*The Annals of Statistics*
2012, Vol. 40, No. 1, 436–465
DOI: 10.1214/11-AOS966

# STATISTICAL ANALYSIS OF FACTOR MODELS OF HIGH DIMENSION

BY JUSHAN BAI[1] AND KUNPENG LI[2]

*Columbia University and the Central University of Finance and Economics, and Tsinghua University and University of International Business and Economics*

This paper considers the maximum likelihood estimation of factor models of high dimension, where the number of variables ($N$) is comparable with or even greater than the number of observations ($T$). An inferential theory is developed. We establish not only consistency but also the rate of convergence and the limiting distributions. Five different sets of identification conditions are considered. We show that the distributions of the MLE estimators depend on the identification restrictions. Unlike the principal components approach, the maximum likelihood estimator explicitly allows heteroskedasticities, which are jointly estimated with other parameters. Efficiency of MLE relative to the principal components method is also considered.

**1. Introduction.** Factor models provide an effective way of summarizing information from large data sets, and are widely used in social and physical sciences.[3] There has also been advancement in the theoretical analysis of factor models of high dimension. Much of this progress has been focused on the principal components method; see, for example, [5, 7, 22] and [23].[4] The advantage of the principal components method is that it is easy to compute and it provides consistent estimators for the factors and factor loadings when both $N$ and $T$ are large. The principal components method implicitly assumes that the idiosyncratic covariance matrix is a scalar multiple of an identity matrix. While the method is robust to heteroscedasticity and weak correlations in the idiosyncratic errors, there are biases associated with the estimates. In fact, if $N$ is fixed, the principal components estimator for the factor loadings is inconsistent, as shown in [5], except under homoscedasticity.

In this paper, we consider the maximum likelihood estimator under the setting of large $N$ and large $T$. The maximum likelihood estimator (MLE) is more efficient

Received June 2011; revised October 2011.

[1] Supported by NSF Grant SES-0962410.

[2] Supported by the new entrant program of UIBE (Grant Number: 11QD11).

*MSC2010 subject classifications.* Primary 62H25; secondary 62F12.

*Key words and phrases.* High-dimensional factor models, maximum likelihood estimation, factors, factor loadings, idiosyncratic variances, principal components.

[3] See, for example, [10, 11, 13, 20, 22, 23].

[4] [15] considers the principal components analysis in the frequency domain for generalized dynamic factor models.

than the principal components method. In addition, MLE is also consistent and efficient under fixed $N$ and large $T$ because this setting falls within the framework of classical inference; see, for example, [3] and [18].

Our estimator coincides with the classical factor analysis. However, the statistical theory does not follow from the existing literature. Classical inferential theory is based on the assumption that $N$ is fixed (or one of the dimensions is fixed). This assumption, to a certain extent, runs counter to the primary purpose of factor analysis, which is to explain the commonality among a large number of variables in terms of a small number of latent factors. Let $M_{zz} = \frac{1}{T-1}\sum_{t=1}^{T}(z_t - \bar{z})(z_t - \bar{z})'$ be the data matrix of $N \times N$ and let $\Sigma_{zz}(\theta) = E(M_{zz})$. A key assumption in classical inference is that $\sqrt{T}\,\mathrm{vech}(M_{zz} - \Sigma_{zz}(\theta))$ is asymptotically normal with a positive definite limiting covariance matrix, as $T \to \infty$. This assumption does not hold as $N$ also goes to infinity. The asymptotic normality is not well defined with an increasing dimension. For example, if $N > T$, $M_{zz}$ is a singular matrix, so it cannot have a normal distribution with a positive covariance matrix. Furthermore, the dimension of the unknown parameters (denoted by $\theta$) is also increasing as $N$ increases. The usual delta method (Taylor expansion) for deriving the limiting distribution of the MLE of $\theta$ will not work. Therefore, the high-dimensional inference for MLE requires a new framework.

Fixing $N$ is for the purpose of tractability for theoretical analysis. Such an assumption is unduly restrictive. Many applications or theoretical models involve data sets with the number of variables comparable with or even greater than the number of observations; see [11, 20, 22] and [23]. Although the large-$N$ analysis is demanding, the limiting distribution of the maximum likelihood estimator has a much simpler form under large $N$ than under fixed $N$.

There exists a small literature on efficient estimation of factors and factor loadings under large $N$. [9] considers a two-step approach by treating both the factors and the factor loadings as the parameters of interest. [12] also considers a two-step approach. The first step uses the principal components method to obtain the residuals and the second step uses a feasible generalized least squares. This method depends on large $N$ and large $T$ to get consistent estimation of the residual variances. MLE is considered by [14]. A certain average consistency is obtained; [14] does not consider consistency for individual parameters nor the limiting distributions.

The present paper is the first to develop a full statistical theory for the maximum likelihood estimator. Our approach is different from the existing literature. The challenge of the analysis lies in the simultaneous estimation of the heteroscedasticities and other model parameters. To estimate the heteroscedasticity, the maximum likelihood estimator does not rely on estimating the individual residuals, which would be the case for two-step procedures. Using residuals to construct variance estimators will be inconsistent when one of the dimension is fixed. The MLE remains consistent under fixed $N$.

The rest of this paper is organized as follows. Section 2 introduces the model and assumptions. Section 3 considers a symmetrical presentation for factor models. Identification conditions are considered in Section 4. Consistency and limiting distributions are derived in Section 5. Section 6 considers the estimation of factor scores. Section 7 compares the efficiency of the MLE relative to the principal components method and Section 8 discusses computational issues. The last section concludes. Proofs of consistency are given in the Appendix and additional proofs are provided in the supplement [6]. Throughout the paper, the norm of a vector or matrix is that of Frobenius, that is, $\|A\| = [\mathrm{tr}(A'A)]^{1/2}$ for vector or matrix $A$; $\mathrm{diag}(A)$ represents a diagonal matrix when $A$ is a vector, but $\mathrm{diag}(A)$ can be either a matrix or a column vector (consisting of the diagonal elements of $A$) when $A$ is a matrix.

## 2. Factor models.

Let $N$ denote the number of variables and $T$ the sample size. For $i = 1, \dots, N$ and $t = 1, \dots, T$, the observation $z_{it}$ is said to have a factor structure if it can be represented as

$$z_{it} = \alpha_i + \lambda_i' f_t + e_{it}, \tag{2.1}$$

where $f_t = (f_{t1}, f_{t2}, \dots, f_{tr})'$ and $\lambda_i = (\lambda_{i1}, \dots, \lambda_{ir})'$; both are $r \times 1$. Let $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_N)'$ be $N \times r$, and $z_t = (z_{1t}, \dots, z_{Nt})'$ be the $N \times 1$ vector of observable variables. Let $e_t$ and $\alpha$ be similarly defined. In matrix form,

$$z_t = \alpha + \Lambda f_t + e_t. \tag{2.2}$$

The vector $z_t$ is observable; none of the right-hand side variables are observable. We make the following assumptions:

ASSUMPTION A. $\{f_t\}$ is a sequence of fixed constants. Let $M_{ff} = \frac{1}{T} \times \sum_{t=1}^{T} (f_t - \bar{f})(f_t - \bar{f})'$ be the sample variance of $f_t$ where $\bar{f} = \frac{1}{T} \sum_{t=1}^{T} f_t$. There exists an $\overline{M}_{ff} > 0$ (positive definite) such that $\overline{M}_{ff} = \lim_{T\to\infty} M_{ff}$.

ASSUMPTION B. $E(e_t) = 0$; $E(e_t e_t') = \Sigma_{ee} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)$; $E(e_{it}^4) \le C^4$ for all $i$ and $t$, for some $C < \infty$. The $e_{it}$ are independent for all $i$ and $t$, and the $N \times 1$ vector $e_t$ is identically distributed over $t$.

ASSUMPTION C. There exists a positive constant $C$ large enough such that:

(C.1) $\|\lambda_j\| \le C$ for all $j$.

(C.2) $C^{-2} \le \sigma_j^2 \le C^2$ for all $j$.

(C.3) The limits $\lim_{N\to\infty} N^{-1} \Lambda' \Sigma_{ee}^{-1} \Lambda = Q$ and $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{-4} (\lambda_i \otimes \lambda_i)(\lambda_i' \otimes \lambda_i') = \Omega$ exist, where $Q$ and $\Omega$ are positive definite matrices.

ASSUMPTION D. The variances $\sigma_j^2$ are estimated in the compact set $[C^{-2}, C^2]$. Furthermore, $M_{ff}$ is restricted to be in a set consisting of all semi-positive definite matrices with all elements bounded in the interval $[-C, C]$, where $C$ is a large constant.

In Assumption A, we assume $f_t$ is a sequence of fixed constants. Our analysis holds if it is a sequence of random variables. In this case, we assume $f_t$ to be independent of all other variables. The analysis can then be regarded as conditioning on $\{f_t\}$. Without loss of generality, we assume that $\bar{f} = \frac{1}{T}\sum_{t=1}^{T} f_t = 0$ [or $E(f_t) = 0$ for random factors] because the model can be rewritten as $z_t = \alpha + \Lambda\bar{f} + \Lambda(f_t - \bar{f}) + e_t = \alpha^* + \Lambda f_t^* + e_t$ with $\alpha^* = \alpha + \Lambda\bar{f}$ and $f_t^* = f_t - \bar{f}$. In Assumption B, we assume $e_t$ to be independent over time. In fact, our consistent result still holds if $e_t$ are serially correlated and heteroscedastic over time or $e_{it}$ are correlated over $i$, provided that these correlations are weak (sufficient conditions are given in [2]). The limiting distribution then would need modification. For simplicity, we shall consider the uncorrelated case. The analysis of the maximum likelihood estimation under high dimension is already difficult; allowing correlation will make the analysis even more cumbersome. We will report the results under general correlation patterns in a separate paper. Assumption D is for theoretical analysis. Like all nonlinear (nonconvex) analysis, parameters are assumed to be in a bounded set.

The second moment of the sample, denoted by $M_{zz}$, is

$$M_{zz} = \frac{1}{T}\sum_{t=1}^{T}(z_t - \bar{z})(z_t - \bar{z})', \tag{2.3}$$

where $\bar{z} = T^{-1}\sum_{t=1}^{T} z_t$. Note that the division by $T$ instead of $T - 1$ is for notational simplicity. Let $\Sigma_{zz}$ be

$$\Sigma_{zz} = \Lambda M_{ff}\Lambda' + \Sigma_{ee}. \tag{2.4}$$

The objective function considered in this paper is

$$\ln L = -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}(M_{zz}\Sigma_{zz}^{-1}). \tag{2.5}$$

The above objective function may be regarded as a quasi likelihood function. To see this, assume $f_t$ is stochastic with mean zero and variance $\Sigma_{ff}$. From $z_t = \alpha + \Lambda f_t + e_t$, the variance matrix of $z_t$, denoted by $\Sigma_{zz}$, is

$$\Sigma_{zz} = \Lambda\Sigma_{ff}\Lambda' + \Sigma_{ee}.$$

So the quasi likelihood function (omitting a constant) can be written as

$$\begin{aligned}\ln L &= -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2NT}\sum_{t=1}^{T}(z_t - \alpha)'\Sigma_{zz}^{-1}(z_t - \alpha)\\ &= -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2NT}\sum_{t=1}^{T}(z_t - \bar{z})'\Sigma_{zz}^{-1}(z_t - \bar{z})\\ &\quad - \frac{1}{2NT}\sum_{t=1}^{T}(\bar{z} - \alpha)'\Sigma_{zz}^{-1}(\bar{z} - \alpha).\end{aligned}$$

Clearly $\hat{\alpha}$ minimizes the likelihood function at $\bar{z}$. So the concentrated quasi likelihood function can now be written as

$$\begin{aligned}\ln L &= -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\left[\frac{1}{T}\sum_{t=1}^{T}(z_t-\bar{z})(z_t-\bar{z})'\Sigma_{zz}^{-1}\right]\\ &= -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}(M_{zz}\Sigma_{zz}^{-1}),\end{aligned}$$

which is the same as (2.5) except that $\Sigma_{ff}$ is in place of $M_{ff}$. Because the factors are fixed constants instead of random variables, as stated in Assumption A, it is natural to use $M_{ff}$ rather than $\Sigma_{ff}$ in (2.4) and (2.5).

