$. Term $A_5 = N^{-1}\mathrm{tr}(\bar{u}\bar{u}'(\Lambda\Lambda'+\Sigma_u)^{-1}) = O_P((\log N)/T)$ uniformly in the parameter space, and hence can be ignored. Let us look at terms A_1, A_2, A_3 and A_4 subsequently.

Note that $\lambda_{\max}(\Sigma_u)$ and $N\lambda_{\min}^{-1}(\Lambda'\Lambda)$ are both bounded from above uniformly in Ξ_δ, we have,

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}\lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}] \le \sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}\frac{\lambda_{\max}(\Sigma_u)}{\lambda_{\min}(\Lambda'\Lambda)} = O(N^{-1}), \qquad (A.7)$$

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}\lambda_{\max}[(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}] \le \sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}\lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}] = O(N^{-1}). \qquad (A.8)$$

In addition, $\|\Lambda\|_F = O(N^{1/2}N^{1/2})$, $\lambda_{\max}(\Sigma_u^{-1}) = O(1)$ uniformly in Ξ_δ, and $\|\Lambda_0\|_F = O(N^{1/2})$. Applying the matrix inversion formula yields

$$\begin{aligned}A_1 &= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)\\ &\quad - \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)\\ &= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0) + O\left(\frac{1}{N}\right), \qquad (A.9)\end{aligned}$$

where $O(\cdot)$ is uniform over $(\Lambda, \Sigma_u) \in \Xi_\delta$. In the second equality above we applied (A.7) and (A.8) and the following inequality:

$$\begin{aligned}&\frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)\\ &\le \frac{1}{N}\|\Lambda_0'\Sigma_u^{-1}\Lambda\|_F^2\lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}]\lambda_{\max}[(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}]\\ &= O(N^{-1}).\end{aligned}$$

By Lemma A.1(iii), and $\lambda_{\max}((\Lambda\Lambda'+\Sigma_u)^{-1}) \le \lambda_{\max}(\Sigma_u^{-1}) = O(1)$ uniformly in Ξ_δ,

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}|A_2| \le \frac{1}{N}\|\Lambda_0'(\Lambda\Lambda'+\Sigma_u)^{-1}\|_F\left\|\frac{1}{T}\sum_{t=1}^{T}f_tu_t'\right\|_F = O_P\left(\sqrt{\frac{\log N}{T}}\right). \qquad (A.10)$$

Similarly, $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}|A_3| = O_P(\sqrt{\frac{\log N}{T}})$. Again by the matrix inversion formula,

$$\begin{aligned}A_4 &= \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1})\\ &\quad - \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}).\end{aligned}$$

The second term on the right hand side is of smaller order (uniformly) than the first term, because it has an additional term $(I+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}$, whose maximum eigenvalue is $O(N^{-1})$ uniformly by (A.8). The first term of A_4 is bounded by (uniformly in Ξ_δ):

$$\begin{aligned}&\frac{c}{N}\|S_u\Sigma_u^{-1}\Lambda\|_F O(N^{-1})\|\Lambda'\Sigma_u^{-1}\|_F\\ &\le O(N^{-1})\lambda_{\max}(S_u) = O\left(\sqrt{\frac{\log N}{T}} + \frac{1}{N}\right),\end{aligned}$$

where we have, $\lambda_{\max}(S_u) \le \lambda_{\max}(\Sigma_u) + \|S_u - \Sigma_u\|_F = O_P(1 + N\sqrt{\log N/T})$. Hence $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta}|A_4| = O(T^{-1/2}(\log N)^{1/2} + N^{-1})$. Results (A.5) and (A.6) then yield

$$\begin{aligned}&\frac{1}{N}\log|\Lambda\Lambda'+\Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda'+\Sigma_u)^{-1})\\ &= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0) + \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u|\\ &\quad - \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)\\ &\quad + O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right)\\ &= \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u| + Q_2(\Lambda, \Sigma_u)\\ &\quad + O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right).\end{aligned}$$

□

Let

$$L_c(\Lambda, \Sigma_u) = L_1(\Lambda, \Sigma_u) - N^{-1}\log|\Sigma_{u0}| - N^{-1}\mathrm{tr}(S_u\Sigma_{u0}^{-1}) - N^{-1}\mu_{N,T}\sum_{i\neq j}w_{ij}|\Sigma_{u0,ij}|.$$

Then the minimizer of L_c is the same as that of L_1. This implies $L_c(\widehat{\Lambda}, \widehat{\Sigma}_u) \le L_c(\Lambda_0, \Sigma_{u0})$. Recall the definitions of $Q_1(\Sigma_u)$, $Q_2(\Lambda, \Sigma_u)$ and $Q_3(\Lambda, \Sigma_u)$ in the Appendix. Then

$$L_c(\Lambda, \Sigma_u) = Q_1(\Sigma_u) + Q_2(\Lambda, \Sigma_u) + Q_3(\Lambda, \Sigma_u).$$

Lemma A.3. *There is a nonnegative nonstochastic sequence* $0 \le d_T = O(N^{-1}\log N + T^{-1/2}(\log N)^{1/2})$ *such that* $Q_1(\widehat{\Sigma}_u) + Q_2(\widehat{\Lambda}, \widehat{\Sigma}_u) \le d_T$ *with probability one.*

Proof. We have $Q_2(\widehat{\Lambda}, \widehat{\Sigma}_u) \ge 0$. In addition, $Q_2(\Lambda_0, \Sigma_{u0}) = Q_1(\Sigma_{u0}) = 0$. Hence

$$\begin{aligned}Q_1(\widehat{\Sigma}_u) + Q_2(\widehat{\Lambda}, \widehat{\Sigma}_u) &= L_c(\widehat{\Lambda}, \widehat{\Sigma}_u) - Q_3(\widehat{\Lambda}, \widehat{\Sigma}_u)\\ &\le L_c(\Lambda_0, \Sigma_{u0}) - Q_3(\widehat{\Lambda}, \widehat{\Sigma}_u)\\ &= Q_3(\Lambda_0, \Sigma_{u0}) - Q_3(\widehat{\Lambda}, \widehat{\Sigma}_u).\end{aligned}$$

By the definition of $\Theta_\lambda \times \Gamma$, there is $\delta > 0$ such that $\Theta_\lambda \times \Gamma \subset \Xi_\delta$. The result then holds for $d_T = 2\sup|Q_3(\Lambda, \Sigma_u)|$ by Lemma A.2. □

Throughout the proofs, we note that the consistency of $\widehat{\Lambda}$ depends crucially on the consistency of the following quantity:

$$J = (\widehat{\Lambda} - \Lambda_0)'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}.$$

Lemma A.4. (i) $\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0 - (I_r - J)\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(I_r - J)' = o_P(N)$

(ii) *First order condition:* $\widehat{\Lambda}'(\widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Sigma}_u)^{-1}(S_y - \widehat{\Lambda}\widehat{\Lambda}' - \widehat{\Sigma}_u) = 0$.

Proof. Since $L_c(\widehat{\Lambda}, \widehat{\Sigma}_u) \le L_c(\Lambda_0, \Sigma_{u0})$, and $Q_1(\Sigma_{u0}) = Q_2(\Lambda_0, \Sigma_{u0}) = 0$, also Lemma A.2 proves that $Q_3 = (N^{-1}\log N + T^{-1/2}(\log N)^{1/2})$ uniformly over Ξ_δ, there is a nonnegative sequence $d_T = O_P(N^{-1}\log N + T^{-1/2}(\log N)^{1/2})$ such that $Q_1(\widehat{\Sigma}_u) + Q_2(\widehat{\Lambda}, \widehat{\Sigma}_u) \le d_T$. Hence $Q_1(\widehat{\Sigma}) + Q_2(\widehat{\Lambda}, \widehat{\Sigma}_u) \le d_T = o_P(1)$, given that $\log N = o(T)$.

