Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models

Alexei Onatski*

Faculty of Economics, University of Cambridge.

September 30, 2014

Abstract

In this paper, we obtain asymptotic approximations to the squared error of the least squares estimator of the common component in large approximate factor models with possibly misspecified number of factors. The approximations are derived under both strong and weak factors asymptotics assuming that the cross-sectional and temporal dimensions of the data are comparable. We develop consistent estimators of these approximations and propose to use them for model comparison and for selection of the number of factors. We show that the estimators of the number of factors that minimize these loss estimators are asymptotically loss efficient in the sense of Shibata (1980), Li (1987), and Shao (1997).

JEL codes: C38, C52

Key words: misspecification, factor model, number of factors, loss efficiency.

*Faculty of Economics, Sidgwick Ave, Cambridge, CB3 9DD. Tel: +44 1223 335240. Fax: +44 1223 335475. E-mail: ao319@cam.ac.uk

1 Introduction

Empirical analyses of high-dimensional economic data often rely on approximate factor models estimated by the principal components method (see Stock and Watson (2011) for a recent survey of related literature). Many of these analyses intend to accurately estimate a low-dimensional common component of the data. For example, the interest may lie in the part of multi-national data that can be attributed to a common business cycle, as in Forni and Reichlin (2001), or in the decomposition of sectoral output growth rates into the common and idiosyncratic parts, as in Foerster et al (2011). Unfortunately, the estimation problem is complicated by the fact that the number of factors is typically unknown and is likely to be misspecified. This paper studies consequences of the misspecification for the squared error of the estimated common component.

Assuming that the cross-sectional and temporal dimensions of the data, n and T, are comparable, we derive asymptotic approximations to the squared error loss through the order $n^{-1} \sim T^{-1}$. We consider both strong and weak factors asymptotics. Under the latter, the asymptotic loss turns out to be minimized not necessarily at the true number of factors.

We develop estimators of the loss which are consistent under strong and under weak factors asymptotics, and propose to use them for model comparison and for selection of the number of factors. We show that estimators of the number of factors that minimize the proposed loss estimates are asymptotically loss efficient in the sense of Shibata (1980), Li (1987), and Shao (1997). The majority of recently proposed estimators of the number of factors, including the popular Bai and Ng (2002) estimators, are asymptotically loss efficient under the strong factors asymptotics, but not under the weak factors one.

The basic framework of our analysis is standard. We consider an approximate factor model

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

where X is an $n \times T$ matrix of data, Λ is an $n \times r$ matrix of factor loadings, F is a $T \times r$ matrix of factors and e is an $n \times T$ matrix of idiosyncratic terms. Throughout the paper, we will treat Λ and F as unknown parameters. Equivalently, our results can be thought of as conditional on the unobserved realizations of random Λ and F.

Suppose that we estimate the first p of the factors and the corresponding loadings by the least squares, and let us denote the estimates as $\hat{F}_{1:p}$ and $\hat{\Lambda}_{1:p}$, respectively. As is well known, $\hat{F}_{1:p}$ and $\hat{\Lambda}_{1:p}$ can equivalently be obtained by the principal components (PC) method. That is, the columns of $\hat{F}_{1:p}/\sqrt{T}$ are unit-length eigenvectors of $X'X$, and $\hat{\Lambda}_{1:p} = X\hat{F}_{1:p}/T$. In the special case where the idiosyncratic terms are i.i.d. $N(0,1)$, these are the maximum likelihood estimates subject to the normalization. Since we do not know the true value of r, p may be smaller, equal, or larger than r. We will say that the number of factors is misspecified if $p \neq r$.