If both $\Lambda$ and $F = (f_1, f_2, \dots, f_T)'$ are treated as parameters, the corresponding likelihood function is

$$-\frac{1}{2N}\ln|\Sigma_{ee}| - \frac{1}{2NT}\sum_{t=1}^{T}(z_t-\alpha-\Lambda f_t)'\Sigma_{ee}^{-1}(z_t-\alpha-\Lambda f_t). \tag{2.6}$$

Since $\Lambda$ has $Nr$ parameters and $F$ has $Tr$ parameters, the number of parameters to be estimated will be very large, which leads to efficiency loss. In contrast, the number of parameters in (2.5) is only $N(r+1)+r(r+1)/2$, which is considerably smaller than the number of parameters in (2.6), which is $N(r+1)+Tr$. The difference is pronounced for small $N$ but large $T$. In fact, when estimating $\Lambda$, $F$ and $\Sigma_{ee}$, the global maximum likelihood estimator does not exist. It can be shown that the likelihood function diverges to infinity by certain choice of parameters (see [2], page 587).

By restricting $\Sigma_{ee} = I_N$ (an identity matrix), the MLE estimator of (2.6) becomes the principal components estimator. That is, the principal components method minimizes the objective function $\sum_{t=1}^{T}(z_t-\alpha-\Lambda f_t)'(z_t-\alpha-\Lambda f_t)$ over $\alpha$, $\Lambda$ and $F$. The estimators cannot be efficient when heteroscedasticity actually exists.

Even though the $f_t$ are fixed constants, we avoid directly estimating $f_t$. Instead we only estimate the sample moment of $f_t$. This considerably reduces the number of parameters and removes the corresponding incidental parameters bias. The estimator is also consistent under fixed $N$, since the setting falls back to the classical factor analysis.

By maximizing (2.5), in combination with (2.4), we can obtain three first-order conditions (see, e.g., [18]):

$$\hat{\Lambda}'\hat{\Sigma}_{zz}^{-1}(M_{zz}-\hat{\Sigma}_{zz}) = 0, \tag{2.7}$$

$$\operatorname{diag}(\hat{\Sigma}_{zz}^{-1}) = \operatorname{diag}(\hat{\Sigma}_{zz}^{-1}M_{zz}\hat{\Sigma}_{zz}^{-1}), \tag{2.8}$$

$$\hat{\Lambda}'\hat{\Sigma}_{zz}^{-1}\hat{\Lambda} = \hat{\Lambda}'\hat{\Sigma}_{zz}^{-1}M_{zz}\hat{\Sigma}_{zz}^{-1}\hat{\Lambda}, \tag{2.9}$$

where $\hat{\Lambda}$, $\hat{M}_{ff}$ and $\hat{\Sigma}_{ee}$ denote the MLE and $\hat{\Sigma}_{zz} = \hat{\Lambda}\hat{M}_{ff}\hat{\Lambda}' + \hat{\Sigma}_{ee}$.

Condition (2.7) is derived from the partial derivatives with respect to $\Lambda$, (2.8) is derived with respect to the diagonal elements of $\Sigma_{ee}$, and (2.9) is derived with respect to $M_{ff}$. Equation (2.9) can be obtained from (2.7) by post-multiplying $\hat{\Sigma}_{zz}^{-1}\hat{\Lambda}$. Since (2.9) is redundant, in order to make the system of three equations solvable, we need to impose further restrictions. These identification restrictions will be discussed in Section 4.

## 3. Symmetry and choice of representations.

Consider the model

$$z_{it} = \delta_t + \lambda_i' f_t + e_{it}.$$

Let $z_i = (z_{i1}, z_{i2}, \dots, z_{iT})'$, $\delta = (\delta_1, \dots, \delta_T)'$, $F = (f_1, f_2, \dots, f_T)'$ and $e_i = (e_{i1}, \dots, e_{iT})'$; then

$$z_i = \delta + F\lambda_i + e_i$$

$(i = 1, 2, \dots, N)$. Define

$$M_{zz} = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})', \qquad \Sigma_{zz} = FM_{\lambda\lambda}F' + \Sigma_{ee}^{\dagger},$$

where $\Sigma_{ee}^{\dagger} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_T^2)$ and

$$M_{\lambda\lambda} = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\lambda_i - \bar{\lambda})'$$

is the $r \times r$ sample variance of the factor loadings.

Although we use the same notation of $M_{zz}$ and $\Sigma_{zz}$, they are now $T \times T$ matrices instead of $N \times N$. The matrix $\Sigma_{ee}^{\dagger}$ contains idiosyncratic variances in the time dimension (time series heteroscedasticity). The quasi maximum likelihood estimator maximizes the likelihood function

$$\ln L = -\frac{1}{2T}\ln|\Sigma_{zz}| - \frac{1}{2T}\operatorname{tr}(M_{zz}\Sigma_{zz}^{-1}).$$

This representation avoids estimating $\lambda_1, \lambda_2, \dots, \lambda_N$ directly, but only the sample moment of $\lambda_i$. The representation has $Tr + T + r(r+1)/2$ number of parameters. If $N$ is much larger than $T$, this representation will give more efficient estimation. In particular, if $T$ is fixed, we can only use this representation to get consistent estimation of $f_1, f_2, \dots, f_T$ and $\Sigma_{ee}^{\dagger}$. This representation will also be useful if one is interested in estimating the heteroscedasticity in the time dimension.

The analysis of one representation will carry over to the other by switching the role of $N$ and $T$ and the role of $\Lambda$ and $F$. So it is sufficient to carefully examine one representation. Bearing this in mind, our analysis focuses on the representation in the previous section. The objective function in (2.5) involves fewer parameters when $N$ is less than $T$, although we make no assumption about the relative size between $N$ and $T$ (except for Theorem 6.1), and in particular, $N$ is allowed to be much larger than $T$.

**4. Identification conditions.** It is well known that the factor models are not identifiable without additional restrictions. For any $r \times r$ invertible matrix $R$, we have $\Lambda M_{ff} \Lambda' = \tilde{\Lambda} \tilde{M}_{ff} \tilde{\Lambda}'$ where $\tilde{\Lambda} = \Lambda R$ and $\tilde{M}_{ff} = R^{-1} M_{ff} R^{'-1}$. Thus observationally equivalent models are obtained. In order to uniquely fix $\Lambda$ and $M_{ff}$ given $\Lambda M_{ff} \Lambda'$, we need $r^2$ restrictions since an invertible $r \times r$ matrix has $r^2$ free parameters. For details of identification conditions, readers are referred to [18] and [4]. There are many ways to impose restrictions. In this paper, we consider five identification strategies which have been used in traditional factor analysis. These restrictions are listed in Table 1. The left pane is for the representation in Section 2, while the right pane is for the representation in Section 3.

We make some comments on these restrictions. Given $\Lambda M_{ff} \Lambda'$, IC1 will uniquely fix $\Lambda$ and $M_{ff}$. So full identification is achieved. But this is not the case for IC2. If we change the sign of any column of $\Lambda$, $\Lambda M_{ff} \Lambda'$ is not changed. This implies that we only identify $\Lambda$ up to a column sign change.

Furthermore, if we switch the positions between the $i$th and $j$th columns of $\Lambda$, and the positions between the $i$th and $j$th diagonal elements of $M_{ff}$, the matrix $\Lambda M_{ff} \Lambda'$ is not changed. This means that we need restrictions on the ordering of the diagonal elements of $M_{ff}$. In this paper, we assume that the diagonal components of $M_{ff}$ are arranged from the largest to the smallest and they must be distinct and positive. Because of this restriction, we naturally require that the diagonal elements of estimator $\hat{M}_{ff}$ are also arranged in this order, which is important for the proof of consistency.

Under IC3, for the same reason, we assume that the diagonal elements of $\frac{1}{N} \Lambda' \Sigma_{ee}^{-1} \Lambda$ are distinct and positive, and are arranged in decreasing order; $\Lambda$ is identified up to a column sign change.

IC4 imposes $\frac{1}{2} r(r+1)$ restrictions on the factor loadings, and $\frac{1}{2} r(r-1)$ restrictions on the factors. Identification is fully achieved like IC1.

Under IC5, we can only identify $\Lambda$ up to a column sign change. In addition, we need nonzero diagonal elements for the lower triangular matrix. The reason is intuitive. If the $i$th diagonal element is zero, both the $i$th and $(i+1)$th columns will share the same structure.

IC1 is related to the measurement error problem; it assumes that the first $r$ observations are noise measurements of the underlying factors. IC2 and IC3 are the usual restrictions for MLE; see [18]. IC4 and IC5 assume a recursive relation: the first factor affects the first variable only, and the first two factors affect the first two variables only, and so on; they are widely used, for example, [4] and [16]. Clearly, IC1, IC4 and IC5 require a careful choice of the first $r$ observations in practice. The inferential theory assumes that the underlying parameters satisfy the restrictions, implying different $\lambda_i$ under different restrictions.

**5. Asymptotic properties of the likelihood estimators.** Since the number of parameters increases as $N$ increases, the usual argument that the objective function

TABLE 1
*Identifying restrictions*

| | **Restrictions on $F$** | **Restrictions on $\Lambda$** | | **Restrictions on $\Lambda$** | **Restrictions on $F$** |
|---|---|---|---|---|---|
| IC1 | Unrestricted | $\Lambda = (I_r, \Lambda_2')'$ | IC1′ | Unrestricted | $F = (I_r, F_2')'$ |
| IC2 | $M_{ff} = \text{diagonal}$<br>(with distinct elements) | $\frac{1}{N}\Lambda'\Sigma_{ee}^{-1}\Lambda = I_r$ | IC2′ | $M_{\lambda\lambda} = \text{diagonal}$<br>(with distinct elements) | $\frac{1}{T}F'\Sigma_{ee}^{\dagger-1}F = I_r$ |
| IC3 | $M_{ff} = I_r$ | $\frac{1}{N}\Lambda'\Sigma_{ee}^{-1}\Lambda = \text{diagonal}$<br>(with distinct elements) | IC3′ | $M_{\lambda\lambda} = I_r$ | $\frac{1}{T}F'\Sigma_{ee}^{\dagger-1}F = \text{diagonal}$<br>(with distinct elements) |
| IC4 | $M_{ff} = \text{diagonal}$ | $\Lambda = (\Lambda_1', \Lambda_2')'$<br>$\Lambda_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \lambda_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{r1} & \lambda_{r2} & \cdots & 1 \end{pmatrix}$ | IC4′ | $M_{\lambda\lambda} = \text{diagonal}$ | $F = (F_1', F_2')'$<br>$F_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ f_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{r1} & f_{r2} & \cdots & 1 \end{pmatrix}$ |
| IC5 | $M_{ff} = I_r$ | $\Lambda' = (\Lambda_1', \Lambda_2')'$<br>$\Lambda_1 = \begin{pmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{r1} & \lambda_{r2} & \cdots & \lambda_{rr} \end{pmatrix}$<br>$\lambda_{ii} \neq 0, i = 1, 2, \ldots, r$ | IC5′ | $M_{\lambda\lambda} = I_r$ | $F' = (F_1', F_2')'$<br>$F_1 = \begin{pmatrix} f_{11} & 0 & \cdots & 0 \\ f_{21} & f_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f_{r1} & f_{r2} & \cdots & f_{rr} \end{pmatrix}$<br>$f_{ii} \neq 0, i = 1, 2, \ldots, r$ |

converges in probability to a fixed nonrandom function and the function achieves its maximum value at the true parameter values will not work. This is because as $N$ and $T$ increase, there will be an infinite number of parameters in the limit. Our idea of consistency is to obtain some average consistency, and then use these initial results to obtain consistency for individual parameters. Even the average consistency requires a novel argument in the presence of an increasing number of parameters.