On the other hand, Lemma B.2 implies there is a stochastic sequence $e_T = o_P(1)$ so that $Q_1(\widehat{\Sigma}_u) \ge e_T$ (The proofs of Lemmas B.1 and B.2 do not depend on Lemma A.4.) This then implies $0 \le Q_2(\widehat{\Sigma}_u, \widehat{\Lambda}) = o_P(1)$. On the other hand,

$$Q_2(\widehat{\Sigma}_u, \widehat{\Lambda}) = \frac{1}{N}\mathrm{tr}\Big[\Lambda_0'\widehat{\Sigma}_u^{-1}\Lambda_0 - \Lambda_0'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\Lambda_0\Big].$$

The matrix in the bracket is semi-positive definite. Hence

$$\frac{1}{N}\Lambda_0'\widehat{\Sigma}_u^{-1}\Lambda_0 - (I_r - J)\frac{1}{N}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(I_r - J)' = o_P(1). \qquad (A.11)$$

Finally, the desired result follows from Lemma B.4.

The first order condition in part (ii) is a straightforward calculation. □

Lemma A.5. (i) $\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(S_y - \widehat{\Lambda}\widehat{\Lambda}' - \widehat{\Sigma}_u) = 0$.

(ii) $(J - I_r)'(J - I_r) - I_r = O_P(N^{-1} + \sqrt{\log N/T})$.

Proof. (i) Using the matrix inverse formula, we have $\widehat{\Lambda}'(\widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Sigma}_u)^{-1} = (I_r + \widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}$. Thus part (i) follows from the first order condition in Lemma A.4.

(ii) Let $H = (\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}$. Part (i) can be equivalently written as $J + J' - J'J + K = 0$ where

$$\begin{aligned} K &= J'\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H + H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\frac{1}{T}\sum_{t=1}^{T} u_t f_t' J \\ &\quad - \frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H - H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\frac{1}{T}\sum_{t=1}^{T} u_t f_t' \\ &\quad - H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(S_u - \widehat{\Sigma}_u)\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H. \end{aligned}$$

Note that for $(\widehat{\Lambda}, \widehat{\Sigma}_u) \in \Xi_\delta$, $H = O_P(N^{-1})$, $J = O_P(1)$ for each element, $\|\widehat{\Sigma}_u^{-1}\| = O_P(1)$, $\|\widehat{\Lambda}\|_F = O_P(N^{1/2})$, hence

$$\begin{aligned} \left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H\right\|_F &\le O_P(N^{-1/2})\left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|_F \\ &= O_P\left(\sqrt{\frac{\log N}{T}}\right). \end{aligned}$$

Moreover, $\|S_u\|^2 = O_P(T^{-1}N^2\log N + 1)$, which implies $H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}S_u\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H = O_P(N^{-1} + T^{-1/2}(\log N)^{1/2})$. Also, $H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Sigma}_u\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H = H = O_P(N^{-1})$. Therefore the last term in K is $O_P(N^{-1} + \sqrt{\log N/T})$. Thus $K = O_P(N^{-1} + T^{-1/2}(\log N)^{1/2})$. It then implies (ii). □

Lemma A.6. *Given that* $\log N = o(T)$, *we have* $J = o_P(1)$.

Proof. By our assumption, both $\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}$ and $\Lambda_0'\Sigma_{u0}^{-1}\Lambda$ are diagonal. Moreover, the eigenvalues of $N^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}$ and $N^{-1}\Lambda_0'\Sigma_{u0}^{-1}\Lambda$ are bounded away from zero. Therefore by Lemma A.4(i) and Lemma A.5(ii), there are two diagonal matrices M_1 and M_2 whose eigenvalues are all bounded away from zero, such that

$$\begin{aligned} &(I_r - J)M_1(I_r - J)' = M_2 + o_P(1), \\ &(J - I_r)'(J - I_r) = I_r + o_P(1) \end{aligned} \qquad (A.12)$$

Applying Lemma A.1 of Bai and Li (2012a,b), we have $J = o_P(1)$ and $M_1 = M_2 + o_P(1)$. We also assumed $\widehat{\Lambda}$ and Λ_0 have the same column signs, as a part of identification condition. □

Appendix B. Proof of Theorem 3.1

Note that the theoretical results of our paper are only about the consistency of the estimated Λ and Σ_u, although some convergence rate of the covariance estimator is presented in Lemma B.2. The presented rate is not minimax optimal, and we would like to understand Lemma B.2 as a guarantee of the consistency.

Throughout, let (recall that $D_N = \sum_{i\neq j,(i,j)\in J_U} 1$).

$$\Delta = \widehat{\Sigma}_u^{-1} - \Sigma_{u0}^{-1}, \qquad K_T = \sum_{(i,j)\in J_L} |\Sigma_{u0,ij}|.$$

Lemma B.1. *For all large enough T and N,*

$$\begin{aligned} NQ_1(\widehat{\Sigma}_u) &\ge \frac{1}{2}\mu_{N,T}\min_{(i,j)\in J_L} w_{ij}\sum_{(i,j)\in J_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| + c\|\Delta\|_F^2 \\ &\quad - 2\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T \\ &\quad - \left(O_P\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N + D_N} + \mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij}\sqrt{D_N}\right)\|\Delta\|_F. \end{aligned}$$

Proof. Let $\Omega_0 = \Sigma_{u0}^{-1}$, $\widehat{\Omega} = \widehat{\Sigma}_u^{-1}$. For any Σ_u, let $\Omega = \Sigma_u^{-1}$. Define a function $f(t) = -\log|\Omega_0 + t\Delta| + \mathrm{tr}(S_u(\Omega_0 + t\Delta))$, $t \ge 0$. Then $-\log|\widehat{\Omega}| + \mathrm{tr}(S_u\widehat{\Omega}) = f(1)$; $-\log|\Omega_0| + \mathrm{tr}(S_u\Omega_0) = f(0)$; and

$$\begin{aligned} NQ_1(\widehat{\Sigma}_u) &= f(1) - f(0) + \mu_{N,T}\sum_{i\neq j} w_{ij}|\widehat{\Sigma}_{u,ij}| \\ &\quad - \mu_{N,T}\sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| \end{aligned} \qquad (B.1)$$

By the integral remainder Taylor expansion, $f(1) - f(0) = f'(0) + \int_0^1 (1-t)f''(t)dt$. We now calculate $f'(0)$ and $f''(t)$. We have, $f'(t) = \mathrm{tr}(S_u\Delta) - \mathrm{tr}((\Omega_0 + t\Delta)^{-1}\Delta)$, which implies,

$$\begin{aligned} f'(0) &= \mathrm{tr}((S_u - \Sigma_{u0})(\widehat{\Omega} - \Omega_0)) \\ &= \mathrm{tr}(\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega}(\Sigma_{u0} - \widehat{\Sigma}_u)) \\ &= \sum_{ij}(\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega})_{ij}(\Sigma_{u0} - \widehat{\Sigma}_u)_{ij}. \end{aligned}$$

Note that both $\|\Omega_0\|_1$ and $\|\widehat{\Omega}\|_1$ are bounded from above for $\Sigma_{u0}, \widehat{\Sigma}_u \in \Gamma$. By Lemma A.1(ii), $\max_{ij}|(\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega})_{ij}| \le \max_{ij}|(S_u - \Sigma_{u0})_{ij}|\|\Omega_0\|_1\|\widehat{\Omega}\|_1 = O_P(\sqrt{\log N/T})$. Therefore, $|f'(0)| = O_P(\sqrt{\log N/T})\sum_{ij}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}|$. In addition,

$$\begin{aligned} f''(t) &= \mathrm{tr}((\Omega_0 + t\Delta)^{-1}\Delta(\Omega_0 + t\Delta)^{-1}\Delta) \\ &= vec(\Delta)'(\Omega_0 + t\Delta)^{-1} \otimes (\Omega_0 + t\Delta)^{-1}vec(\Delta), \end{aligned}$$

where vec denotes the vectorization operator and $\otimes$ denotes the Kronecker product. Since both $(\widehat{\Lambda}, \widehat{\Sigma}_u)$ and (Λ_0, Σ_{u0}) are inside $\Theta_\lambda \times \Gamma$, $\sup_{0\le t\le 1}\lambda_{\max}(t\widehat{\Sigma}_u^{-1} + (1-t)\Sigma_{u0}^{-1})$ is bounded from above, which then implies $\inf_{0\le t\le 1}\lambda_{\min}[(\Omega_0 + t\Delta)^{-1}] = \inf_{0\le t\le 1}\lambda_{\max}^{-1}(t\widehat{\Sigma}_u^{-1} + (1-t)\Sigma_{u0}^{-1})$ is bounded below by a positive constant c. Hence $\inf_{0\leq t\leq 1} f''(t) \geq c\|\Delta\|_F^2$. From (B.1) and $f(1) - f(0) \geq -|f'(0)| + c\|\Delta\|_F^2$, we have

$$\begin{aligned}
NQ_1(\widehat{\Sigma}_u) &\geq \mu_{N,T}\sum_{i\neq j} w_{ij}|\widehat{\Sigma}_{u,ij}| - \mu_{N,T}\sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| \\
&\quad + c\|\Delta\|_F^2 - O_P\left(\sqrt{\frac{\log N}{T}}\right)\sum_{ij}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&= \mu_{N,T}\sum_{(i,j)\in J_L} w_{ij}|\widehat{\Sigma}_{u,ij}| + \mu_{N,T}\sum_{i\neq j,(i,j)\in J_U} w_{ij}|\widehat{\Sigma}_{u,ij}| \\
&\quad - \mu_{N,T}\sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| + c\|\Delta\|_F^2 - O_P\left(\sqrt{\frac{\log N}{T}}\right) \\
&\quad \times \sum_{\Sigma_{u0,ij}\in J_U}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \; - O_P\left(\sqrt{\frac{\log N}{T}}\right) \\
&\quad \times \sum_{(i,j)\in J_L}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}|.
\end{aligned}$$