We are interested in the effect of the misspecification on the quality of the PC estimate $\hat{\Lambda}_{1:p}\hat{F}'_{1:p}$ of the common component $\Lambda F'$ of the data. This quality is measured by the average (over time and cross-section) squared error

$$L_p = \operatorname{tr}\left[(\hat{\Lambda}_{1:p}\hat{F}'_{1:p} - \Lambda F')(\hat{\Lambda}_{1:p}\hat{F}'_{1:p} - \Lambda F')'\right] / (nT) \,. \tag{2}$$

Our interest in L_p is motivated by several reasons. First, accurate extraction of the common component is important in many applications. Second, in the special case where the idiosyncratic terms are i.i.d. $N(0,1)$, L_p is proportional to the Kullback-Leibler distance between the true model (1) and the factor model with factors $\hat{F}_{1:p}$ and loadings $\hat{\Lambda}_{1:p}$. Recall that the expected value of such a distance is usually approximated by Akaike's (1973) information criterion (AIC). In Section 3, we show that the AIC approximation does not hold in the large factor model setting, and propose a valid alternative.

Finally, loss functions similar to L_p are widely used in the context of linear regression models. For example, Mallows' (1973) "measure of adequacy for prediction" of linear regression model $Y = Z_{1:p}\beta_{1:p} + \varepsilon$ when the true model is $Y = Z\beta + u$ is given by $(\hat{Z}_{1:p}\hat{\beta}_{1:p} - Z\beta)'(\hat{Z}_{1:p}\hat{\beta}_{1:p} - Z\beta)$. The problems of prediction, model selection, and model averaging with this loss function were extensively studied by Phillips (1979), Kunitomo and Yamamoto (1985), Shao (1997), and Hansen (2007), to name just a few studies.

Since $\Lambda F'$ is unobserved, L_p can not be evaluated directly. In Section 2, we derive asymptotic approximations for L_p that are easy to analyze and estimate. Subsection 2.1 considers the standard strong factors asymptotic regime (Bai and Ng (2008)).

The strong factors asymptotics has been criticized by Boivin and Ng (2006), Heaton and Solo (2006), DeMol et al (2008), Onatski (2010, 2012), Kapetanios and Marcellino (2010), and Chudik et al (2011) for not providing accurate finite sample approximations in applications where the factors are moderately or weakly influential. Therefore, in Subsection 2.2 we derive asymptotic approximations for L_p using

Onatski's (2012) weak factors assumptions.

Using the derived asymptotic approximations, Section 3 develops four different estimators of L_p. All these estimators use a preliminary estimator $\hat{r}$ of the true number of factors r. Under the strong factors asymptotics, if $\hat{r} \xrightarrow{p} r$, all the corresponding loss estimators are consistent for L_p after a shift by a constant that does not depend on p.

Under the weak factors asymptotics, in general, no preliminary estimator $\hat{r}$ can consistently estimate r. As explained in Onatski (2012, p. 250), one can, instead, estimate the number q of theoretically detectable, or "effective", factors. If $\hat{r} \xrightarrow{p} q$, then two of the corresponding proposed loss estimators provide the asymptotic upper and lower bounds on the shifted loss. We show that the minimizers of these estimators bracket the actual loss minimizer with probability approaching one as n and T go to infinity. The other two loss estimators are consistent for the shifted loss when there is either no cross-sectional or no temporal correlation in the idiosyncratic terms. In these special cases, the number of factors that minimizes the corresponding estimator of the loss is consistent for the number of factors that minimizes the actual loss. The latter is not necessarily equal to the true number of factors r or to the "effective" number of factors q.

All the proposed loss estimators are simple functions of the eigenvalues of the sample covariance matrix. Monte Carlo exercises in Section 4 show that their quality is excellent when simulated factors are relatively strong. When the factors become weaker, the quality gradually deteriorates, but remains reasonably good in intermediate cases.

In Section 5, we provide an empirical example of model comparison based on our loss estimators. We compare a two- and a three-factor model of excess stock returns, and find that estimating the third factor leads to a loss deterioration for the monthly data covering the period from 2001 to 2012. That is, a PC estimate of the three-factor model provides a worse description of the undiversifiable risk portion of the excess returns than a PC estimate of the two-factor model. Interestingly, this loss-based ordering is reversed when we use the data from 1989 to 2000, which suggests a decrease in the signal-to-noise ratio in the more recent excess returns data.