PROPOSITION 5.1. *Let $\hat{\theta}$ be the MLE by maximizing* (2.5), *where $\hat{\theta} = (\hat{\lambda}_1, \dots, \hat{\lambda}_N, \hat{\sigma}_1^2, \dots, \hat{\sigma}_N^2, \hat{M}_{ff})$. Under Assumptions* A–D, *when $N, T \to \infty$, with any one of the identification conditions* IC1–IC5, *we have*

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\|\hat{\lambda}_i - \lambda_i\|^2 \xrightarrow{p} 0, \tag{5.1a}$$

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i^2 - \sigma_i^2)^2 \xrightarrow{p} 0, \tag{5.1b}$$

$$\hat{M}_{ff} - M_{ff} \xrightarrow{p} 0. \tag{5.1c}$$

Establishing the above result requires a considerable amount of work. Developing and identifying appropriate strategies have taken an even greater amount of efforts. The difficulty lies in the problem of infinite number of parameters in the limit and the nonlinearity of objective function. The infinite number of parameters problem in this paper is fundamentally different from those in the existing literature. For example, consider an $AR(\infty)$ process $X_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j}$. Although there exist an infinite number of parameters $\{a_j\}_{j=1}^{\infty}$, the assumption that $a_j \to 0$, as $j \to \infty$, effectively limits the number of parameters. For example, $\hat{a}_j \equiv 0$ is consistent for $a_j$ for $j \geq \ln(T)$. The assumption that $a_j \to 0$ may be viewed as one form of smoothing restriction. However, in the present context and in the absence of any form of smoothness, all parameters are free parameters, and there will be an infinite number of them in the limit. This is the source of difficulty.

While there is also an infinite number of parameters problem in the analysis of the principal components (PC) estimator, the method does not estimate heteroscedasticity, and it minimizes an objective function stated in Section 2. Its degree of nonlinearity is much less than the likelihood function (2.5). It is the joint estimation of heteroscedasticity that makes the analysis difficult. In the Appendix, we provide a novel proof of consistency, which constitutes a departure from the usual analysis, say, in [19] and [24].

The proofs of (5.1a) and (5.1c) depend heavily on the identification conditions. If we denote $A \equiv (\hat{\Lambda} - \Lambda)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}$, the proof of consistency centers on proving $A \xrightarrow{p} 0$. However, the proof of $A \xrightarrow{p} 0$ is quite different with different identification conditions. Under IC2, for example, $(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}/N)^{-1} = I_r$. Under other identification conditions, the proof of even $(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}/N)^{-1} = O_p(1)$ is extremely demanding. Under IC2, IC3 and IC5, we need to assume that the estimator $\hat{\Lambda}$ has the same column signs as those of $\Lambda$ in order to have consistency. Having the same column signs is regarded as part of the identification restrictions under IC2, IC3 and IC5.

In order to derive the inferential theory for the estimated parameters, we need to strengthen Proposition 5.1. We state the result as a theorem:

THEOREM 5.1. *Under the assumptions of Proposition* 5.1, *we have*

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\|\hat{\lambda}_i - \lambda_i\|^2 = O_p(T^{-1}), \tag{5.2a}$$

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i^2 - \sigma_i^2)^2 = O_p(T^{-1}), \tag{5.2b}$$

$$\|\hat{M}_{ff} - M_{ff}\|^2 = O_p(T^{-1}). \tag{5.2c}$$

It is interesting to compare our results with those in classical factor analysis. If $N$ is fixed, the existing literature has already shown that $\hat{\lambda}_j$ and $\hat{\sigma}_j^2$ converge to $\lambda_j$ and $\sigma_j^2$ at the rate of $\sqrt{T}$ for any $j$. Since $N$ is fixed, the classical result implies (5.2a) and (5.2b). In fact $\|\hat{M}_{ff} - M_{ff}\|^2 = O_p(T^{-1})$ holds also since it can be derived from the first two (the results analogous to (5.2c) under IC1 when $N$ is finite can be seen in [1]). Theorem 5.1 shows that these results still hold in the large-$N$ setting despite estimating an increasing number of elements. However, we point out that the rate stated in Theorem 5.1 is not the sharpest. If IC2 or IC3 is adopted as identification conditions, then $\|\hat{M}_{ff} - M_{ff}\|^2 = O_p(T^{-1})$ can be refined as $\|\hat{M}_{ff} - M_{ff}\|^2 = O_p(N^{-1}T^{-1}) + O_p(T^{-2})$. Because (5.2c) is sufficient for the inferential theory to be developed, we only state this general result.

As pointed out earlier, the behavior of $A \equiv (\hat{\Lambda} - \Lambda)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}$ is important in establishing consistency. In fact, matrix $A$ plays a key role in the inferential theory as well. The convergence rate of $A$ depends on identification conditions. We use $A_k$ in place of $A$ under IC$k$ ($k = 1, 2, \ldots, 5$). Under IC2 and IC3, the convergence rate of $A$ is $\min(\sqrt{NT}, T)$. However, under other sets of identification conditions, the convergence rate of $A$ is $\sqrt{T}$. This difference in convergence rate affects the limiting distribution of $\hat{M}_{ff}$, which also makes the limiting distributions of $\hat{\lambda}_j$ different.

In Section C of the supplement [6], we give the asymptotic representations of $\sqrt{T}(\hat{\lambda}_j - \lambda_j)$ under IC1–IC5. The main representations are given in (C.5), (C.12), (C.17) and (C.24), respectively. The following theorem is a consequence of these representations.

THEOREM 5.2. *Under the assumptions of Proposition* 5.1, *for each* $j = 1, 2, \dots, N$, *as* $N, T \to \infty$, *we have*:

*Under* IC1,

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}(\lambda_j' \Sigma_{eer} \lambda_j + \sigma_j^2)). \tag{5.3a}$$

*Under* IC2 *or* IC3,

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}\sigma_j^2). \tag{5.3b}$$

*Under* IC4, *for* $j > r$

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, \Pi + (\overline{M}_{ff})^{-1}\sigma_j^2) \tag{5.3c}$$

*and for* $2 \le j \le r$

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, \Pi + 2I_r^{j-1}(\overline{M}_{ff})^{-1}\sigma_j^2 - (\overline{M}_{ff})^{-1}\sigma_j^2).$$

*Under* IC5, *for* $j > r$

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, \Xi + I_r\sigma_j^2) \tag{5.3d}$$

*and for* $1 \le j \le r$

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, \Xi + 2I_r^j\sigma_j^2 - I_r\sigma_j^2),$$

*where* $\Pi = (\lambda_j' \otimes I_r)\tilde{D}(\overline{M}_{ff})\Phi\tilde{D}(\overline{M}_{ff})'(\lambda_j \otimes I_r)$, $\Xi = (\lambda_j' \otimes I_r)\overline{D}\Gamma\overline{D}'(\lambda_j \otimes I_r)$. $\Sigma_{eer}$ *is an* $r \times r$ *diagonal matrix with the* $j$*th diagonal element* $\sigma_j^2$; $I_r^j$ *is an* $r \times r$ *diagonal matrix with the first* $j$ *diagonal elements being* 1 *and the rest being* 0. *The meanings of* $\tilde{D}(\overline{M}_{ff})$, $\overline{D}$, $\Phi$ *and* $\Gamma$ *are explained below.*

The matrix $\tilde{D}(M)$ (with $M = \overline{M}_{ff}$) is a generalized duplication matrix of $r^2 \times \frac{1}{2}r(r+1)$ depending on the diagonal matrix $M$; $\tilde{D}(M)$ can be constructed row by row in the following way. Given the number $k$, $1 \le k \le r^2$, we denote $j = \lfloor (k-1)/r \rfloor + 1$ and $i = k - (j-1)r$, where $\lfloor \cdot \rfloor$ denotes the largest integer no greater than the argument. If $i \ge j$, all elements of the $k$th row are zero, except that the $(\frac{1}{2}(2r - j + 2)(j-1) + i - j + 1)$th element is 1; if $i < j$, all elements are zero, except that the $(\frac{1}{2}(2r - i + 2)(i-1) - i + j + 1)$th element is $-m_j m_i^{-1}$, where $m_j$ is the $j$th diagonal element of $M$. The $r^2 \times \frac{1}{2}r(r-1)$ matrix $\overline{D}$ under IC5 is also a generalized duplication matrix. Let $A$ be a skew-symmetric matrix and let $\operatorname{veck}(A)$ be the operator that stacks the elements of $A$ strictly below the diagonal into a vector (excluding diagonal elements). Then $\overline{D}$ is defined as $\operatorname{vec}(A) = \overline{D}\operatorname{veck}(A)$.

Here are some examples for $\tilde{D}(M)$. If $M$ is a scalar, then $\tilde{D}(M) = 1$. If $M = \operatorname{diag}(m_1, m_2)$, then

$$\tilde{D}(M) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_2/m_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

If $M = \operatorname{diag}(m_1, m_2, m_3)$, then

$$\tilde{D}(M) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -m_2/m_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -m_3/m_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -m_3/m_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here are some examples for $\overline{D}$, which only depends on the dimension of $\overline{M}_{ff}$ (i.e., the number of factors). If $r = 1$, then $\overline{D} = 0$. If $r = 2$, then $\overline{D} = (0, 1, -1, 0)'$. If $r = 3$, then

$$\overline{D} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\Phi$ in (5.3c) is the limiting covariance of $\operatorname{vech}(A_4)$, where $A_4$ is the $A$ matrix defined earlier under IC4. The asymptotic representation of $A_4$ is given by (C.15) in Section C of the supplement [6]. From this asymptotic representation, the elements of $\Phi$ can be easily computed and are given by (C.16). The matrix $\Gamma$ in (5.3d) is the limiting covariance of $\operatorname{veck}(A_5)$, where $A_5$ is the matrix $A$ under IC5, and is asymptotically skew-symmetric. The asymptotic representation of $A_5$ is given by (C.22). The elements of $\Gamma$ are determined by (C.23) in Section C of the supplement [6].

*Remarks*: In classical factor analysis, the MLE usually imposes the restriction of IC3. The limiting distribution $\hat{\lambda}_j$ in classical factor analysis (fixed $N$) is very complicated; see [18]. Most textbooks on multivariate statistics do not even present the limiting distributions owing to its complexity. As pointed out by Anderson ([2], page 583), the limiting distribution is "too complicated to derive or even present here." In contrast, the limiting distribution under IC3 with large $N$ is

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}\sigma_j^2).$$

This is as efficient as the case in which the $f_t$ are observable. For if the $f_t$ are observable, the estimator of $\lambda_j$ by applying OLS to (2.1) is $\hat{\lambda}_j^{\text{ols}} = (T^{-1}\sum_{t=1}^T (f_t - \bar{f})(f_t - \bar{f})')^{-1}(T^{-1}\sum_{t=1}^T (f_t - \bar{f})(z_{jt} - \bar{z}_j))$. It is easy to show that $\sqrt{T}(\hat{\lambda}_j^{\text{ols}} -$

$\lambda_j) \xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}\sigma_j^2)$, the same as the MLE estimator under IC2 and IC3. The OLS estimator is the best linear unbiased estimator under Assumption B, as the error terms of the regression equation are i.i.d.

Now we state the limiting distributions of $\hat{M}_{ff}$ and $\hat{\sigma}_j^2$ for each $j$. In Section C of the supplement [6], we derive the asymptotic representations for $\hat{M}_{ff} - M_{ff}$ under IC1, IC2 and IC4, respectively, which are given by (C.7), (C.11) and (C.19). The asymptotic representation for $\hat{\sigma}_j^2 - \sigma_j^2$ is given by (C.4). The next two theorems follow these asymptotic representations.