Since $|\widehat{\Sigma}_{u,ij}| \geq |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| - |\Sigma_{u0,ij}|$, and $\sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| = \sum_{i\neq j,(i,j)\in J_U} w_{ij}|\Sigma_{u0,ij}| + \sum_{(i,j)\in J_L} w_{ij}|\Sigma_{u0,ij}|$. It follows that

$$\begin{aligned}
NQ_1(\widehat{\Sigma}_u) &\geq \mu_{N,T}\sum_{(i,j)\in J_L} w_{ij}|\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| \\
&\quad - O_P\left(\sqrt{\frac{\log N}{T}}\right)\sum_{(i,j)\in J_L}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| + c\|\Delta\|_F^2 \\
&\quad - \mu_{N,T}\sum_{(i,j)\in J_L} w_{ij}|\Sigma_{u0,ij}| - O_P\left(\sqrt{\frac{\log N}{T}}\right)\sum_{\Sigma_{u0,ij}\in J_U}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&\quad - \mu_{N,T}\sum_{i\neq j,(i,j)\in J_U} w_{ij}[|\Sigma_{u0,ij}| - |\widehat{\Sigma}_{u,ij}|] - \mu_{N,T}\sum_{(i,j)\in J_L} w_{ij}|\Sigma_{u0,ij}| \\
&\geq \left(\mu_{N,T}\min_{(i,j)\in J_L} w_{ij} - O_P\left(\sqrt{\frac{\log N}{T}}\right)\right)\sum_{(i,j)\in J_L}|\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| \\
&\quad + c\|\Delta\|_F^2 - 2\mu_{N,T}\sum_{(i,j)\in J_L} w_{ij}|\Sigma_{u0,ij}| - O_P\left(\sqrt{\frac{\log N}{T}}\right) \\
&\quad \times \sum_{\Sigma_{u0,ij}\in J_U}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| - \mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij} \\
&\quad \times \sum_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&\geq \frac{1}{2}\mu_{N,T}\min_{(i,j)\in J_L} w_{ij}\sum_{(i,j)\in J_L}|\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| + c\|\Delta\|_F^2 \\
&\quad - 2\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T - O_P\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N + D_N}\|\Delta\|_F \\
&\quad - \mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij}\|\Delta\|_F\sqrt{D_N},
\end{aligned}$$

which implies the desired result. □

The following lemma presents a non-optimal rate of convergence, which is for the consistency only.

Lemma B.2.

$$\begin{aligned}
&\frac{1}{N}\|\Sigma_{u0} - \widehat{\Sigma}_u\|_F^2 \\
&= O_P\left(\frac{1}{N}\left(\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T + \log N + \mu_{N,T}^2\max_{i\neq j,(i,j)\in J_U} w_{ij}^2 D_N\right)\right) \\
&\quad + O_P\left(\frac{D_N\log N}{NT} + \sqrt{\frac{\log N}{T}}\right).
\end{aligned}$$

Proof. Lemma B.1 implies

$$\begin{aligned}
NQ_1(\widehat{\Sigma}_u) &\geq c\|\Delta\|_F^2 - 2\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T \\
&\quad - \left(O_P\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N + D_N} + \mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij}\sqrt{D_N}\right)\|\Delta\|_F.
\end{aligned}$$

Lemma A.3 gives $NQ_1(\widehat{\Sigma}_u) \leq O_P(\log N + N\sqrt{\log N/T})$. Hence we have

$$\begin{aligned}
\|\Delta\|_F^2 &= O_P\left(\left(\sqrt{\frac{(N + D_N)\log N}{T}} + \mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij}\sqrt{D_N}\right)^2\right) \\
&\quad + O_P\left(\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}\right) \\
&= O_P\left(\frac{(N + D_N)\log N}{T} + \mu_{N,T}^2\max_{i\neq j,(i,j)\in J_U} w_{ij}^2 D_N\right. \\
&\quad \left. + \mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}\right) \\
&= O_P\left(\frac{D_N\log N}{T} + \mu_{N,T}^2\max_{i\neq j,(i,j)\in J_U} w_{ij}^2 D_N\right. \\
&\quad \left. + \mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}\right).
\end{aligned}$$

Note that $\Sigma_{u0} - \widehat{\Sigma}_u = \widehat{\Sigma}_u\Delta\Sigma_{u0}$. Hence the desired result follows from $\|\widehat{\Sigma}_u\| < M$ almost surely and $\|\Sigma_{u0}\| < M$.

In addition, it also implies

$$\begin{aligned}
Q_1(\widehat{\Sigma}_u) &\geq c\frac{1}{N}\|\Delta\|_F^2 - \frac{2\mu_{N,T}}{N}\max_{J_L} w_{ij}K_T \\
&\quad - \left(\frac{1}{N}O_P\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N + D_N} + \frac{\mu_{N,T}}{N}\max_{i\neq j,(i,j)\in J_U} w_{ij}\sqrt{D_N}\right)\| \\
&\quad \times \Delta\|_F.
\end{aligned}$$

Combining with $NQ_1(\widehat{\Sigma}_u) \leq O_P(\log N + N\sqrt{\log N/T})$, we have

$$\begin{aligned}
&\frac{1}{N}\|\Delta\|_F^2 \\
&= O_P\left(\frac{1}{N}\left(\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T + \log N + \mu_{N,T}^2\max_{i\neq j,(i,j)\in J_U} w_{ij}^2 D_N\right)\right) \\
&\quad + O_P\left(\frac{D_N\log N}{NT} + \sqrt{\frac{\log N}{T}}\right)
\end{aligned}$$

which is $o_P(1)$. The second term $\frac{2\mu_{N,T}}{N}\max_{J_L} w_{ij}K_T = o(1)$, and the third term on the right hand side is also straightforward to be verified as $o_P(1)$.

Finally, recall that $\Delta = \widehat{\Sigma}_u^{-1} - \Sigma_{u0}^{-1}$,

$$\begin{aligned}
\frac{1}{N}\|\Sigma_{u0} - \widehat{\Sigma}_u\|_F^2 &= \frac{1}{N}\|\Sigma_{u0}(\Sigma_{u0}^{-1} - \widehat{\Sigma}_u^{-1})\widehat{\Sigma}_u\|_F^2 \\
&\leq \|\Sigma_{u0}\|\|\widehat{\Sigma}_u\|\frac{1}{N}\|\Delta\|_F^2.
\end{aligned}$$

Hence $\frac{1}{N}\|\Sigma_{u0} - \widehat{\Sigma}_u\|_F^2$ has the same convergence rate as that of $\frac{1}{N}\|\Delta\|_F^2$ due to $\|\Sigma_{u0}\| < M$ and $\|\widehat{\Sigma}_u\| < M$. □

Lemma B.3. $N^{-1}\sum_{(i,j)\in J_L}|\widehat{\Sigma}_{u,ij}-\Sigma_{u0,ij}|=o_P(1)$.