Section 6 discusses possible extensions, establishes a connection with the literature on sparse models (see, for example, Belloni et al (2012)), and concludes. All proofs are given in the Appendix.

2 Asymptotic approximation for the loss

2.1 Strong factors asymptotics

In what follows, $\mu_i(M)$ denotes the i-th largest eigenvalue of a Hermitian matrix M. Further, $A_{\cdot j}$ and $A_{j\cdot}$ denote the j-th column and j-th row of a matrix A, respectively. We make the following assumptions.

A1 There exists a diagonal matrix D_n with elements $d_{1n} \geq d_{2n} \geq \ldots \geq d_{rn} > 0$ along the diagonal, such that $F'F/T = I_r$ and $\Lambda'\Lambda/n = D_n$.

This assumption is a convenient normalization. The only non-trivial constraint it implies is the requirement that $\operatorname{rank} F = r$ and $\operatorname{rank} \Lambda = r$.

A2 As $n \to \infty$, $\Lambda'\Lambda/n \to D$, where D is a diagonal matrix with decreasing elements $d_1 > d_2 > ... > d_r > 0$ along the diagonal.

Assumption A2 is sometimes called the factor pervasiveness assumption. It requires that the cumulative explanatory power of factors, measured by the diagonal elements of $\Lambda'\Lambda$, increases proportionally to n. The assumption is standard, but may be too strong in some applications. In Subsection 2.2, we consider an alternative assumption that allows $\Lambda'\Lambda$ to remain bounded as $n \to \infty$.

Let $n, T \to_c \infty$ denote the situation where both n and T diverge to infinity so that $n/T \to c \in (0, \infty)$. This asymptotic regime is particularly useful for the analysis of data with comparable cross-sectional and temporal dimensions, such as many financial and macroeconomic datasets. It also does not preclude situations where n/T is small or large as long as n/T does not go to zero or to infinity.

A3 As $n, T \to_c \infty$, (i) there exists $\varepsilon > 0$ such that $\Pr\left(\mathrm{tr}[ee']/(nT) > \varepsilon\right) \to 1$; (ii) for any $j, k \le r$, $\Lambda'_{\cdot j} e F_{\cdot k}/\sqrt{nT} = O_{\mathrm{P}}(1)$; (iii) $\mu_1\left(ee'/T\right) = O_{\mathrm{P}}(1)$.

Part (i) of A3 rules out uninteresting cases where the idiosyncratic terms e_{it} are zero or very close to zero for most of i and t. Part (ii) of A3 is in the spirit of assumptions E (d,e) in Bai and Ng (2008). Validity of the central limit theorem for sequences $\{\Lambda_{ij} e_{it} F_{tk}; i, t \in \mathbb{N}\}$ with $j, k \le r$ is sufficient but not necessary for A3 (ii). Part (iii) of A3 further bounds the amount of dependence in the idiosyncratic terms.

Assumption A3 (iii) is technically very convenient and has been previously used by Moon and Weidner (2010). They provide several examples of primitive conditions implying A3 (iii). Proposition 6, which we formulate and prove in the Appendix, shows that A3 (iii) holds for very wide classes of stationary processes $\{e_{\cdot t}, t \in \mathbb{Z}\}$.

Proposition 1 *Let $P_{i:j}$ be a $T \times T$ matrix of projection on the space spanned by $F_{\cdot i}, ..., F_{\cdot j}$, and let $Q_{i:j}$ be an $n \times n$ matrix of projection on the space spanned by $\Lambda_{\cdot i}, ..., \Lambda_{\cdot j}$. Under assumptions A1-A3, as $n, T \to_c \infty$, $L_p = L_p^{(1)} + o_{\mathrm{P}}(1/T)$, where*

$$L_p^{(1)} = \begin{cases} \sum_{j=p+1}^{r} d_{jn} + \operatorname{tr}\left[eP_{1:p}e' + e'Q_{1:p}e\right]/(nT) & \text{if } p \leq r \\ L_r^{(1)} + \sum_{j=r+1}^{p} \mu_j (X'X)/(nT) & \text{if } p > r \end{cases}. \tag{3}$$