THEOREM 5.3. *Under the assumptions of Proposition* 5.1, *we have*:
*Under* IC1,

$$\sqrt{T}\,\mathrm{vech}(\hat{M}_{ff} - M_{ff}) \xrightarrow{d} \mathcal{N}\bigl(0, 4D_r^{+}(\Sigma_{eer} \otimes \overline{M}_{ff})D_r^{+\prime}\bigr).$$

*Under* IC2, $N/T \to 0$ *and normality of* $e_{it}$,

$$\sqrt{NT}\,\mathrm{diag}(\hat{M}_{ff} - M_{ff})$$
$$\xrightarrow{d} \mathcal{N}\bigl(0, J_r[2(I_r \otimes \overline{M}_{ff})\Omega(I_r \otimes \overline{M}_{ff}) + 4(Q \otimes \overline{M}_{ff})]J_r'\bigr).$$

*Under* IC4,

$$\sqrt{T}\,\mathrm{diag}(\hat{M}_{ff} - M_{ff}) \xrightarrow{d} \mathcal{N}\bigl(0, 4J_r[(\Lambda_1'\Sigma_{eer}^{-1}\Lambda_1)^{-1} \otimes \overline{M}_{ff}]J_r'\bigr),$$

*where* $D_r^{+}$ *is the Moore–Penrose inverse of the duplication matrix* $D_r$; $J_r$ *is an* $r \times r^2$ *matrix, which satisfies, for any* $r \times r$ *matrix* $M$, $\mathrm{diag}\{M\} = J_r \,\mathrm{vec}(M)$, *where* $\mathrm{diag}\{\cdot\}$ *is the operator which stacks the diagonal elements into a vector.*

Note under IC3 and IC5, $M_{ff}$ is known and thus not estimated. Normality under IC2 is used only for calculating the limiting variance. Given the asymptotic representation of $\hat{M}_{ff} - M_{ff}$, it is easy to derive the limiting distribution under nonnormality.

THEOREM 5.4. *Under the assumptions of Proposition* 5.1, *with any set of the identification conditions, we have*

$$\sqrt{T}(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{d} \mathcal{N}\bigl(0, \sigma_j^4(2 + \kappa_j)\bigr), \tag{5.4}$$

*where* $\kappa_j$ *is the excess kurtosis of* $e_{jt}$. *Under normality of* $e_{it}$, *the limiting distribution becomes* $\mathcal{N}(0, 2\sigma_j^4)$.

Our analysis assumes that the underlying parameters satisfy the identification restrictions, which is also the classical framework of [18]. A consequence is that we are directly estimating the underlying true parameters instead of rotations of them. The rotation matrix used in [23] and [7] degenerates into an identity matrix. This result itself is interesting.

**6. Asymptotic properties for the estimated factors.** The factors $f_t$ can be estimated by two different methods. One is the projection formula and the other is the generalized least squares (GLS). These methods are discussed in [2].

If the factor $f_t$ is normally distributed with mean zero and variance $\Sigma_{ff}$, and is independent of $e_t$, then the joint distribution of $(f_t, z_t)$, by (2.2), can be written as

$$\begin{bmatrix} f_t \\ z_t \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \begin{matrix} \Sigma_{ff} & \Sigma_{ff}\Lambda' \\ \Lambda\Sigma_{ff} & \Lambda\Sigma_{ff}\Lambda' + \Sigma_{ee} \end{matrix}\right]. \tag{6.1}$$

Given $z_t$, the best predictor of $f_t$, $f_t^p$, is $f_t^p = \Sigma_{ff}\Lambda'(\Lambda\Sigma_{ff}\Lambda' + \Sigma_{ee})^{-1}(z_t - \alpha)$. By the basic result $(\Lambda\Sigma_{ff}\Lambda' + \Sigma_{ee})^{-1} = \Sigma_{ee}^{-1} - \Sigma_{ee}^{-1}\Lambda(\Sigma_{ff}^{-1} + \Lambda'\Sigma_{ff}^{-1}\Lambda)^{-1} \times \Lambda'\Sigma_{ee}^{-1}$, we have

$$f_t^p = (\Sigma_{ff}^{-1} + \Lambda'\Sigma_{ee}^{-1}\Lambda)^{-1}\Lambda'\Sigma_{ee}^{-1}(z_t - \alpha). \tag{6.2}$$

Although (6.2) is deduced under the assumption of normality of $f_t$ and in this paper the $f_t$ are fixed constants, equation (6.2) can still be used to estimate $f_t$ by replacing the parameters with their corresponding estimates. So the estimator $\tilde{f}_t$ is

$$\tilde{f}_t = (\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(z_t - \bar{z}). \tag{6.3}$$

An alternative procedure is the GLS. If the $\lambda_j$ and $\sigma_j^2$ are observable, the GLS estimator of $f_t$ is $(\Lambda'\Sigma_{ee}^{-1}\Lambda)^{-1}\Lambda'\Sigma_{ee}^{-1}(z_t - \bar{z})$. The unknown variables can be replaced by their estimates. We define the GLS estimator of $f_t$ as

$$\hat{f}_t = (\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(z_t - \bar{z}). \tag{6.4}$$

Under large $N$, not much difference exists between (6.3) and (6.4). In fact, they are asymptotically equivalent and have the same limiting distributions. But for relatively small $N$, the difference may not be ignorable.

PROPOSITION 6.1. *Under the assumptions of Proposition* 5.1, $\tilde{f}_t = \hat{f}_t + O_p(1/N)$.

Since $\tilde{f}_t$ and $\hat{f}_t$ have the same limiting distribution, we only state the distribution for (6.4). In Section D of the supplement [6] we derive the asymptotic representations for $\sqrt{N}(\hat{f}_t - f_t)$, which are given by (D.3), (D.4) and (D.5), respectively. From these representations, we obtain:

THEOREM 6.1. *Let* $\Delta \in [0, \infty)$. *Under the assumptions of Proposition* 5.1 *and* $\sqrt{N}/T \to 0$, *we have*:

*Under* IC1 *and* $N/T \to \Delta$,

$$\sqrt{N}(\hat{f}_t - f_t) \xrightarrow{d} \mathcal{N}(0, \Delta f_t'(\overline{M}_{ff})^{-1} f_t \Sigma_{eer} + Q^{-1}). \tag{6.5a}$$

*Under* IC2,

$$\sqrt{N}(\hat{f}_t - f_t) \xrightarrow{d} \mathcal{N}(0, I_r). \tag{6.5b}$$

*Under* IC3,

$$\sqrt{N}(\hat{f}_t - f_t) \xrightarrow{d} \mathcal{N}(0, Q^{-1}). \tag{6.5c}$$

*Under* IC4 *and* $N/T \to \Delta$,

$$\begin{aligned}&\sqrt{N}(\hat{f}_t - f_t)\\ &\quad \xrightarrow{d} \mathcal{N}\bigl(0, \Delta(I_r \otimes f_t')\tilde{D}(\overline{M}_{ff})\Phi\tilde{D}'(\overline{M}_{ff})(I_r \otimes f_t) + Q^{-1}\bigr).\end{aligned} \tag{6.5d}$$

*Under* IC5 *and* $N/T \to \Delta$,

$$\sqrt{N}(\hat{f}_t - f_t) \xrightarrow{d} \mathcal{N}\bigl(0, \Delta(I_r \otimes f_t')\overline{D}\Gamma\overline{D}'(I_r \otimes f_t) + Q^{-1}\bigr). \tag{6.5e}$$

*The matrix $Q$ is defined in Assumption* C. *The matrices $\Sigma_{eer}$, $\Phi$, $\Gamma$, $\tilde{D}(M_{ff})$ and $\overline{D}$ are defined in Theorem* 5.2.

If $\Delta = 0$, Theorem 6.1 shows that $\hat{f}_t$ has the same limiting distribution regardless of the identification restrictions. In this case, the variance is equal to $Q = \lim_{N\to\infty} N^{-1}\Lambda'\Sigma_{ee}^{-1}\Lambda$. Note that under IC2, $Q = I_r$. Irrespective of whether $\Delta$ is zero, $\hat{f}_t$ is efficient under IC2 and IC3 in the sense that the limiting variance coincides with the situation in which both the factor loadings and the variances $\sigma_t^2$ are observable and GLS is applied to a cross-sectional regression for each fixed $t$. This result requires $\sqrt{N}/T \to 0$.

Recall that a factor model has two symmetrical representations. We also discussed earlier which presentation should be used in practice. If $N$ is smaller than $T$, we should estimate the factor loadings by the maximum likelihood method because this representation has fewer number of parameters. The opposite is true if $T$ is smaller than $N$. This intuitive argument is borne out by the results of Theorems 5.2 and 6.1. By Theorem 5.2, the magnitudes of $N$ and $T$ do not affect the limiting covariance of $\hat{\lambda}_j$ (other than the rate of convergence) but they do affect the limiting covariance of $\hat{f}_t$ as shown by Theorem 6.1. If $\Delta$ is large, $\hat{f}_t$ cannot be estimated well under IC1, IC4 and IC5. Note that $\hat{f}_t$ is not the maximum likelihood estimator. In this case, we can use the representation of Section 3 and directly estimate $f_t$ by the maximum likelihood method.

## 7. Comparison with the principal components method.

The method of principal components (PC) does not assume a factor model, and the method is usually regarded as a dimension reduction technique. But PC can be used to estimate factor models under large $N$ and large $T$; see [7, 11, 13] and [22]. Let $\hat{\lambda}_j^{\text{pc}}$ and $\hat{f}_t^{\text{pc}}$ denote the PC estimators for $\lambda_j$ and $f_t$, respectively. The results of [5] and [8] imply the following asymptotic representation for the principal components estimators. As $N, T \to \infty$, if $\sqrt{N}/T \to 0$, then

$$\sqrt{T}(\hat{\lambda}_j^{\text{pc}} - \lambda_j) = \left(\frac{1}{T}\sum_{t=1}^{T} f_t f_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t e_{jt}\right) + o_p(1)$$

$$\xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}\sigma_j^2)$$

and if $\sqrt{N}/T \to 0$, then

$$\sqrt{N}(\hat{f}_t^{\text{pc}} - f_t) = \left(\frac{1}{N}\sum_{i=1}^{N} \lambda_i \lambda_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \lambda_i e_{it}\right) + o_p(1)$$

$$\xrightarrow{d} \mathcal{N}(0, (\overline{M}_{\lambda\lambda})^{-1}\Upsilon(\overline{M}_{\lambda\lambda})^{-1}),$$

where $\overline{M}_{\lambda\lambda} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \lambda_i\lambda_i'$ and $\Upsilon = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \lambda_i\lambda_i'\sigma_i^2$.

The PC estimator in [5] uses the identification restriction IC3. Under IC3, we already show that the MLE satisfies, as $N, T \to \infty$,

$$\sqrt{T}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0, (\overline{M}_{ff})^{-1}\sigma_j^2).$$

Theorem 6.1 above shows that, under IC3, $\sqrt{N}(\hat{f}_t - f_t) \xrightarrow{d} \mathcal{N}(0, Q^{-1})$. This result requires $\sqrt{N}/T \to 0$, which is satisfied if $N/T \to \Delta$.

While $\hat{\lambda}_j^{\text{pc}}$ and $\hat{\lambda}_j$ have the same limiting distribution, $\hat{f}_t^{\text{pc}}$ is less efficient than $\hat{f}_t$. This follows because the sandwich form of the covariance matrix $(\overline{M}_{\lambda\lambda})^{-1}\Upsilon(\overline{M}_{\lambda\lambda})^{-1}$ is no smaller than $Q^{-1}$, where $Q$ is the limit of $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma_i^2} \times \lambda_i\lambda_i'$. Moreover, under IC3, the MLE $\hat{\lambda}_j$ only requires $N, T \to \infty$, but the PC estimator $\hat{\lambda}_j^{\text{pc}}$ requires an additional assumption that $\sqrt{T}/N \to 0$. Furthermore, the maximum likelihood estimator $\hat{\lambda}_j$ is consistent under fixed $N$, but $\hat{\lambda}_j^{\text{pc}}$ requires both $N$ and $T$ to be large in order to have consistency. Of course, under fixed $N$, the limiting distribution of MLE will have a different (more complicated) asymptotic covariance matrix; see [18].

To estimate $\sigma_j^2$, the PC method would need to estimate the individual residuals $\hat{e}_{it} = X_{it} - \hat{\alpha}_i - (\hat{\lambda}_i^{\text{pc}})'\hat{f}_t^{\text{pc}}$ and then construct $\hat{\sigma}_j^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_{jt}^2$. In case that $N$ is fixed, $f_t$ cannot be consistently estimated, so $\hat{e}_{it}$ is inconsistent for $e_{it}$. This further implies that $\hat{\sigma}_j^2$ is inconsistent for $\sigma_j^2$. In comparison, the MLE does not estimate the individuals $\hat{e}_{it}$. The variances are estimated jointly with the factor loadings $\lambda_j$ and with the matrix $M_{ff}$. The variance estimator remains consistent under fixed $N$.