Proof. Lemma B.1 implies

$$\begin{aligned}&\frac{1}{2}\mu_{N,T}\min_{(i,j)\in J_L}w_{ij}\sum_{(i,j)\in J_L}|\widehat{\Sigma}_{u,ij}-\Sigma_{u0,ij}|\\&\quad\le NQ_1(\widehat{\Sigma}_u)+2\mu_{N,T}\max_{(i,j)\in J_L}w_{ij}K_T\\&\quad+\left(O_P\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N+D_N}+\mu_{N,T}\max_{i\neq j,(i,j)\in J_U}w_{ij}\sqrt{D_N}\right)\\&\quad\times\|\Delta\|_F.\end{aligned}\tag{B.2}$$

We have $NQ_1(\widehat{\Sigma}_u)\le O_P(\log N+N\sqrt{\log N/T})$. By Lemma B.2,

$$\begin{aligned}\|\Delta\|_F&=O_P\left(\sqrt{\frac{D_N\log N}{T}}+\mu_{N,T}\max_{i\neq j,(i,j)\in J_U}w_{ij}\sqrt{D_N}\right)\\&\quad+O_P\left(\sqrt{\mu_{N,T}\max_{(i,j)\in J_L}w_{ij}K_T}+\sqrt{\log N}+\sqrt{N}\left(\frac{\log N}{T}\right)^{1/4}\right)\end{aligned}$$

which implies the desired result under Assumption 3.4. □

Lemma B.4. $N^{-1}\Lambda_0'(\widehat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1})\Lambda_0=o_P(1)$.

Proof. Let $\Delta_1=\widehat{\Sigma}_u-\Sigma_{u0}$, $\Xi=\Lambda_0'\Sigma_{u0}^{-1}=(\xi_1,\ldots,\xi_N)$, and $\widehat{V}=\widehat{\Sigma}_u^{-1}\Lambda_0$. Since the ℓ_1 norms of $\widehat{\Sigma}_u^{-1}$ and Σ_{u0}^{-1} are bounded away from infinity, we have, $\sup_{i\le N}\|\widehat{V}_i\|=O_P(1)$ and $\sup_{i\le N}\|\xi_i\|=O(1)$. Then

$$\begin{aligned}&\frac{1}{N}\Lambda_0'(\Sigma_{u0}^{-1}-\widehat{\Sigma}_u^{-1})\Lambda_0\\&\quad=\frac{1}{N}\Xi\Delta_1\widehat{V}=\frac{1}{N}\sum_{(i,j)\in J_L}\xi_i\widehat{V}_j'\Delta_{1,ij}+\frac{1}{N}\sum_{\Sigma_{u0,ij}\in J_U}\xi_i\widehat{V}_j'\Delta_{1,ij}\\&\quad\le O_P\left(\frac{1}{N}\right)\sum_{(i,j)\in J_L}|\Delta_{1,ij}|+O_P\left(\frac{1}{N}\right)\sum_{\Sigma_{u0,ij}\in J_U}|\Delta_{1,ij}|.\end{aligned}$$

The first term on the right hand side is $o_P(1)$ by Lemma B.3, and the second is bounded by $N^{-1}\|\widehat{\Sigma}_u-\Sigma_{u0}\|_F\sqrt{N+D_N}$ (using Cauchy–Schwarz inequality). By Lemma B.2,

$$\begin{aligned}&\frac{N+D_N}{N^2}\|\Sigma_{u0}-\widehat{\Sigma}_u\|_F^2=O_P\left(\frac{N+D_N}{N^2}\left(\mu_{N,T}\max_{(i,j)\in J_L}w_{ij}K_T\right.\right.\\&\quad\left.\left.+\log N+\mu_{N,T}^2\max_{i\neq j,(i,j)\in J_U}w_{ij}^2D_N\right)\right)\\&\quad+\frac{N+D_N}{N^2}O_P\left(\frac{D_N\log N}{T}+N\sqrt{\frac{\log N}{T}}\right)\end{aligned}$$

which is also $o_P(1)$ and Assumption 3.4. Hence the result follows. □

Proof of Theorem 3.1. $N^{-1}\|\widehat{\Sigma}_u-\Sigma_{u0}\|_F^2=o_P(1)$ follows from Lemma B.2 and Assumption 3.4. On the other hand, Eq. (A.11) also implies

$$\frac{1}{N}(\widehat{\Lambda}-\Lambda_0)'\widehat{\Sigma}_u^{-1}(\widehat{\Lambda}-\Lambda_0)-J\frac{1}{N}H^{-1}J'=o_P(1).$$

where $H^{-1}=\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}=O_P(N)$. By Lemma A.6, $N^{-1}JH^{-1}J'=o_P(1)$. Hence $N^{-1}(\widehat{\Lambda}-\Lambda_0)'\widehat{\Sigma}_u^{-1}(\widehat{\Lambda}-\Lambda_0)=o_P(1)$, which implies the consistency $N^{-1}\|\widehat{\Lambda}-\Lambda_0\|^2=o_P(1)$ because the eigenvalues of $\widehat{\Sigma}_u^{-1}$ are bounded away from zero. □

To prove the consistency of $\widehat{f_t}$, we note that

$$\widehat{f_t}-f_t=-J'f_t+(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(u_t-\bar{u}).$$

Since $J=o_P(1)$ by Lemma A.6, and $\bar{u}$ is of smaller order than u_t for each fixed t. Hence $\widehat{f_t}-f_t=O_P(N^{-1})\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}u_t+o_P(1)$. Moreover, since $\|\widehat{\Sigma}_u^{-1}\|$ and $\|\widehat{\Sigma}_u\|$ are both $O_P(1)$ and $\|\widehat{\Lambda}\|_F=O_P(N^{1/2})$ by the restriction of the parameter space $\Theta_\lambda\times\Gamma$, we have

$$\begin{aligned}&N^{-1}\|(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}-\Lambda_0\Sigma_{u0}^{-1})u_t\|_F\\&\quad\le N^{-1}\|(\widehat{\Lambda}'-\Lambda')\widehat{\Sigma}_u^{-1}u_t\|_F+N^{-1}\|\Lambda(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})u_t\|_F.\end{aligned}$$

The first term is bounded by $N^{-1}\|\widehat{\Lambda}-\Lambda\|_F\|u_t\|O_P(1)=O_P(N^{-1/2+1/2})N^{-1/2}\|\widehat{\Lambda}-\Lambda\|_F=o_P(1)$. On the other hand, let $\hat{\xi}_i$ be the i th column of $\Lambda'\widehat{\Sigma}_u^{-1}$, and e_{jt} be the jth entry of $\Sigma_u^{-1}u_t$. We have $\max_i\|\hat{\xi}_i\|=O_P(1)$ and $\max_j|e_{jt}|=O_P(\log N)$.

$$\begin{aligned}N^{-1}\Lambda'(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})u_t&=N^{-1}\sum_{ij}\hat{\xi}_ie_{jt}(\Sigma_{ij}-\hat{\Sigma}_{ij})\\&\le O_P(\log N)N^{-1}\sum_{J_L}|\Sigma_{ij}-\hat{\Sigma}_{ij}|+O_P(\log N)N^{-1}\sum_{J_U}|\Sigma_{ij}-\hat{\Sigma}_{ij}|\\&\le O_P(\log N)N^{-1}\sum_{J_L}|\Sigma_{ij}-\hat{\Sigma}_{ij}|\\&\quad+O_P(\log N)N^{-1}\|\Sigma_u-\widehat{\Sigma}_u\|_F\sqrt{N+D_N}.\end{aligned}$$

It follows from (B.2) that the first term is $o_P(1)$. In addition, it follows from Lemma B.2 that the second term is also $o_P(1)$. Therefore,

$$\begin{aligned}\widehat{f_t}-f_t&=O_P(N^{-1})\Lambda_0'\Sigma_{u0}^{-1}u_t+o_P(1)\\&=O_P(N^{-1})\sum_{i=1}^N\xi_iu_{it}+o_P(1)=O_P(N^{-1/2})+o_P(1)\\&=o_P(1).\end{aligned}$$

Appendix C. Proof of Theorem 3.2

Let $\widehat{\Sigma}_{u,ij}^*$ be the PCA estimator of $\Sigma_{u0,ij}$. Write

$$Re=\max_{i\le N,j\le N}|\widehat{\Sigma}_{u,ij}^*-\Sigma_{u0,ij}|=O_P(\omega_T)$$

where $\omega_T=\sqrt{\frac{\log N}{T}}+\frac{1}{\sqrt{N}}$.