It is instructive to compare (3) to the loss in the case of known factors. This case is similar to the standard OLS regression with factor loadings playing the role of the regression coefficients. If the known factors satisfy A1, then a simple regression algebra shows that

$$L_p^{known} = \begin{cases} \sum_{j=p+1}^{r} d_{jn} + \operatorname{tr}\left[eP_{1:p}e'\right]/(nT) & \text{if } p \leq r \\ L_r^{known} + \operatorname{tr}\left[XP_{r+1:p}X'\right]/(nT) & \text{if } p > r \end{cases},$$

where the superscript '*known*' is introduced to distinguish the case of known factors from that of latent factors.

Comparing L_p^{known} to $L_p^{(1)}$, we see that $L_p^{(1)}$ contains an extra term $\operatorname{tr}\left[e'Q_{1:p}e\right]/(nT)$. The reason is that, in Proposition 1, not only loadings, but also factors are estimated. Hence, the expression for the loss becomes symmetric with respect to interchanging factors and factor loadings. More important, for $p > r$, the term $\operatorname{tr}\left[XP_{r+1:p}X'\right]/(nT)$ in L_p^{known} is replaced by the term $\sum_{j=r+1}^{p} \mu_j (X'X)/(nT)$ in $L_p^{(1)}$. It is because when the over-specified factors are not known, they are chosen so as to explain as much variation as possible. In other words, the projection $P_{r+1:p}$ in $\operatorname{tr}\left[XP_{r+1:p}X'\right]/(nT)$ is replaced by the projection on the space spanned by the $r+1, ..., p$-th principal eigenvectors of $X'X$.

If we further assume homoscedasticity, $E(e'e) = n\sigma^2 I_T$, then we can write

$$EL_p^{known} = \begin{cases} \sum_{j=p+1}^{r} d_{jn} + \sigma^2 p/T & \text{if } p \leq r \\ \sigma^2 p/T & \text{if } p > r \end{cases}. \tag{4}$$

Hence, the expected loss, or risk, consists of the bias term $\sum_{j=p+1}^{r} d_{jn}$ and the variance term $\sigma^2 p/T$, with the bias term disappearing under correct or over-specification. In the case of latent factors, we have

Corollary 1 *Suppose that the elements of* e *are i.i.d. zero mean random variables with variance* σ^2 *and a finite fourth moment. Then, under assumptions A1-A2, as* $n, T \to_c \infty$, $L_p = L_p^{(1)} + o_{\mathrm{P}}(1/T)$, *where*

$$EL_p^{(1)} = \begin{cases} \sum_{j=p+1}^{r} d_{jn} + \sigma^2 p (1/T + 1/n) & \text{if } p \leq r \\ \sigma^2 (p-r) (1/\sqrt{T} + 1/\sqrt{n})^2 + \sigma^2 r (1/T + 1/n) & \text{if } p > r \end{cases}.$$

Comparing $EL_p^{(1)}$ to EL_p^{known}, we see that the variance term of $EL_p^{(1)}$ is symmetric with respect to interchanging n and T. More important, in contrast to the case of known factors, the marginal effect on the variance term of $EL_p^{(1)}$ from adding p-th factor depends on whether the model is under- or over-specified. It is $\sigma^2 (1/T + 1/n)$ in the under-specified, but $\sigma^2(1/\sqrt{T} + 1/\sqrt{n})^2$ in the over-specified case.

The unusual form of the term $\sigma^2(1/\sqrt{T} + 1/\sqrt{n})^2$ can be linked to the a.s. convergence $\mu_1(ee'/T) \to \sigma^2 (1 + \sqrt{c})^2$ as $n, T \to_c \infty$ (Yin et al, 1988). Replacing c in $\sigma^2 (1 + \sqrt{c})^2$ by n/T, and dividing the obtained expression by n, we get

$\sigma^2(1/\sqrt{T}+1/\sqrt{n})^2$ (see the proof of Corollary 1 in the Appendix for more details on the link).