Finally, the PC estimator for $\lambda_i$ satisfies (see [5]) $\frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_i^{\text{pc}} - \lambda_i\|^2 = O_p(\frac{1}{N}) + O_p(\frac{1}{T})$, while the MLE for $\lambda_i$ satisfies $\frac{1}{N}\sum_{i=1}^{N}\|\hat{\lambda}_i - \lambda_i\|^2 = O_p(\frac{1}{T})$.

**8. Computational issues.** The maximum likelihood estimation can be implemented via the EM algorithm and is considered by [21]. The EM algorithm is an iterated approach. To be specific, consider the identification condition IC3. Once the estimator under IC3 is obtained, estimators under other identification restrictions can be easily obtained (to be discussed below). Under IC3, we only need to estimate $\Lambda$ and $\Sigma_{ee}$ since $M_{ff} = I_r$.

Let $\theta^{(k)} = (\Lambda^{(k)}, \Sigma_{ee}^{(k)})$ denote the estimator at the $k$th iteration. The EM algorithm updates the estimator according to

$$\Lambda^{(k+1)} = \left[\frac{1}{T}\sum_{t=1}^{T} E(z_t f_t' | Z, \theta^{(k)})\right]\left[\frac{1}{T}\sum_{t=1}^{T} E(f_t f_t' | Z, \theta^{(k)})\right]^{-1},$$

$$\Sigma_{ee}^{(k+1)} = \operatorname{diag}(M_{zz} - \Lambda^{(k+1)}\Lambda^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1} M_{zz}),$$

where $\Sigma_{zz}^{(k)} = \Lambda^{(k)}\Lambda^{(k)\prime} + \Sigma_{ee}^{(k)}$, and

$$\frac{1}{T}\sum_{t=1}^{T} E(f_t f_t' | Z, \theta^{(k)}) = \Lambda^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1} M_{zz} (\Sigma_{zz}^{(k)})^{-1}\Lambda^{(k)} + I_r - \Lambda^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1}\Lambda^{(k)},$$

$$\frac{1}{T}\sum_{t=1}^{T} E(z_t f_t' | Z, \theta^{(k)}) = M_{zz}(\Sigma_{zz}^{(k)})^{-1}\Lambda^{(k)}.$$

This gives $\theta^{(k+1)} = (\Lambda^{(k+1)}, \Sigma_{ee}^{(k+1)})$. The iteration continues until $\|\theta^{(k+1)} - \theta^{(k)}\|$ is smaller than a preset tolerance. In the simulation reported below, we use the principal components estimator as the starting value. Let $(\Lambda^{\dagger}, \Sigma^{\dagger})$ denote the final round of iteration. Let $\mathcal{V}$ be the orthogonal matrix consisting of the eigenvectors of $\frac{1}{N}\Lambda^{\dagger\prime}(\Sigma_{ee}^{\dagger})^{-1}\Lambda^{\dagger}$ corresponding to descending eigenvalues. Let $\hat{\Lambda} = \Lambda^{\dagger}\mathcal{V}$ and $\hat{\Sigma}_{ee} = \Sigma_{ee}^{\dagger}$. Then $\hat{\theta} = (\hat{\Lambda}, \hat{\Sigma}_{ee})$ satisfies IC3. For general models, [25] shows that the EM solutions are stationary points of the likelihood functions. For completeness, we provide a direct and simple proof of this claim for factor models in the supplement [6] (Section E).

It is interesting to note that, under large $N$ and large $T$, the number of iterations needed to achieve convergence is smaller than under either a small $N$ or a small $T$. In Section E of the supplement [6], we also explain how to write a computer program so it runs fast.

Let $(\hat{\Lambda}, \hat{\Sigma}_{ee})$ denote the MLE under IC3. We discuss how to obtain estimators that satisfy other identification restrictions. First, note that $\hat{\Sigma}_{ee}$ is identical under IC1–IC5. We only need to discuss how to obtain $\Lambda$ and $M_{ff}$. Let $\hat{\Lambda}^{\ell}$ and $\hat{M}_{ff}^{\ell}$ denote the MLE under IC$\ell$ ($\ell = 1, \dots, 5$). Let $\hat{\Lambda}_1$ denote the first $r \times r$ block of $\hat{\Lambda}$. For IC1, let $\hat{\Lambda}^1 = \hat{\Lambda}(\hat{\Lambda}_1)^{-1}$ and $\hat{M}_{ff}^1 = \hat{\Lambda}_1\hat{\Lambda}_1'$. This new estimator satisfies IC1. For IC2, let $\hat{\Lambda}^2 = \hat{\Lambda}(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1/2}$ and $\hat{M}_{ff}^2 = \frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}$. Then this estimator satisfies IC2. For IC4, $M_{ff} = I_r$ is known. Let $\Lambda_1' = \mathcal{Q}\mathcal{R}$ be the QR decomposition of $\Lambda_1'$ with $\mathcal{Q}$ an orthogonal matrix and $\mathcal{R}$ an upper triangular matrix. Define $\hat{\Lambda}^4 = \hat{\Lambda}\mathcal{Q}$. Then $\hat{\Lambda}^4$ satisfies IC4. Finally, consider IC5. Let $\mathcal{W}$ be the diagonal matrix with its diagonal elements the same as the first $r \times r$ block of $\hat{\Lambda}^4$. Let $\hat{\Lambda}^5 = \hat{\Lambda}^4\mathcal{W}^{-1}$ and $\hat{M}_{ff}^5 = \mathcal{W}\mathcal{W}'$. Then IC5 is satisfied.

We now consider the finite sample properties of the MLE. Data are generated according to $z_{it} = \lambda_i' f_t + e_{it}$ with $r = 2$, where $\lambda_i$, $f_t$ are i.i.d. $\mathcal{N}(0, I_2)$ and $e_{it}$ follows $\mathcal{N}(0, \sigma_i^2)$ with $\sigma_i^2 = 0.1 + 10 \times U_i$, and $U_i$ are i.i.d. uniform on [0, 1]. Adding 0.1 to the variance avoids near-zero values. We consider combinations of $T = 30, 50, 100$ and $N = 10, 30, 50, 100, 150$. Estimators under different identification conditions only differ up to a rotation matrix, so only IC3 will be considered. We also compute the principal components (PC) estimator for comparison. To measure the accuracy between $\hat{\Lambda}$ and $\Lambda$ (both are $N \times 2$), we compute the second (the smallest nonzero) canonical correlation between them. Canonical correlation is widely used as a measure of goodness-of-fit in factor analysis; see, for example, [14] and [17]. Similarly, we also compute the second canonical correlation between $\hat{F}$ and $F$. For the estimated variances, we calculate the squared correlation between $\mathrm{diag}(\hat{\Sigma}_{ee})$ and $\mathrm{diag}(\Sigma_{ee})$. The corresponding values for the principal components estimators are also computed. Table 2 reports the average canonical correlations based on 5000 repetitions for each $(N, T)$ combination.

TABLE 2
*The performance of MLE and PC*

| | | **MLE** | | | **PC** | | |
|---|---|---|---|---|---|---|---|
| $N$ | $T$ | $\Lambda$ | $F$ | $\Sigma_{ee}$ | $\Lambda$ | $F$ | $\Sigma_{ee}$ |
| 10 | 30 | 0.4818 | 0.3473 | 0.8432 | 0.4058 | 0.2744 | 0.7991 |
| 30 | 30 | 0.7276 | 0.7995 | 0.9273 | 0.6391 | 0.6450 | 0.9223 |
| 50 | 30 | 0.7676 | 0.8973 | 0.9303 | 0.7221 | 0.7953 | 0.9302 |
| 100 | 30 | 0.7874 | 0.9555 | 0.9308 | 0.7679 | 0.9006 | 0.9312 |
| 150 | 30 | 0.7941 | 0.9719 | 0.9310 | 0.7823 | 0.9347 | 0.9315 |
| 10 | 50 | 0.6080 | 0.4153 | 0.8951 | 0.4875 | 0.2975 | 0.8187 |
| 30 | 50 | 0.8383 | 0.8407 | 0.9583 | 0.7751 | 0.7113 | 0.9499 |
| 50 | 50 | 0.8589 | 0.9161 | 0.9590 | 0.8306 | 0.8341 | 0.9569 |
| 100 | 50 | 0.8722 | 0.9624 | 0.9592 | 0.8613 | 0.9198 | 0.9591 |
| 150 | 50 | 0.8764 | 0.9764 | 0.9592 | 0.8697 | 0.9475 | 0.9593 |
| 10 | 100 | 0.7563 | 0.4939 | 0.9448 | 0.5878 | 0.3298 | 0.8345 |
| 30 | 100 | 0.9182 | 0.8614 | 0.9793 | 0.8789 | 0.7519 | 0.9700 |
| 50 | 100 | 0.9292 | 0.9245 | 0.9798 | 0.9135 | 0.8572 | 0.9770 |
| 100 | 100 | 0.9362 | 0.9668 | 0.9798 | 0.9305 | 0.9308 | 0.9792 |
| 150 | 100 | 0.9383 | 0.9788 | 0.9799 | 0.9349 | 0.9545 | 0.9798 |

The results suggest that the precision of $\hat{\Lambda}$ is closely tied to the size of $T$ and the precision of $\hat{F}$ is tied to $N$. This is consistent with the theory. For all $(N, T)$ combinations, the MLE dominates PC. The domination becomes less important for $N \geq 50$ and $T \geq 50$ for estimating factor loadings. But for small $N$, no matter how large is $T$, MLE noticeably outperforms PC. For the estimated factors, there is still noticeable outperformance even under large $N$ and $T$. These are all consistent with the theory.

**9. Conclusion.** In this paper we have developed an inferential theory for factor models of high dimension. We study the maximum likelihood estimator under five different sets of identification restrictions. Consistency, rate of convergence and the limiting distributions are derived. Unlike the principal component methods, the estimators are shown to be efficient under the model assumptions. While both the factor loadings and factors are treated as parameters (nonrandom), the key to efficiency is not to simultaneously estimate both the factor loadings and the factors. If $N$ is relatively small compared with $T$, the efficient approach is to estimate the individual factor loadings ($\lambda_i$) and the sample moment of the factor scores ($f_t$), not the individual scores. The sample moment contains only $r(r+1)/2$ unknown elements. If the factor scores $f_t$ are of interest, they can be estimated by the generalized least squares in a separate stage. The estimated factor scores are also shown to be efficient under the model assumptions. The opposite procedure should be adopted if $N$ is much larger than $T$. In the latter case, we estimate the individual factor scores and the sample moment of the factor loadings. If $N$ and $T$ are comparable, the choice of procedures boils down which heteroscedasticity, cross-sectional dimension or the time dimension, is the object of interest. The paper also provides a novel approach to consistency in the presence of a large and increasing number of parameters.

## APPENDIX: PROOF OF PROPOSITION 5.1

The following notation will be used throughout:

$$\begin{aligned}
\hat{H} &= (\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1},\\
\hat{H}_N &= N \cdot \hat{H} = (N^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1},\\
\hat{G} &= (\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1},\\
\hat{G}_N &= N \cdot \hat{G},\\
\xi_t &= (e_{1t}, e_{2t}, \ldots, e_{rt})'.
\end{aligned}$$

From $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$, we have $\hat{H} = \hat{G}(I - \hat{M}_{ff}^{-1}\hat{G})^{-1}$. From $\Sigma_{zz} = \Lambda M_{ff}\Lambda' + \Sigma_{ee}$, we have

$$\Sigma_{zz}^{-1} = \Sigma_{ee}^{-1} - \Sigma_{ee}^{-1}\Lambda(M_{ff}^{-1} + \Lambda'\Sigma_{ee}^{-1}\Lambda)^{-1}\Lambda'\Sigma_{ee}^{-1}. \tag{A.1}$$

It follows

$$\begin{aligned}\hat{\Lambda}'\hat{\Sigma}_{zz}^{-1} &= \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1} - \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1} \\ &= \hat{M}_{ff}^{-1}\hat{G}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}.\end{aligned} \tag{A.2}$$

To prove Proposition 5.1, we use a superscript "*" to denote the true parameters, for example, $\Lambda^*$, $\Sigma_{ee}^*$, $f_t^*$, etc. The variables without the superscript "*" denote the function arguments (input variables) in the likelihood function.