We now verify Assumption 3.4 for the adaptive lasso.

Lemma C.1. *For adaptive lasso,*

(i) $\min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}|\max_{i\neq j,(i,j)\in J_U}w_{ij}=O_P(1)$.

(ii) $\delta_T\max_{(i,j)\in J_L}w_{ij}=O_P(1)$.

Proof. By the assumption that $\min_{(i,j)\in J_U}|\Sigma_{u0,ij}|\gg\omega_T$, we have result (i). For any $(i,j)\in J_L$, the following inequality holds: $\delta_T^{-1}\le w_{ij}^{-1}\le|\Sigma_{u0,ij}|+|\Sigma_{u0,ij}-\widehat{\Sigma}_{u,ij}^*|+\delta_T$, which then implies result (ii), due to the assumptions that $\delta_T=o(\omega_T)$, and $|\Sigma_{u0,ij}|=O(\omega_T)$ for $(i,j)\in J_L$. □

Proof of Assumption 3.4 for adaptive lasso

Note that

$$\eta_T=\frac{\max_{i\neq j,(i,j)\in J_U}(\delta_T+|\widehat{\Sigma}_{ij}^*|)^{-1}}{\min_{(i,j)\in J_L}(\delta_T+|\widehat{\Sigma}_{ij}^*|)^{-1}}\le\frac{\delta_T+\max_{(i,j)\in J_L}|\Sigma_{0,ij}|+Re}{\min_{i\neq j,(i,j)\in J_U}|\Sigma_{0,ij}|-Re}=o_P(1),$$

$$\beta_T=\frac{\max_{(i,j)\in J_L}w_{ij}}{\min_{(i,j)\in J_L}w_{ij}}=\frac{\max_{(i,j)\in J_L}(\delta_T+|\widehat{\Sigma}_{ij}^*|)}{\min_{(i,j)\in J_L}(\delta_T+|\widehat{\Sigma}_{ij}^*|)}\le\frac{\delta_T+Re+\max_{(i,j)\in J_L}|\Sigma_{0,ij}|}{\delta_T}.$$

By the assumption that $D_N = O(N)$,

$$\zeta = \min\left\{\sqrt{\frac{T}{\log N}\frac{N}{D_N}}, \left(\frac{T}{\log N}\right)^{1/4}\sqrt{\frac{N}{D_N}}, \frac{N}{\sqrt{D_N\log N}}\right\}$$
$$\gg \min\left\{\left(\frac{T}{\log N}\right)^{1/4}, \sqrt{\frac{N}{\log N}}\right\}.$$

Hence $\eta_T = O_P(\zeta)$. Moreover,

$$\beta_T\frac{1}{N}\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}| \le \frac{1}{N}\sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|(1+\delta_T^{-1}(Re+\max_{(i,j)\in J_L}|\Sigma_{0,ij}|)).$$

This together with the lower bound assumption on δ_T yields Assumption 3.4(i).

For part (ii), note that $\eta_T = o_P(1)$ implies that with probability approaching one,

$$\min\left\{N, \frac{N^2}{D_N}, \frac{N^2}{D_N}\eta_T^{-2}\right\} = N,$$
$$\min\left\{\frac{N}{D_N}, \sqrt{\frac{N}{D_N}}, \frac{N}{D_N}\eta_T^{-1}\right\} = \sqrt{\frac{N}{D_N}}.$$

By Lemma C.1(ii), (recall that $K_T = \sum_{(i,j)\in J_L}|\Sigma_{u0,ij}|$) and the lower bound $\delta_T \gg \omega_T K_T/N$, $\mu_{N,T}\max_{(i,j)\in J_L} w_{ij}K_T = O_P(\mu_{N,T}\delta_T^{-1}K_T) = o_P(\omega_T\delta_T^{-1}K_T) = o_P(N)$.

By the assumptions that $D_N = O(N)$ and $\min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}| \gg \omega_T$, we have

$$\mu_{N,T}\max_{i\neq j,(i,j)\in J_U} w_{ij} = \mu_{N,T}O_P(\min_{i\neq j,(i,j)\in J_U}|\Sigma_{u0,ij}|)^{-1}$$
$$= O_P(\mu_{N,T}\omega_T^{-1}) = o_P(N^{-(1-a)}) = o_P(\sqrt{N/D_N}),$$

due to the upper bound on $\mu_{N,T} = o(\omega_T)$. Finally, with probability approaching one,

$$\mu_{N,T}\min_{(i,j)\in J_L} w_{ij} \ge \frac{\mu_{N,T}}{2\omega_T+Re} \ge \frac{\mu_{N,T}}{3\omega_T} \gg \sqrt{\log N/T} + (\log N)/N.$$

Proof of Assumption 3.4 for SCAD

Since $\mu_{N,T}/\min_{i\neq j,(i,j)\in J_U}|R_{ij}| = o_P(1)$ and $\max_{(i,j)\in J_L}|R_{ij}| = O_P(\mu_{N,T})$, it is easy to verify that with probability approaching one, $\max_{i\neq j,(i,j)\in J_U} w_{ij} = 0$, $\min_{(i,j)\in J_L} w_{ij} = \max_{(i,j)\in J_L} w_{ij} = 1$. Hence $\eta_T = 0$ and $\beta_T = 1$. This immediately implies the desired result.

Appendix D. Proof of Theorem 4.1

Recall that $\widehat{\Lambda}_{k+1} = AM^{-1}$, where $M = \widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k + I_r - \widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k$,

$$A = S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k, \quad \widehat{\Sigma}_{y,k} = \widehat{\Lambda}_k\widehat{\Lambda}_k' + \widehat{\Sigma}_{u,k}.$$

Lemma D.1. (i) $\|S_y - \widehat{\Sigma}_{y,k}\|_\infty = o_P(\mu_{N,T})$,

(ii) $\lambda_{\min}^{-1}(\widehat{\Lambda}_k'\widehat{\Lambda}_k) = O_P(N^{-1})$,

(iii) $\|(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\| = O_P(N^{-1})$,

(iv) $\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k = \widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}$. *Hence* $\|\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k\|_\infty \le \|\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k\|_\infty O_P(N^{-1})$.

(v) $\|M^{-1} - I_r\|_\infty = o_P(\mu_{N,T})$,

(vi) $\|\widehat{\Sigma}_{y,k}^{-1}\|_1 = O_P(1)$.

Proof. (i) On one hand, note that for the sample covariance S_y, and $\Sigma_y = \Lambda_0\Lambda_0' + \Sigma_{u0}$, under our conditions, $\|S_y - \Sigma_{y,0}\|_\infty = O_P(\sqrt{\frac{\log N}{T}})$ (e.g., Fan et al., 2008). On the other hand,

$$\|\widehat{\Sigma}_{y,k} - \Sigma_{y,0}\|_\infty \le \|\widehat{\Lambda}_k\widehat{\Lambda}_k' - \Lambda_0\Lambda_0'\|_\infty + \|\widehat{\Sigma}_{u,k} - \Sigma_{u,0}\|_\infty$$
$$= o_P(\mu_{N,T}).$$

Hence the result follows from the triangular inequality and that $\sqrt{\frac{\log N}{T}} = o(\mu_{N,T})$.