2.2 Weak factors asymptotics

In this subsection we derive an asymptotic approximation to L_p using alternative weak factor assumptions proposed and discussed in detail in Onatski (2012).

A1w There exists a diagonal matrix Δ_n with elements $\delta_{1n} \geq \delta_{2n} \geq ... \geq \delta_{rn} > 0$ along the diagonal, such that $F'F/T = I_r$ and $\Lambda'\Lambda = \Delta_n$. As $n \to \infty$, $\Delta_n \to \Delta$, where Δ is a diagonal matrix with decreasing elements $\delta_1 > \delta_2 > ... > \delta_r > 0$ along the diagonal.

By definition, δ_{jn} equals the cross-sectional sum of the squared loadings of the j-th factor. Hence, δ_{jn} measures the cumulative explanatory power, or strength, of factor j. The convergence $\delta_{jn} \to \delta_j$ stays in contrast to assumption A2, which implies that $\delta_{jn} = nd_{jn} \to \infty$. As explained in detail in Onatski (2012), the asymptotic regime described by A1w is meant to provide an adequate approximation to empirically relevant finite sample situations where a few of the largest eigenvalues of the sample covariance matrix are not overwhelmingly larger than the rest of the eigenvalues.

A2w There exist $n \times n$ and $T \times T$ deterministic matrices A_n and B_T such that $e = A_n \varepsilon B_T$, where (i) ε is an $n \times T$ matrix with i.i.d. $N(0, \sigma^2)$ entries; (ii) A_n is such that $\operatorname{tr}(A_n A_n') = n$ and $(A_n A_n')\Lambda = \Lambda$; (iii) B_T is such that $\operatorname{tr}(B_T' B_T) = T$ and $(B_T' B_T) F = F$.

The idiosyncratic matrices of the form $e = A_n \varepsilon B_T$ were previously considered in Bai and Ng (2006), Onatski (2010, 2012), and Ahn and Horenstein (2013). When A_n and B_T are not identity matrices, the idiosyncratic terms are both cross-sectionally and serially correlated. The assumption restricts the covariance matrix of the vectorized e to be of the Kronecker product form $\sigma^2 B_T' B_T \otimes A_n A_n'$. This can be viewed as an approximation to more realistic covariance structures. For a general discussion of the quality of approximations with Kronecker products see Van Loan and Pitsianis (1993).

As explained in Onatski (2012, p. 247), A2w (ii), (iii) are simplifying technical assumptions. They allow Onatski (2012, Theorem 1) to obtain explicit expressions for the bias of the PC estimator under the weak factors asymptotics. The analysis below will rely on these explicit expressions. The Monte Carlo exercises in Section 4 show that the quality of the loss approximation $L_p^{(1)}$ derived under A2w remains good if A2w (ii) and (iii) are relaxed. A theoretical investigation of this phenomenon requires a substantial additional technical effort. We leave such an investigation for future research.

The Gaussianity assumption made in A2w (i) is certainly very strong. We can relax this assumption to a non-Gaussian (nG) version at the expense of making matrices A_n and B_T more special. Let U_A and V_B be, respectively, $n \times n$ and $T \times T$ orthogonal matrices such that the matrix of the first r columns of U_A equals $(\Lambda' \Lambda)^{-1/2} \Lambda$ and that of the first r columns of V_B' equals $(F'F)^{-1/2} F$. Further, let $\mathcal{A}_0 = \operatorname{diag}(a_1, ..., a_n)$, where $a_i \geq 0$ for all i, $a_i = 1$ for $i \leq r$, and $\sum_{i=1}^n a_i^2 = n$. Similarly, let $\mathcal{B}_0 = \operatorname{diag}(b_1, ..., b_T)$, where $b_i \geq 0$ for all i, $b_i = 1$ for $i \leq r$, and

$\sum_{i=1}^{n} b_i^2 = T$.