Let $\theta = (\lambda_1, \dots, \lambda_n, \sigma_1^2, \dots, \sigma_n^2, M_{ff})$ and let $\Theta$ be a parameter set such that $C^{-2} \leq \sigma_t^2 \leq C^2$, $M_{ff}$ is positive definite matrices with elements bounded. We assume $\theta^* = (\lambda_1^*, \dots, \lambda_n^*, \sigma_1^{2*}, \dots, \sigma_n^{2*}\ , M_{ff}^*)$ is an interior point of $\Theta$. For simplicity, we also write $\theta = (\Lambda, \Sigma_{ee}, M_{ff})$ and $\theta^* = (\Lambda^*, \Sigma_{ee}^*, M_{ff}^*)$.

PROOF OF PROPOSITION 5.1. The centered likelihood function can be written as

$$L(\theta) = \overline{L}(\theta) + R(\theta),$$

where

$$\overline{L}(\theta) = -\frac{1}{N}\ln|\Sigma_{zz}| - \frac{1}{N}\operatorname{tr}(\Sigma_{zz}(\theta^*)\Sigma_{zz}^{-1}) + 1 + \frac{1}{N}\ln|\Sigma(\theta^*)|$$

and $R(\theta) = -\frac{1}{N}\operatorname{tr}((M_{zz} - \Sigma_{zz}(\theta^*))\Sigma_{zz}^{-1})$. Note that $1 + \frac{1}{N}\sum_{i=1}^N \ln|\Sigma(\theta^*)|$ does not depend on any unknown parameters and is for the purpose of centering.

Lemma A.2 [6] implies that $\sup_\theta |R(\theta)| = o_p(1)$. In particular, we have $|R(\hat{\theta})| = o_p(1)$ and $|R(\theta^*)| = o_p(1)$. So $|R(\theta^*) - R(\hat{\theta})| = o_p(1)$. Since $\hat{\theta}$ maximizes $L(\theta)$, it follows $\overline{L}(\hat{\theta}) + R(\hat{\theta})) \geq \overline{L}(\theta^*) + R(\theta^*)$. Hence we have $\overline{L}(\hat{\theta}) \geq \overline{L}(\theta^*) + R(\theta^*) - R(\hat{\theta}) \geq \overline{L}(\theta^*) - |o_p(1)|$. However, the function $\overline{L}(\theta)$ achieves its maximum at $\theta^*$, so $\overline{L}(\hat{\theta}) \leq \overline{L}(\theta^*)$. Since $\overline{L}(\theta^*)$ is normalized to zero, we have $\overline{L}(\hat{\theta}) \geq -|o_p(1)|$ and $\overline{L}(\hat{\theta}) \leq 0$. It follows that $\overline{L}(\hat{\theta}) = o_p(1)$.

Notice $|\Sigma_{zz}| = |\Sigma_{ee}| \cdot |I_r + M_{ff}\Lambda'\Sigma_{ee}^{-1}\Lambda|$. But $|I_r + M_{ff}\Lambda'\Sigma_{ee}^{-1}\Lambda| = O(N)$. Similarly $|\Sigma_{zz}(\theta^*)| = |\Sigma_{ee}^*| \cdot |I_r + M_{ff}^*\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^*|$, thus uniformly on $\Theta$,

$$-\frac{1}{N}\ln|\Sigma_{zz}| + \frac{1}{N}\ln|\Sigma_{zz}(\theta^*)| = -\frac{1}{N}\ln|\Sigma_{ee}| + \frac{1}{N}\ln|\Sigma_{ee}^*| + O\left(\frac{\ln(N)}{N}\right).$$

Next, from $\Sigma_{zz}(\theta^*) = \Lambda^* M_{ff}^* \Lambda^{*\prime} + \Sigma_{ee}^*$, we have $\Sigma_{zz}(\theta^*)\Sigma_{zz}^{-1} = \Lambda^* M_{ff}^* \Lambda^{*\prime} \times \Sigma_{zz}^{-1} + \Sigma_{ee}^*\Sigma_{zz}^{-1}$. Using the formula for $\Sigma_{zz}^{-1}$, we have $\operatorname{tr}(\Sigma_{ee}^*\Sigma_{zz}^{-1}) = \operatorname{tr}(\Sigma_{ee}^*\Sigma_{ee}^{-1}) + O(1)$, because $\operatorname{tr}[\Sigma_{ee}^*\Sigma_{ee}^{-1}\Lambda(M_{ff}^{-1} + \Lambda'\Sigma_{ee}^{-1}\Lambda)^{-1}\Lambda'\Sigma_{ee}^{-1}] = O(1)$. The latter follows since the matrix in the square bracket is bounded in norm by $C^4\|\Lambda'\Sigma_{ee}^{-1}\Lambda \times (M_{ff}^{-1} + \Lambda'\Sigma_{ee}^{-1}\Lambda)^{-1}\| \leq C^4\|I_r\|$ due to the bound on $\sigma_i^2$ and $\sigma_i^{*2}$. Thus divided by $N$, we have

$$\frac{1}{N}\operatorname{tr}[\Sigma_{zz}(\theta^*)\Sigma_{zz}^{-1}] = \frac{1}{N}\operatorname{tr}[\Lambda^* M_{ff}^* \Lambda^{*\prime}\Sigma_{zz}^{-1}] + \frac{1}{N}\operatorname{tr}(\Sigma_{ee}^*\Sigma_{ee}^{-1}) + O\left(\frac{1}{N}\right).$$

Notice $\ln|\Sigma_{ee}| = \sum_{i=1}^{N} \ln \sigma_i^2$ and $\mathrm{tr}(\Sigma_{ee}^{*}\Sigma_{ee}^{-1}) = \sum_{i=1}^{N} \sigma_i^{*2}/\sigma_i^2$; we have proved that

$$\overline{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(\ln \sigma_i^2 + \frac{\sigma_i^{*2}}{\sigma_i^2} - 1 - \ln \sigma_i^{*2}\right) - \frac{1}{N}\mathrm{tr}(\Lambda^* M_{ff}^* \Lambda^{*\prime}\Sigma_{zz}^{-1}) + O\left(\frac{\ln N}{N}\right)$$

uniformly on $\Theta$. By $\overline{L}(\hat{\theta}) = o_p(1)$, it follows that

$$-\frac{1}{N}\sum_{i=1}^{N}\left(\ln \hat{\sigma}_i^2 + \frac{\sigma_i^{*2}}{\hat{\sigma}_i^2} - 1 - \ln \sigma_i^{*2}\right) - \frac{1}{N}\mathrm{tr}(\Lambda^* M_{ff}^* \Lambda^{*\prime}\hat{\Sigma}_{zz}^{-1}) \xrightarrow{p} 0.$$

A key observation is that both terms are nonpositive; it follows

$$\text{(A.3)} \qquad \frac{1}{N}\sum_{i=1}^{N}\left(\ln \hat{\sigma}_i^2 + \frac{\sigma_i^{*2}}{\hat{\sigma}_i^2} - 1 - \ln \sigma_i^{*2}\right) \xrightarrow{p} 0,$$

$$\text{(A.4)} \qquad \frac{1}{N}\mathrm{tr}(\Lambda^* M_{ff}^* \Lambda^{*\prime}\hat{\Sigma}_{zz}^{-1}) \xrightarrow{p} 0.$$

Consider the function $f(x) = \ln x + \frac{\sigma_i^{*2}}{x} - \ln \sigma_i^{*2} - 1$. Given that $0 < C^{-2} \le \sigma_i^2 \le C^2 < \infty$ for $C > 1$, for any $x \in [C^{-2}, C^2]$, there exists a constant $b$ (e.g., take $b = \frac{1}{4C^4}$), such that $f(x) \ge b(x - \sigma_i^{*2})^2$. It follows

$$o_p(1) = \frac{1}{N}\sum_{i=1}^{N}\left(\ln \hat{\sigma}_i^2 + \frac{\sigma_i^{*2}}{\hat{\sigma}_i^2} - 1 - \ln \sigma_i^{*2}\right) \ge b\frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i^2 - \sigma_i^{*2})^2.$$

The above implies

$$\text{(A.5)} \qquad \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i^2 - \sigma_i^{*2})^2 \xrightarrow{p} 0.$$

Now we turn to (A.4). By (A.1), we have

$$\begin{aligned}
&\frac{1}{N}\mathrm{tr}(\Lambda^* M_{ff}^* \Lambda^{*\prime}\hat{\Sigma}_{zz}^{-1}) \\
&\qquad = \frac{1}{N}\mathrm{tr}\big(M_{ff}^*\Lambda^{*\prime}[\hat{\Sigma}_{ee}^{-1} - \hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}]\Lambda^*\big).
\end{aligned}$$

From $(\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1} = (\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1} - (\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{M}_{ff}^{-1}(\hat{M}_{ff}^{-1} + \hat{\Lambda}' \times \hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}$, we obtain

$$\begin{aligned}
&\frac{1}{N}\mathrm{tr}(\Lambda^* M_{ff}^* \Lambda^{*\prime}\hat{\Sigma}_{zz}^{-1}) \\
&\quad = \frac{1}{N}\mathrm{tr}[M_{ff}^*\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^* - M_{ff}^*\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^*] \\
&\qquad + \mathrm{tr}[M_{ff}^*\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{M}_{ff}^{-1}(\hat{M}_{ff}^{-1} + \hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^*].
\end{aligned}$$

Both expressions are nonnegative; by (A.4), we must have

$$\frac{1}{N}\operatorname{tr}\bigl(M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*}-M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}\bigr)\xrightarrow{p}0 \tag{A.6}$$

and

$$\begin{aligned}&\frac{1}{N}\operatorname{tr}\bigl(M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\\&\qquad\times\hat{M}_{ff}^{-1}(\hat{M}_{ff}^{-1}+\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}\bigr)\xrightarrow{p}0.\end{aligned}\tag{A.7}$$

By (A.5) and Lemma A.4 [6], $\frac{1}{N}\operatorname{tr}(M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*})\xrightarrow{p}C^{*}>0$, say. From (A.6)

$$\frac{1}{N}\operatorname{tr}(M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*})\xrightarrow{p}C^{*}>0$$

with the same $C^{*}$. The preceding result and (A.7) imply

$$\hat{M}_{ff}^{-1}(\hat{M}_{ff}^{-1}+\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}=o_p(1).$$

By assumption, we confine $M_{ff}$ on a compact set, that is, $\hat{M}_{ff}=O_p(1)$. By the definition of $\hat{G}$, we have $\hat{G}=o_p(1)$. From $\hat{H}=\hat{G}(I-\hat{M}_{ff}^{-1}\hat{G})^{-1}$, we have $\hat{H}=o_p(1)$. We obtain the following result:

$$\hat{G}=o_p(1);\qquad \hat{H}=o_p(1).\tag{A.8}$$

The matrix on the left-hand side of (A.6) is semi-positive definite and is finite dimensional ($r\times r$), its trace is $o_p(1)$ if and only if every entry is $o_p(1)$. Thus we have

$$\frac{1}{N}\bigl(M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*}-M_{ff}^{*}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}\bigr)\xrightarrow{p}0.$$

Pre-multiplying both sides by $M_{ff}^{*-1}$ gives

$$\frac{1}{N}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*}-\frac{1}{N}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}\xrightarrow{p}0.\tag{A.9}$$

The second term on the left-hand side can be rewritten as

$$[\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}]\biggl(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\biggr)[(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}].$$

Let $A\equiv(\hat{\Lambda}-\Lambda^{*})'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}$, where $\hat{H}=(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}$. It follows that $\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\times\hat{\Lambda}(\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}=(I_r-A)$. So (A.9) is equivalent to

$$\frac{1}{N}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*}-(I_r-A)\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(I_r-A)'\xrightarrow{p}0.$$

However, $\frac{1}{N}\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\Lambda^{*} = \frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^{*} + o_p(1)$ by Lemma A.4 [6] and (A.5), thus

$$\frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^{*} - (I_r - A)\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}(I_r - A)' \xrightarrow{p} 0. \tag{A.10}$$

Because the first term is of full rank in the limit, the second term is also of full rank. This implies that $I_r - A$ in the limit is of full rank.