(ii) Note that

$$\left\|\frac{1}{N}\widehat{\Lambda}_k'\widehat{\Lambda}_k - \frac{1}{N}\Lambda_0'\Lambda_0\right\| = O_P(\|\widehat{\Lambda}_k - \Lambda_0\|_\infty) = o_P(\mu_{N,T}).$$

Since $\lambda_{\min}(\frac{\Lambda_0'\Lambda_0}{N}) > \delta^{-1}$,

$$\lambda_{\min}^{-1}(\widehat{\Lambda}_k'\widehat{\Lambda}_k) \le \frac{1}{N}\left(\lambda_{\min}\left(\frac{\Lambda_0'\Lambda_0}{N}\right) - \left\|\frac{1}{N}\widehat{\Lambda}_k'\widehat{\Lambda}_k - \frac{1}{N}\Lambda_0'\Lambda_0\right\|\right)^{-1}$$
$$= \frac{1}{N}(\delta + o_P(\mu_{N,T})).$$

(iii) It follows from $\|\widehat{\Sigma}_{u,k}\|_1 < M$ that $\lambda_{\min}(\widehat{\Sigma}_{u,k}^{-1}) = \lambda_{\max}^{-1}(\widehat{\Sigma}_{u,k}) > M^{-1}$. Hence

$$\|(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\| = \lambda_{\min}^{-1}(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)$$
$$\le \lambda_{\min}^{-1}(\widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)$$
$$\le \lambda_{\min}^{-1}(\widehat{\Sigma}_{u,k}^{-1})\lambda_{\min}^{-1}(\widehat{\Lambda}_k'\widehat{\Lambda}_k) = O_P(N^{-1}).$$

(iv) The desired equality follows directly from the matrix inversion formula:

$$\widehat{\Sigma}_{y,k}^{-1} = \widehat{\Sigma}_{u,k}^{-1} - \widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}. \quad \text{(D.1)}$$

(v) We first bound $\|M - I_r\|_\infty$. Note that $M - I_r = \widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}(S_y - \widehat{\Sigma}_{y,k})\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k$. So Lemma D.1 implies

$$\|M - I_r\|_\infty \le N^2\|S_y - \widehat{\Sigma}_{y,k}\|_\infty\|\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k\|_\infty^2$$
$$\le \|S_y - \widehat{\Sigma}_{y,k}\|_\infty\|\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k\|_\infty^2$$
$$\le \|S_y - \widehat{\Sigma}_{y,k}\|_\infty\|\widehat{\Sigma}_{u,k}^{-1}\|_1^2\|\widehat{\Lambda}_k\|_\infty^2 = o_P(\mu_{N,T}).$$

Hence $\lambda_{\min}(M) \ge 1 - o_P(\mu_{N,T})$, yielding $\|M^{-1}\|_\infty \le \|M^{-1}\| = \lambda_{\min}^{-1}(M) = O_P(1)$. So

$$\|M^{-1} - I_r\|_\infty \le r\|M - I_r\|_\infty\|M^{-1}\|_\infty = o_P(\mu_{N,T}).$$

(vi) Note that

$$\|\widehat{\Lambda}_k(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\widehat{\Lambda}_k'\|_1$$
$$\le \|\widehat{\Lambda}_k\|_\infty^2\|(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\|_\infty Nr^2$$
$$= O_P(1).$$

Hence by (D.1) and that $\|\widehat{\Sigma}_{u,k}^{-1}\|_1 = O_P(1)$,

$$\|\widehat{\Sigma}_{y,k}^{-1}\|_1 \le \|\widehat{\Sigma}_{u,k}^{-1}\|_1 + \|\widehat{\Sigma}_{u,k}^{-1}\|_1$$
$$\times\|\widehat{\Lambda}_k(I_r + \widehat{\Lambda}_k'\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k)^{-1}\widehat{\Lambda}_k'\|_1\|\widehat{\Sigma}_{u,k}^{-1}\|_1$$
$$= O_P(1). \quad \square \quad \text{(D.2)}$$

D.1. Proving $\|\widehat{\Lambda}_{k+1} - \Lambda_0\|_\infty = o_P(\mu_{N,T})$

Recall that $A = S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k$ and $\widehat{\Lambda}_{k+1} = AM^{-1}$.

Step 1: proving $\|\widehat{\Lambda}_{k+1} - A\|_\infty = o_P(\mu_{N,T})$

We first show $\|A\|_\infty = O_P(1)$. By Lemma D.1,

$$\|A\|_\infty \le \|S_y\|_\infty\|\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k\|_\infty N \le \|S_y\|_\infty\|\widehat{\Sigma}_{u,k}^{-1}\widehat{\Lambda}_k\|_\infty O_P(1) = O_P(1).$$

Hence still by Lemma D.1, we have

$$\begin{aligned}\|\widehat{\Lambda}_{k+1}-A\|_\infty &= \|A(M^{-1}-I_r)\|_\infty\\ &\le \|A\|_\infty\|M^{-1}-I_r\|_\infty r = o_P(\mu_{N,T}).\end{aligned}$$

Step 2: proving $\|\widehat{\Lambda}_k-A\|_\infty = o_P(\mu_{N,T})$
By Lemma D.1(i)(iv),

$$\begin{aligned}\|\widehat{\Lambda}_k-A\|_\infty &= \|\widehat{\Lambda}_k-A\|_\infty = \|(S_y-\widehat{\Sigma}_{y,k})\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k\|_\infty\\ &\le \|S_y-\widehat{\Sigma}_{y,k}\|_\infty\|\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k\|_\infty N = o_P(\mu_{N,T}).\end{aligned}$$

The desired result then follows from the triangular inequality and that $\|\widehat{\Lambda}_k-\Lambda_0\|_\infty = o_P(\mu_{N,T})$.

D.2. Proving $\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_\infty = o_P(\mu_{N,T})$, $\quad\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1 = o_P(\mu_{N,T}m_N)$

To simplify the technicality, we consider the case

$$\max_{i,j\in J_L}|\Sigma_{u0,ij}| = 0, \qquad \min_{i,j\in J_U}|\Sigma_{u0,ij}| \gg \mu_{N,T}.$$

Recall that $\widehat{\Sigma}_{u,k+1}$ is obtained by applying soft-thresholding on $B = \widehat{\Sigma}_{u,k} - tG$, where $G := \widehat{\Sigma}_{y,k}^{-1} - \widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}$. We also prove only for the scad weight, since it is asymptotically unbiased (in the sense to be described below). The proof for the adaptive lasso will be quite similar except that t has to be chosen as a decreasing sequence $t_{N,T}$.

Step 1: proving $\|G\|_\infty = o_P(\mu_{N,T})$
By Lemma D.1(vi), $\|\widehat{\Sigma}_{y,k}^{-1}\|_1 = O_P(1)$. Hence

$$\begin{aligned}\|G\|_\infty &= \|\widehat{\Sigma}_{y,k}^{-1}(S_y-\widehat{\Sigma}_{y,k})\widehat{\Sigma}_{y,k}^{-1}\|_\infty\\ &\le \|\widehat{\Sigma}_{y,k}^{-1}\|_1^2\|S_y-\widehat{\Sigma}_{y,k}\|_\infty = o_P(\mu_{N,T}).\end{aligned}$$

Step 2: proving $\|B-\Sigma_{u0}\|_\infty = o_P(\mu_{N,T})$
Note that $\|B-\Sigma_{u0}\|_\infty \le \|\widehat{\Sigma}_{u,k}-\Sigma_{u0}\|_\infty + t\|G\|_\infty = o_P(\mu_{N,T})$.

Step 3: proving $\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_\infty = o_P(\mu_{N,T})$
Consider three cases:

Case 1: $i=j$: $(\widehat{\Sigma}_{u,k+1})_{ij} = B_{ij}$
In this case, $|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}| \le \|B-\Sigma_{u0}\|_\infty = o_P(\mu_{N,T})$.

Case 2: $i\neq j$: $(i,j)\in J_L$.
In this case, $\Sigma_{0,ij}=0$, and $\max_{(i,j)\in J_L}|B_{ij}| = o_P(\mu_{N,T})$. For scad, note that with probability approaching one, $w_{ij} = 1\ \forall(i,j)\in J_L$, hence $|B_{ij}| < \mu_{N,T}w_{ij}t$ for all $(i,j)\in J_L$ with probability approaching one. This implies,

$$P((\widehat{\Sigma}_{u,k+1})_{ij}=0, \quad \forall(i,j)\in J_L)\to 1,$$

$$P(|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}|=0, \quad \forall(i,j)\in J_L)\to 1. \tag{D.3}$$

Case 3: $i\neq j$: $(i,j)\in J_U$
Note that with probability approaching one, the scad weights satisfy: $\max_{i\neq j,(i,j)\in J_U} w_{ij} = 0$. Hence $P(|B_{ij}| > \omega_{ij}t\mu_{N,T}, \forall(i,j)\in J_U)\to 1$, and

$$P((\widehat{\Sigma}_{u,k+1})_{ij} = B_{ij}, \quad \forall(i,j)\in J_U, i\neq j)\to 1.$$

We see that the soft thresholding with scad weights is asymptotically unbiased.