A2w (nG) There exist $n \times n$ and $T \times T$ deterministic matrices A_n and B_T such that $e = A_n \varepsilon B_T$, where (i) ε is an $n \times T$ matrix with i.i.d. entries ε_{it}, such that $E\varepsilon_{it} = 0$, $E\varepsilon_{it}^2 = \sigma^2$, and $E\varepsilon_{it}^4 < \infty$; (ii) $A_n = U_A \mathcal{A}_0$; (iii) $B_T = \mathcal{B}_0 V_B$.

Part (i) of assumption A2w (nG) requires only the existence of the fourth moments of ε_{it}, which is much less demanding than the Gaussianity. Although matrices A_n and B_T in parts (ii) and (iii) are more special than their counterparts in assumption A2w, their special form is not constraining the covariance matrix $\sigma^2 B_T' B_T \otimes A_n A_n'$ of the vectorized e. Indeed, any A_n and B_T that satisfy A2w (ii) and (iii) must have singular value decompositions of the form $A_n = U_A \mathcal{A}_0 V_A$ and $B_T = U_B \mathcal{B}_0 V_B$, where V_A and U_B are, respectively, $n \times n$ and $T \times T$ orthogonal matrices. Therefore, $A_n A_n' = U_A \mathcal{A}_0^2 U_A'$ and $B_T' B_T = V_B' \mathcal{B}_0^2 V_B$, which is the same as for A_n and B_T that satisfy A2w (nG) (ii) and (iii).

Our last assumption describes the asymptotic behavior of matrices A_n and B_T. Let $G_A(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[\mu_i\left(A_n A_n'\right) \leq x\right]$ and $G_B(x) = \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}\left[\mu_i\left(B_T' B_T\right) \leq x\right]$, where $\mathbf{1}[\cdot]$ denotes the indicator function. Hence, $G_A(x)$ and $G_B(x)$ are the empirical distribution functions of the eigenvalues of $A_n A_n'$ and $B_T' B_T$, respectively.

A3w There exist probability distributions $\mathcal{G}_A$ and $\mathcal{G}_B$ with bounded supports $[\underline{x}_A, \bar{x}_A]$ and $[\underline{x}_B, \bar{x}_B]$, cumulative distribution functions $\mathcal{G}_A(x)$ and $\mathcal{G}_B(x)$, and densities $\frac{d}{dx}\mathcal{G}_A(x)$ and $\frac{d}{dx}\mathcal{G}_B(x)$ at every interior point of support $x \in (\underline{x}_A, \bar{x}_A)$ and $x \in (\underline{x}_B, \bar{x}_B)$, respectively, such that, as $n, T \to_c \infty$: (i) $G_A(x) \to \mathcal{G}_A(x)$ and $G_B(x) \to \mathcal{G}_B(x)$ for all $x \in \mathbb{R}$, (ii) $\mu_1\left(A_n A_n'\right) \to \bar{x}_A$ and $\mu_1\left(B_T' B_T\right) \to \bar{x}_B$, and (iii) $\inf_{x \in (\underline{x}_A, \bar{x}_A)} \frac{d}{dx}\mathcal{G}_A(x) > 0$ and $\inf_{x \in (\underline{x}_B, \bar{x}_B)} \frac{d}{dx}\mathcal{G}_B(x) > 0$.

Assumption A3w holds for a broad range of matrices $A_n A_n'$ and $B_T' B_T$. For example, it is satisfied for large classes of widely used Toeplitz matrices.

Onatski (2012) shows that, under assumptions A1w-A3w, there is an asymptotic relationship between the true strength of the j-th factor, δ_j, and the j-th sample covariance eigenvalue. Precisely, as $n, T \rightarrow_c \infty$,

$$\mu_j \left(XX'/T\right) \xrightarrow{p} \sigma^2 f\left(\delta_j/\sigma^2\right), \ j \leq r, \tag{5}$$

where function $f(\cdot)$ depends only on $\mathcal{G}_A$ and $\mathcal{G}_B$ and can be evaluated numerically. In contrast, under the strong factor assumptions A1-A3, the r largest eigenvalues of XX'/T, which are sometimes referred to as "factor eigenvalues", diverge to infinity. Function $f(\cdot)$ plays an important role in the analysis below. Its salient features are summarized in the following lemma.