Meanwhile, equation (A.9) can be expressed alternatively as

$$\begin{aligned}&\frac{1}{N}(\hat{\Lambda} - \Lambda^{*})'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^{*})\\ &\qquad - \frac{1}{N}(\hat{\Lambda} - \Lambda^{*})'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\left(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\right)^{-1}\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^{*}) \xrightarrow{p} 0,\end{aligned}$$

which can also be written as, in terms of $A$,

$$\frac{1}{N}(\hat{\Lambda} - \Lambda^{*})'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^{*}) - A\left(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\right)A' \xrightarrow{p} 0. \tag{A.11}$$

Both (A.10) and (A.11) will be useful in establishing consistency.

We now make use of the first-order conditions. The first-order condition (2.7), by (A.2), can be simplified as $\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(M_{zz} - \hat{\Sigma}_{zz}) = 0$. This gives

$$\begin{aligned}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Bigg(&\Lambda^{*}M_{ff}^{*}\Lambda^{*\prime} + \Lambda^{*}\frac{1}{T}\sum_{t=1}^{T}f_t^{*}e_t' + \frac{1}{T}\sum_{t=1}^{T}e_t f_t^{*\prime}\Lambda^{*\prime}\\ &+ \frac{1}{T}\sum_{t=1}^{T}(e_t e_t' - \Sigma_{ee}^{*}) + \Sigma_{ee}^{*} - \hat{\Lambda}\hat{M}_{ff}\hat{\Lambda}' - \hat{\Sigma}_{ee}\Bigg) = 0.\end{aligned}$$

For simplicity, we neglect the smaller-order term $\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\bar{e}\bar{e}'$. The $j$th column of the above equation can be written as (after some algebra),

$$\begin{aligned}\hat{\lambda}_j - \lambda_j^{*} = {}& -\hat{M}_{ff}^{-1}(\hat{M}_{ff} - M_{ff}^{*})\lambda_j^{*}\\ & - \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^{*})M_{ff}^{*}\lambda_j^{*}\\ & + \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^{*}\left(\frac{1}{T}\sum_{t=1}^{T}f_t^{*}e_{jt}\right)\\ & + \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}e_t f_t^{*\prime}\right)\lambda_j^{*} - \hat{M}_{ff}^{-1}\hat{H}\hat{\lambda}_j\frac{1}{\hat{\sigma}_j^2}(\hat{\sigma}_j^2 - \sigma_j^{*2})\\ & + \hat{M}_{ff}^{-1}\hat{H}\left(\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\hat{\lambda}_i\frac{1}{T}\sum_{t=1}^{T}[e_{it}e_{jt} - E(e_{it}e_{jt})]\right).\end{aligned} \tag{A.12}$$

Consider the first-order condition (2.9). By the method analogous to the one in deducing (A.12), we have

$$
\begin{aligned}
\hat{M}_{ff} - M^*_{ff} = &-\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*)M^*_{ff} - M^*_{ff}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H} \\
&+ \hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*)M^*_{ff}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H} \\
&+ \hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^*\left(\frac{1}{T}\sum_{t=1}^{T} f^*_t e'_t\right)\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H} \\
&+ \hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} e_t f^{*\prime}_t\right)\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H} \qquad \text{(A.13)} \\
&+ \hat{H}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\hat{\sigma}_i^2\hat{\sigma}_j^2}\hat{\lambda}_i\hat{\lambda}'_j\frac{1}{T}\sum_{t=1}^{T}[e_{it}e_{jt} - E(e_{it}e_{jt})]\right)\hat{H} \\
&- \hat{H}\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^4}\hat{\lambda}_i\hat{\lambda}'_i(\hat{\sigma}_i^2 - \sigma_i^{*2})\hat{H}.
\end{aligned}
$$

Substituting (A.13) into (A.12), we obtain

$$
\begin{aligned}
\hat{\lambda}_j - \lambda^*_j = &\hat{M}_{ff}^{-1}M^*_{ff}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}\lambda^*_j \\
&- \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*)M^*_{ff}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}\lambda^*_j \\
&- \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^*\left(\frac{1}{T}\sum_{t=1}^{T} f^*_t e'_t\right)\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}\lambda^*_j \\
&- \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} e_t f^{*\prime}_t\right)\Lambda^{*\prime}\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}\lambda^*_j \\
&- \hat{M}_{ff}^{-1}\hat{H}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\hat{\sigma}_i^2\hat{\sigma}_j^2}\hat{\lambda}_i\hat{\lambda}'_j\frac{1}{T}\sum_{t=1}^{T}[e_{it}e_{jt} - E(e_{it}e_{jt})]\right)\hat{H}\lambda^*_j \\
&+ \hat{M}_{ff}^{-1}\hat{H}\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^4}\hat{\lambda}_i\hat{\lambda}'_i(\hat{\sigma}_i^2 - \sigma_i^{*2})\hat{H}\lambda^*_j \\
&+ \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} e_t f^{*\prime}_t\right)\lambda^*_j \\
&+ \hat{M}_{ff}^{-1}\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\Lambda^*\left(\frac{1}{T}\sum_{t=1}^{T} f^*_t e_{jt}\right) \qquad \text{(A.14)}
\end{aligned}
$$

$$\begin{aligned}&+\hat{M}_{ff}^{-1}\hat{H}\left(\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}\hat{\lambda}_i\frac{1}{T}\sum_{t=1}^{T}[e_{it}e_{jt}-E(e_{it}e_{jt})]\right)\\&-\hat{M}_{ff}^{-1}\hat{H}\hat{\lambda}_j\frac{1}{\hat{\sigma}_j^2}(\hat{\sigma}_j^2-\sigma_j^{*2}).\end{aligned}$$

Consider (A.13). The fifth term of the right-hand side can be written as

$$\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}e_tf_t^{*\prime}\right)-\hat{H}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}e_tf_t^{*\prime}\right)A,$$

where $A\equiv(\hat{\Lambda}-\Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\hat{H}$ is defined following (A.9). The first term is $\|N^{1/2}\hat{H}^{1/2}\|\cdot O_p(T^{-1/2})$ by Lemma A.3(b) [6] and the second term is

$$A\cdot\|N^{1/2}\hat{H}^{1/2}\|\cdot O_p(T^{-1/2}).$$

The fourth term is the transpose of the fifth. The sixth term is given in Lemma A.3(d). The seventh term is bounded by $\|\hat{H}\|\cdot\|\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^4}\hat{\lambda}_i\hat{\lambda}_i'(\hat{\sigma}_i^2-\sigma_i^2)\hat{H}\|$. The term $\|\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^4}\hat{\lambda}_i\hat{\lambda}_i'(\hat{\sigma}_i^2-\sigma_i^2)\hat{H}\|$ is bounded by $2C^4\sqrt{r}$ due to $|\frac{1}{\hat{\sigma}_i^2}(\hat{\sigma}_i^2-\sigma_i^{*2})|\le 2C^4$ because of the boundedness of $\hat{\sigma}_i^2,\sigma_i^{*2}$. So the seventh term is $o_p(1)$ by (A.8). Given these results, in terms of $A$, equation (A.13) can be rewritten as

$$\begin{aligned}\hat{M}_{ff}-M_{ff}^*={}&-A'M_{ff}^*-M_{ff}^*A+A'M_{ff}^*A+\|N^{1/2}\hat{H}^{1/2}\|\cdot O_p(T^{-1/2})\\&-A\cdot\|N^{1/2}\hat{H}^{1/2}\|\cdot O_p(T^{-1/2})\\&+\|N^{1/2}\hat{H}^{1/2}\|^2\cdot O_p(T^{-1/2})+o_p(1).\end{aligned}\tag{A.15}$$

However, by the definition of $\hat{H}$, $N\hat{H}=(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}$. Equation (A.10) yields $(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda})^{-1}=(I_r-A)'(\frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^*)^{-1}(I_r-A)+o_p(\|I_r-A\|^2)$. So we have

$$\begin{aligned}\|N^{1/2}\hat{H}^{1/2}\|^2&=\operatorname{tr}[N\hat{H}]\\&=\operatorname{tr}\left[(I_r-A)'\left(\frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^*\right)^{-1}(I_r-A)+o_p(\|I_r-A\|^2)\right].\end{aligned}$$

The right-hand side is at most $O_p(A^2)$, implying that $\|N^{1/2}\hat{H}^{1/2}\|=O_p(A)$.

Given the above result, we argue that the matrix $A$ must be stochastically bounded. First, notice that the left-hand side of (A.15) is stochastically bounded by Assumption D. So if $A$ is not stochastically bounded, the right-hand side is dominated by $A'M_{ff}^*A$ in view of $\|N^{1/2}\hat{H}^{1/2}\|=O_p(A)$, But $A'M_{ff}^*A$ will be unbounded since $M_{ff}^*$ is positive definite. A contradiction is obtained. Thus $A=O_p(1)$; it follows that $\|N^{1/2}\hat{H}^{1/2}\|=O_p(A)=O_p(1)$. Given this result, we have

$$\hat{M}_{ff}-M_{ff}^*=-A'M_{ff}^*-M_{ff}^*A+A'M_{ff}^*A+o_p(1).\tag{A.16}$$

Next consider (A.14). The seventh to ninth terms of the right-hand side of (A.14) are all $o_p(1)$ by Lemma A.3 [6] and $\|N^{1/2}\hat{H}^{1/2}\| = O_p(1)$. The last term is bounded by $2C^3\|\hat{M}_{ff}^{-1}\| \cdot \|\frac{1}{\hat{\sigma}_j}\hat{\lambda}_j\hat{H}^{1/2}\| \cdot \|H^{1/2}\|$. Since $\sum_{j=1}^N \|\frac{1}{\hat{\sigma}_j}\hat{\lambda}_j\hat{H}^{1/2}\|^2 = r$, $\|\hat{M}_{ff}^{-1}\| = O_p(1)$ by Assumption D and $\hat{H} = o_p(1)$ by (A.8), it follows that the last term is $o_p(1)$. The third and fourth terms are similar to the fourth and fifth terms in (A.13), and hence are $o_p(1)$ due to $A = O_p(1)$, $\|\lambda_j^*\| \le C$ for all $j$ and $\hat{M}_{ff}$ being bounded. Using the same arguments for (A.16), we have

$$\hat{\lambda}_j - \lambda_j^* = \hat{M}_{ff}^{-1}M_{ff}^*A\lambda_j^* - \hat{M}_{ff}^{-1}A'M_{ff}^*A\lambda_j^* + o_p(1). \tag{A.17}$$

We next prove consistency by using the identification conditions.

*Under* IC1: Since the identification condition is $\Lambda^* = [I_r, \Lambda_2^{*\prime}]'$ and $\hat{\Lambda} = [I_r, \hat{\Lambda}_2']'$, the first $r \times r$ upper block of $\Lambda^*$ is the same as that of $\hat{\Lambda}$, that is, $[\lambda_1^*, \lambda_2^*, \dots, \lambda_r^*] = [\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_r] = I_r$. By (A.17) with $\hat{\lambda}_j - \lambda_j^* = 0$, $j = 1, 2, \dots, r$,

$$M_{ff}^*A - A'M_{ff}^*A \xrightarrow{p} 0.$$

We now attach a subscript to matrix $A$ to signify which identification condition is used. For IC$k$ ($k = 1, 2, \dots, 5$), we use $A_k$ to denote the corresponding $A$. So the above equation implies $M_{ff}^*A_1 - A_1'M_{ff}^*A_1 \xrightarrow{p} 0$. Taking transpose, we have $A_1'M_{ff}^* - A_1'M_{ff}^*A_1 \xrightarrow{p} 0$. Thus $M_{ff}^*A_1 - A_1'M_{ff}^* \xrightarrow{p} 0$. Post-multiplying $A_1$, we obtain $M_{ff}^*A_1^2 - A_1'M_{ff}^*A_1 \xrightarrow{p} 0$. But we also have $M_{ff}^*A_1 - A_1'M_{ff}^*A_1 \xrightarrow{p} 0$. Thus $M_{ff}^*A_1^2 - M_{ff}^*A_1 \xrightarrow{p} 0$. Since $M_{ff}^*$ is positive definite, we have $A_1(I_r - A_1) \xrightarrow{p} 0$. Since we have proved that $I_r - A_1$ converges in probability to a nonsingular matrix, it follows that $A_1 \xrightarrow{p} 0$.