This implies, with probability approaching one, $\max_{i\neq j,(i,j)\in J_U}|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}| = \max_{i\neq j,(i,j)\in J_U}|B_{ij}-\Sigma_{u0,ij}|$, yielding

$$\max_{i\neq j,(i,j)\in J_U}|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}| = o_P(\mu_{N,T}).$$

Summarizing steps 1–3, we conclude

$$\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_\infty = o_P(\mu_{N,T}) = o_P(\mu_{N,T}).$$

Step 4: proving $\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1 = o_P(\mu_{N,T}m_N)$.
We have:

$$\begin{aligned}\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1 \le{}& \max_{i\le p}\sum_{j:(i,j)\in J_L}|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}|\\ &+\max_{i\le p}\sum_{j:(i,j)\in J_U}|(\widehat{\Sigma}_{u,k+1})_{ij}-\Sigma_{u0,ij}|.\end{aligned}$$

By (D.3), the first term on the right hand side equals zero with probability approaching one. The second term on the right hand side is bounded by (recall that $m_N = \max_{i\le N}\sum_{j=1}^N 1\{\Sigma_{u0,ij}\neq 0\}$)

$$\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_\infty m_N = o_P(\mu_{N,T}m_N).$$

Thus $\|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1 = o_P(\mu_{N,T}m_N)$.

D.3. Bounding $\|\widehat{\Sigma}_{u,k+1}^{-1}\|_1$, $\|\widehat{\Sigma}_{u,k+1}\|_1$ and eigenvalues of $N^{-1}\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1}$

First of all,

$$\begin{aligned}\|\widehat{\Sigma}_{u,k+1}\|_1 &\le \|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1 + \|\Sigma_{u0}\|_1\\ &= \|\Sigma_{u0}\|_1 + o_P(\mu_{N,T}m_N).\end{aligned}$$

Secondly,

$$\begin{aligned}&\|\widehat{\Sigma}_{u,k+1}^{-1}-\Sigma_{u0}^{-1}\|_1 = \|\widehat{\Sigma}_{u,k+1}^{-1}(\widehat{\Sigma}_{u,k+1}-\Sigma_{u0})\Sigma_{u0}^{-1}\|_1\\ &\quad\le \|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1\|\Sigma_{u0}^{-1}\|_1^2 + \|\widehat{\Sigma}_{u,k+1}-\Sigma_{u0}\|_1\\ &\quad\quad\times\|\widehat{\Sigma}_{u,k+1}^{-1}-\Sigma_{u0}^{-1}\|_1\|\Sigma_{u0}^{-1}\|_1,\end{aligned}$$

which implies $(1-o_P(\mu_{N,T}m_N))\|\widehat{\Sigma}_{u,k+1}^{-1}-\Sigma_{u0}^{-1}\|_1 = o_P(\mu_{N,T}m_N)$. Hence

$$\begin{aligned}\|\widehat{\Sigma}_{u,k+1}^{-1}\|_1 &\le \|\widehat{\Sigma}_{u,k+1}^{-1}-\Sigma_{u0}^{-1}\|_1 + \|\Sigma_{u0}^{-1}\|_1\\ &= \|\Sigma_{u0}^{-1}\|_1 + o_P(\mu_{N,T}m_N).\end{aligned}$$

Finally, from $\|N^{-1}\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1} - N^{-1}\Lambda_0'\Lambda_0\| = O_P(\|\widehat{\Lambda}_k-\Lambda_0\|_\infty) = o_P(\mu_{N,T})$, we conclude that

$$\lambda_{\max}(N^{-1}\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1}) \le \lambda_{\max}(N^{-1}\Lambda_0'\Lambda_0) + o_P(\mu_{N,T}),$$

and

$$\lambda_{\min}(N^{-1}\widehat{\Lambda}_{k+1}'\widehat{\Lambda}_{k+1}) \ge \lambda_{\min}(N^{-1}\Lambda_0'\Lambda_0) - o_P(\mu_{N,T}).$$

The following lemma shows that the proposed algorithm, if converges, will converge to a stationary point of the penalized ML problem.

Lemma D.2. *Suppose* $\{\widehat{\Lambda}_k, \widehat{\Sigma}_{u,k}\}$ *converges to* $(\bar{\Lambda}, \bar{\Sigma}_u)$ *as* $k\to\infty$. *Then*

$$(S_y-\bar{\Sigma}_y)\bar{\Sigma}_u^{-1}\bar{\Lambda} = 0, \qquad \bar{\Sigma}_y = \bar{\Lambda}\bar{\Lambda}' + \bar{\Sigma}_u$$

$$(\bar{\Sigma}_y^{-1} - \bar{\Sigma}_y^{-1}S_y\bar{\Sigma}_y^{-1})_{ij} + \mu_{N,T}w_{ij}\bar{\rho}_{ij} = 0,$$

$$\bar{\rho}_{ij} = sign((\bar{\Sigma}_u)_{ij})\ if\ (\bar{\Sigma}_u)_{ij}\neq 0.$$

That means, $(\bar{\Lambda}, \bar{\Sigma}_u)$ *satisfies the Karush–Kuhn–Tucker (KKT) conditions of the penalized ML problem.*

Proof. By the iteration, we have $\bar{\Lambda} = S_y\bar{\Sigma}_y^{-1}\bar{\Lambda}M^{-1}$, $M = \bar{\Lambda}'\bar{\Sigma}_y^{-1}S_y\bar{\Sigma}_y^{-1}\bar{\Lambda} + I - \bar{\Lambda}'\bar{\Sigma}_y^{-1}\bar{\Lambda}$, which is equivalent to

$$S_y\bar{\Sigma}_y^{-1}\bar{\Lambda} = \bar{\Lambda}(\bar{\Lambda}'\bar{\Sigma}_y^{-1}S_y\bar{\Sigma}_y^{-1}\bar{\Lambda} + I - \bar{\Lambda}'\bar{\Sigma}_y^{-1}\bar{\Lambda}).$$

The left hand side is $(S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} + \bar{\Lambda}$. The right hand side equals

$$\begin{aligned}&\bar{\Lambda}(\bar{\Lambda}'\bar{\Sigma}_y^{-1}(S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} + I)\\ &\quad= \bar{\Lambda}\bar{\Lambda}'\bar{\Sigma}_y^{-1}(S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} + \bar{\Lambda}\\ &\quad= (\bar{\Sigma}_y-\bar{\Sigma}_u)\bar{\Sigma}_y^{-1}(S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} + \bar{\Lambda}\\ &\quad= (S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} - \bar{\Sigma}_u\bar{\Sigma}_y^{-1}(S_y-\bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} + \bar{\Lambda}.\end{aligned}$$

Hence $\bar{\Sigma}_u \bar{\Sigma}_y^{-1}(S_y - \bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} = 0$, which is equivalent to $(S_y - \bar{\Sigma}_y)\bar{\Sigma}_y^{-1}\bar{\Lambda} = 0$. Note that $\bar{\Sigma}_y^{-1}\bar{\Lambda} = \bar{\Sigma}_u^{-1}\bar{\Lambda}(I + \bar{\Lambda}'\bar{\Sigma}_u^{-1}\bar{\Lambda})^{-1}$. Hence $(S_y - \bar{\Sigma}_y)\bar{\Sigma}_u^{-1}\bar{\Lambda} = 0$.

On the other hand, $\widehat{\Sigma}_{u,k+1}$ solves the problem

$$\min_{\Sigma_u} \frac{1}{2t}\|\Sigma_u - \widehat{\Sigma}_{u,k} + t[\widehat{\Sigma}_{y,k}^{-1} - \widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}]\|_F^2 \\ + \sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}|,$$

whose KKT condition is, for $B = \widehat{\Sigma}_{u,k} - t[\widehat{\Sigma}_{y,k}^{-1} - \widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}]$,

$$\frac{1}{t}(\widehat{\Sigma}_{u,k+1} - B)_{ij} + \mu_{N,T}w_{ij}\rho_{ij} = 0, \\ \rho_{ij} = sign((\widehat{\Sigma}_{u,k+1})_{ij}) \quad \text{if } (\widehat{\Sigma}_{u,k+1})_{ij} \neq 0.$$

Let $k \to \infty$, we have, for $\bar{B} = \bar{\Sigma}_u - t[\bar{\Sigma}_y^{-1} - \bar{\Sigma}_y^{-1}S_y\bar{\Sigma}_y^{-1}]$, and some ρ_{ij} such that $\rho_{ij} = sign((\bar{\Sigma}_u)_{ij})$ if $(\bar{\Sigma}_u)_{ij} \neq 0$, $\frac{1}{t}(\bar{\Sigma}_u - B)_{ij} + \mu_{N,T}w_{ij}\rho_{ij} = 0$. It simplifies to

$$(\bar{\Sigma}_y^{-1} - \bar{\Sigma}_y^{-1}S_y\bar{\Sigma}_y^{-1})_{ij} + \mu_{N,T}w_{ij}\bar{\rho}_{ij} = 0, \\ \bar{\rho}_{ij} = sign((\bar{\Sigma}_u)_{ij}) \quad \text{if } (\bar{\Sigma}_u)_{ij} \neq 0,$$

which is also the KKT condition for the penalized ML

$$\min \frac{1}{N}\log\left|\det\left(\Lambda\Lambda' + \Sigma_u\right)\right| + \frac{1}{N}\mathrm{tr}\left(S_y(\Lambda\Lambda' + \Sigma_u)^{-1}\right) \\ + \sum_{i\neq j}\mu_{N,T}w_{ij}|\Sigma_{u,ij}|.$$

Therefore, $(\bar{\Lambda}, \bar{\Sigma}_u)$, if exists, is a stationary point of the algorithm. □

References

Alessi, L., Barigozzi, M., Capassoc, M., 2010. Improved penalization for determining the number of factors in approximate factor models. Statist. Probab. Lett. 80, 1806–1813.

An, L., Tao, P., 2005. The dc difference of convex functions programming and dca revisited with dc models of real world nonconvex optimization problems. Ann. Oper. Res. 133, 23–46.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Bai, J., Li, K., 2012a. Statistical analysis of factor models of high dimension. Ann. Statist. 40, 436–465.

Bai, J., Li, K., 2012b. Maximum Likelihood Estimation and Inference for Approximate Factor Models of High Dimension. MPRA Paper No. 42099. Forthcoming in *Review of Economics and Statistics*.

Bai, J., Liao, Y., 2013. Statistical Inferences Using Large Estimated Covariances for Panel Data and Factor Models. Available at SSRN 2353396.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.

Bai, J., Wang, P., 2015. Identification and Bayesian estimation of dynamic factor models. J. Bus. & Econom. Statist. 33 (2), 221–240.

Bickel, P., Levina, E., 2008. Covariance regularization by thresholding. Ann. Statist. 36, 2577–2604.

Bien, J., Tibshirani, R., 2011. Sparse estimation of a covariance matrix. Biometrika 98, 807–820.

Boivin, J., Ng, S., 2005. Understanding and comparing factor based macroeconomic forecasts. Int. J. Cent. Bank. 1 (3), 117–152.

Breitung, J., Tenhofen, J., 2011. GLS estimation of dynamic factor models. J. Amer. Statist. Assoc. 106, 1150–1166.

Cai, T., Zhou, H., 2012. Optimal rates of convergence for sparse covariance matrix estimation. Ann. Statist. 40, 2359–2763.

Caner, M., Fan, M., 2011. A Near Minimax Risk Bound: Adaptive Lasso With Heteroskedastic Data in Instrumental Variable Selection. *Manuscript*. North Carolina State University.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean–variance analysis in large asset markets. Econometrica 51, 1305–1324.

Choi, I., 2012. Efficient estimation of factor models. Econometric Theory 28, 274–308.

Connor, G., Korajczyk, R., 1993. A test for the number of factors in an approximate factor model. J. Finance 48, 1263–1291.

Deng, X., Tsui, K., 2013. Penalized covariance matrix estimation using a matrix-logarithm transformation. Journal of Computational and Graphical Statistics 22, 494–512.

Dias, F., Pinherio, M., Rua, A., 2013. Determining the number of global and country-specific factors in the euro area. Stud. Nonlinear Dyn. Econom. 17, 573–618.

Doz, C., Giannone, D., Reichlin, L., 2012. A quasi-maximum likelihood approach for large, approximate dynamic factor models. Rev. Econ. Stat. 94, 1014–1024.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econometrics 164, 188–205.

El Karoui, N., 2008. Spectrum estimation for large dimensional covariance matrices using random matrix theory. Ann. Statist. 36, 2757–2790.

Fan, J., Fan, Y., Lv, J., 2008. High dimensional covariance matrix estimation using a factor model. J. Econometrics 147, 186–197.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Liao, Y., Mincheva, M., 2011. High dimensional covariance matrix estimation in approximate factor models. Ann. Statist. 39, 3320–3356.

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

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic factor model: identification and estimation. Rev. Econ. Statist. 82, 540–554.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: one-sidedestimation and forecasting. J. Amer. Statist. Assoc. 100, 830–840.

Forni, M., Lippi, M., 2001. The generalized dynamic factor model: representation theory. Econometric Theory 17, 1113–1141.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. J. Monetary Econ. 55, 665–676.

Hallin, M., Liška, R., 2007. Determining the number of factors in the general dynamic factor model. J. Amer. Statist. Assoc. 102, 603–617.

Han, X., 2012. Determining the Number of Factors with Potentially Strong Cross-sectional Correlation in Idiosyncratic Shocks. *Manuscript*. North Carolina State University.

Huang, J., Ma, S., Zhang, C., 2008. Adaptive lasso for sparse high-dimensional regression models. Statist. Sinica 18, 1603–1618.

Jung, S., Marron, J.S., 2009. PC consistency in high dimension, low sample size context. Ann. Statist. 37, 4104–4130.

Kapetanios, G., 2010. A testing procedure for determining the number of factors in approximate factor models with large datasets. J. Bus. Econom. Statist. 28, 397–409.

Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37, 4254–4278.

Lam, C., Yao, Q., 2012. Factor modelling for high-dimensional time series: inference for the number of factors. Ann. Statist. 40, 694–726.

Lawley, D., Maxwell, A., 1971. Factor Analysis as a Statistical Method, second ed. Butterworths, London.

Ledoit, O., Wolf, M., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann. Statist 40, 1024–1060.

Luciani, M., 2014. Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. Int. J. Forecasting 30 (1), 20–29.

Ludvigson, S., Ng, S., 2011. A factor analysis of bond risk premia. In: Ulah, A., Giles, D. (Eds.), Handbook of Empirical Economics and Finance. Chapman and Hall, pp. 313–372.

Natalia, B., Kapetanios, G., Pesaran, H., 2012. Exponent of Cross-sectional Dependence: Estimation and Inference. Manuscript.

Neyman, J., Scott, E., 1948. Consistent estimation from partially consistent observations. Econometrica 16, 1–32.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92, 1004–1016.

Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. J. Econometrics 168, 244–258.

Pati, D., Bhattacharya, A., Pillai, N., Dunson, D., 2012. Posterior contraction in Sparse Bayesian Factor Models for Massive Covariance Matrices. *Manuscript*, Duke University.

Ravikumar, P., M, Wainwright, G, Raskutti, Yu, B., 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980.

Rohde, A., Tsybakov, A., 2011. Estimation of high-dimensional low-rank matrices. Ann. Statist 39, 887–930.

Rothman, A., 2012. Positive definite estimators of large covariance matrices. Biometrika 99, 733–740.

Rothman, A., Bickel, P., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.

Stock, J., Watson, M., 1998. Diffusion Indexes, NBER Working Paper 6702.

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

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B Stat. Methodol. 58, 267–288.

Tsai, H., Tsay, R., 2010. Constrained factor model. J. Amer. Statist. Assoc. 105, 1593–1605.

van de Geer, S., Bühlmann, P., Zhou, S., 2011. The adaptive and the thresholded lasso for potentially misspecified models and a lower bound for the lasso. Electron. J. Stat. 5, 688–749.

Wang, P., 2009. Large Dimensional Factor Models with a Multi-level Factor Structure: Identification, Estimation and Inference. *Manuscript*. Hong Kong University of Science and Technology.

Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.

Xue, L., Ma, S., Zou, H., 2012. Positive-definite ℓ_1-penalized estimation of large covariance matrices. J. Amer. Statist. Assoc. 107, 1480–1491.

Yuan, M., 2010. High dimensional inverse covariance matrix estimation via linear programming. J. Mach. Learn. Res. 2010, 2261–2286.

Zhou, S., Rütimann, P., Xu, M., Bühlmann, P., 2011. High-dimensional covariance estimation based on Gaussian graphical models. J. Mach. Learn. Res. 12, 2975–3026.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.