From (A.11) and $A_1 \xrightarrow{p} 0$, we obtain $\frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) \xrightarrow{p} 0$, which is equivalent to (5.1a). From (A.16) and $A_1 \xrightarrow{p} 0$, we obtain $\hat{M}_{ff} - M_{ff}^* \xrightarrow{p} 0$, which is (5.1c). This proves Proposition 5.1 under IC1.

*Under* IC2: From the identification condition $\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda} = \frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^* = I_r$, by adding and subtracting terms, we have the identity

$$\begin{aligned} &\frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda} + \frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) \\ &\quad = -\frac{1}{N}\Lambda^{*\prime}(\hat{\Sigma}_{ee}^{-1} - \Sigma_{ee}^{*-1})\Lambda^* + \frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*). \end{aligned} \tag{A.18}$$

By (A.5) and Lemma A.4 [6], the term $\frac{1}{N}\Lambda^{*\prime}(\hat{\Sigma}_{ee}^{-1} - \Sigma_{ee}^{*-1})\Lambda^*$ is $o_p(1)$. Thus

$$\frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda} + \frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) - \frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) \xrightarrow{p} 0.$$

The above can be written in terms of matrix $A$ (i.e., $A_2$ under IC2),

$$A_2 + A_2' - \frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) \xrightarrow{p} 0.$$

With $\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda} = I_r$, (A.11) implies $A_2 A_2' - \frac{1}{N}(\hat{\Lambda} - \Lambda^*)'\hat{\Sigma}_{ee}^{-1}(\hat{\Lambda} - \Lambda^*) \xrightarrow{p} 0$. These two results imply $A_2 + A_2' - A_2 A_2' \xrightarrow{p} 0$, which is equivalent to

$$(A_2 - I_r)(A_2 - I_r)' - I_r \xrightarrow{p} 0.$$

However, (A.16) is equivalent to

$$\hat{M}_{ff} - (A_2 - I_r)' M_{ff}^* (A_2 - I_r) \xrightarrow{p} 0.$$

Under IC2, $M_{ff}^*$ is a diagonal matrix with distinct elements. Also, $\hat{M}_{ff}$ is a diagonal matrix by restriction. Applying Lemma A.1 [6] with $Q = A_2 - I_r$, $V = M_{ff}^*$, and $D = \hat{M}_{ff}$, we conclude that $Q$ and thus $A_2 - I_r$ converge in probability to a diagonal matrix with elements being either $-1$ or 1. Equivalently, $A_2$ converges to a diagonal matrix with diagonal elements being either 0 or 2. By assuming that $\hat{\Lambda}$ and $\Lambda^*$ have the same column signs, we rule out 2 as the diagonal element So $A_2 = o_p(1)$. The rest of the proof is identical to IC1, implying Proposition 5.1 under IC2.

*Under* IC3: IC3 requires $\hat{M}_{ff} = M_{ff}^* = I_r$, and so by (A.16), $(A_3 - I_r)(A_3 - I_r)' - I_r \xrightarrow{p} 0$. From (A.10),

$$\frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^* - (A_3 - I_r)'\left(\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}\right)(A_3 - I_r) \xrightarrow{p} 0.$$

Under IC3, $\frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^*$ is diagonal with distinct elements, and $\frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}$ is also diagonal by estimation restriction. The latter matrix has distinct diagonal elements with probability 1. It follows that $A_3 - I_r$ converges in probability to a diagonal matrix with diagonal elements either 1 or $-1$ by Lemma A.1 [6] applied with $Q = (A_3 - I_r)'$, $V = \frac{1}{N}\hat{\Lambda}'\hat{\Sigma}_{ee}^{-1}\hat{\Lambda}$, and $D = \frac{1}{N}\Lambda^{*\prime}\Sigma_{ee}^{*-1}\Lambda^*$. The remaining proof is identical to that of IC2 and hence omitted. So we have proved Proposition 5.1 under IC3.

*Under* IC4: By the identification condition, both $\Lambda_1^*$ and $\hat{\Lambda}_1$ are lower triangular matrices, where $\Lambda_1$ is first $r \times r$ submatrix of $\Lambda$. Consider the first $r$ equations of (A.17),

$$\hat{\Lambda}_1' - \Lambda_1^{*\prime} - \hat{M}_{ff}^{-1}(M_{ff}^* A_4 - A_4' M_{ff}^* A_4)\Lambda_1^{*\prime} \xrightarrow{p} 0.$$

By (A.16), we have $\hat{M}_{ff} - M_{ff}^* + A_4' M_{ff}^* + M_{ff}^* A_4 - A_4' M_{ff}^* A_4 \xrightarrow{p} 0$, which can be rewritten as

$$\hat{M}_{ff} - (I_r - A_4)' M_{ff}^* (I_r - A_4) \xrightarrow{p} 0.$$

The above two equations imply

$$(I_r - A_4)' M_{ff}^* (I_r - A_4)(\hat{\Lambda}_1' - \Lambda_1^{*\prime}) - (I_r - A_4)' M_{ff}^* A_4 \Lambda_1^{*\prime} \xrightarrow{p} 0. \tag{A.19}$$

Since both $\hat{M}_{ff}$ and $M_{ff}$ are of full rank, $\hat{M}_{ff} - (I_r - A_4)' M^*_{ff}(I_r - A_4) \xrightarrow{p} 0$ implies $I_r - A_4$ is of full rank. Pre-multiplying $[(I_r - A_4)' M^*_{ff}]^{-1}$, we obtain

$$(I_r - A_4)\hat{\Lambda}_1' - \Lambda_1^{*\prime} \xrightarrow{p} 0. \tag{A.20}$$

Since both $\hat{\Lambda}_1$ and $\Lambda_1^*$ are lower triangular with diagonal elements all 1, both matrices are invertible. It follows that $I_r - A_4 - \Lambda_1^{*\prime}(\hat{\Lambda}_1')^{-1} \xrightarrow{p} 0$. Since both $\hat{\Lambda}_1$ and $\Lambda_1^*$ are lower triangular, we have $I_r - A_4$ converges to an upper triangular matrix. However, $\hat{M}_{ff}$ and $M^*_{ff}$ are both diagonal matrices and invertible. For $\hat{M}_{ff} - (I_r - A_4)' M^*_{ff}(I_r - A_4) \xrightarrow{p} 0$ to hold, given that $I_r - A_4$ is an upper triangular matrix, it implies that $I_r - A_4$ converges to a diagonal matrix. Because both $\hat{\Lambda}_1$ and $\Lambda_1^*$ are matrices with diagonal elements 1, and given the asymptotic diagonality of $A_4$, it follows by (A.20) that $I_r - A_4 \xrightarrow{p} I_r$. So we have $A_4 \xrightarrow{p} 0$. The remaining proof is the same as in IC1 and is omitted. This completes the proof for IC4.

*Under* IC5: Both $\hat{M}_{ff}$ and $M^*_{ff}$ are identity matrices; it follows from (A.16) that $(I_r - A_5)'(I_r - A_5) - I_r \xrightarrow{p} 0$. The derivation of (A.20) only involves the full rank of $I_r - A$, so it is applicable for IC5, that is, $(I_r - A_5)\hat{\Lambda}_1' - \Lambda_1' \xrightarrow{p} 0$. Since both $\hat{\Lambda}_1$ and $\Lambda_1$ are lower triangular and invertible, it follows that $I_r - A_5$ converges to an upper triangular matrix. Given this result and $(I_r - A_5)'(I_r - A_5) - I_r \xrightarrow{p} 0$, it follows that $A_5$ converges to a diagonal matrix with diagonal elements either 0 or 2. By assuming that the column signs of $\hat{\Lambda}$ and $\Lambda^*$ are the same, we have $A_5 \xrightarrow{p} 0$. The remaining proof is the same as in IC1 and is omitted. So we have proved Proposition 5.1 under IC5. This completes the proof of Proposition 5.1. □

The proofs for other results are provided in the supplement [6].

**Acknowledgments.** The authors thank two anonymous referees, an Associate Editor and an Editor for constructive comments.

## SUPPLEMENTARY MATERIAL

**Supplement to "Statistical analysis of factor models of high dimension"** (DOI: 10.1214/11-AOS966SUPP; .pdf). In this supplement we provide the detailed proofs for Theorems 5.1–5.4 and 6.1. We also give a simple and direct proof that the EM solutions satisfy the first order conditions. Remarks are given on how to make use of matrix properties to write a faster computer program.

## REFERENCES

[1] Amemiya, Y., Fuller, W. A. and Pantula, S. G. (1987). The asymptotic distributions of some estimators for a factor analysis model. *J. Multivariate Anal.* **22** 51–64. MR0890881

[2] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[3] ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. MR0947576

[4] ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley. MR0084943

[5] BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

[6] BAI, J. and LI, K. (2012). Supplement to "Statistical analysis of factor models of high dimension." DOI:10.1214/11-AOS966SUPP.

[7] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

[8] BAI, J. and NG, S. (2010). Principal components estimation and identification of the factors. Unpublished manuscript, Columbia Univ.

[9] BREITUNG, J. and TENHOFEN, J. (2008). GLS estimation of dynamic factor models. Working paper, Univ. Bonn.

[10] CAMPBELL, J. Y., LO, A. W. and MACKINLAY, A. C. (1997). *The Econometrics of Financial Markets*. Princeton Univ. Press, Princeton, NJ.

[11] CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050

[12] CHOI, I. (2007). Efficient estimation of factor models. Working paper. Available at http://hompi.sogang.ac.kr/inchoi/workingpaper/efficient_estimation_in_choi_et1.pdf.

[13] CONNOR, G. and KORAJCZYK, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics* **21** 255–289.

[14] DOZ, C., GIANNONE, D. and REICHLIN, L. (2006). A quasi-maximum likelihood approach for large approximate dynamic factor models. Discussion Paper 5724, CEPR.

[15] FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econom. Statist.* **82** 540–554.

[16] GEWEKE, J. and ZHOU, G. (1996). Measuring the price of the arbitrage pricing theory. *The Review of Financial Studies* **9** 557–587.

[17] GOYAL, A., PERIGNON, C. and VILLA, C. (2008). How common are common return factors across the NYSE and Nasdaq? *Journal of Financial Economics* **90** 252–271.

[18] LAWLEY, D. N. and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. Elsevier, New York. MR0343471

[19] NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, *Vol. IV* (R. F. Engle and D. McFadden, eds.). *Handbooks in Economics* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

[20] ROSS, S. A. (1976). The arbitrage theory of capital asset pricing. *J. Econom. Theory* **13** 341–360. MR0429063

[21] RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. MR0668505

[22] STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271

[23] STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257

[24] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[25] WU, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

DEPARTMENT OF ECONOMICS
COLUMBIA UNIVERSITY
420 WEST 118TH STREET
NEW YORK, NEW YORK 10027
USA
E-MAIL: Jushan.bai@columbia.edu

DEPARTMENT OF ECONOMICS
SCHOOL OF ECONOMICS AND MANAGEMENT
TSINGHUA UNIVERSITY
AND
DEPARTMENT OF QUANTITATIVE ECONOMICS
SCHOOL OF INTERNATIONAL TRADE
AND ECONOMICS
UNIVERSITY OF INTERNATIONAL BUSINESS
AND ECONOMICS
BEIJING, 100084
CHINA
E-MAIL: likp.07@sem.tsinghua.edu.cn