Lemma 1 *Suppose that assumptions A1w, A2w or A2w (nG), and A3w hold. Then, (i) there exists* $\bar{\delta} > 0$*, that depends on* $\mathcal{G}_A$ *and* $\mathcal{G}_B$*, such that* $\sigma^2 f\left(\delta/\sigma^2\right) = \operatorname{plim} \mu_1\left(ee'/T\right)$ *for any* $\delta \in \left[0, \bar{\delta}\right]$*; (ii) As a function of* z*,* $f(z)$ *is non-decreasing and continuous, and larger than* z *on* $z \geq 0$*. Furthermore, it is differentiable on* $z < \bar{\delta}/\sigma^2$ *and on* $z > \bar{\delta}/\sigma^2$*, and is such that* $f(z)/z \to 1$ *as* $z \to \infty$*; (iii) the elasticity* $\mathrm{d} \ln f(z)/\mathrm{d} \ln z$ *increases on* $z > \bar{\delta}/\sigma^2$ *and converges to one as* $z \to \infty$*.*

Proposition 2 *Suppose that assumptions A1w, A2w or A2w (nG), and A3w hold. Furthermore, suppose that* $\delta_j \neq \bar{\delta}$ *for* $j = 1, ..., r$ *and let* q *be the largest* $p \in \{0, 1, ..., r\}$ *such that* $\delta_p > \bar{\delta}$*, where* $\delta_0 = \infty$*. Then, as* $n, T \rightarrow_c \infty$*,* $L_p = L_p^{(1)} +$

$o_{\mathrm{P}}(1/T)$, *where*

$$L_p^{(1)} = \begin{cases} \sum_{j=1}^{r} \delta_{jn}/n + \sum_{j=1}^{p} \mu_j (XX') / (nT) - 2\sum_{j=1}^{p} \delta_{jn} f' (\delta_{jn}/\sigma^2) / n & \textit{for } p \leq q \\ L_q^{(1)} + \sum_{j=q+1}^{p} \mu_j (XX') / (nT) & \textit{for } p > q \end{cases}. \tag{6}$$

Here $f'(z)$ denotes the derivative of $f(z)$.

For $p > r \geq q$, the increment to $L_p^{(1)}$ due to over-specifying p factors relative to $p-$ 1 factors is approximated by $\mu_p (XX') / (nT)$. As can be seen from (3), this coincides with the increment to $L_p^{(1)}$ due to the marginal increase in the over-specification under the strong factors asymptotics.

For $p \leq r$, the weak and strong factors asymptotic approximations to the loss are substantially different. Under the weak factors asymptotics, the increment $L_p^{(1)} - L_{p-1}^{(1)}$ remains $\mu_p (XX') / (nT)$ for $p > q$. In such cases, the p-th factor is so weak that $\lim_{n\to\infty} \delta_{pn} \leq \bar{\delta}$. For $p \leq q$, the increment becomes $\mu_p (XX') / (nT) - 2\delta_{pn} f' (\delta_{pn}/\sigma^2) / n$. As formula (5) shows, this equals $\sigma^2 f (\delta_p/\sigma^2) / n - 2\delta_p f' (\delta_p/\sigma^2) / n + o_{\mathrm{P}}(1/T)$, which becomes negative for sufficiently large n and T if and only if

$$\mathrm{d} \ln f(z)/\mathrm{d} \ln z > 1/2 \text{ at } z = \delta_p/\sigma^2. \tag{7}$$

By Lemma 1 (iii), the elasticity $\mathrm{d} \ln f(z)/\mathrm{d} \ln z$ increases with z. Thus, asymptotically, the loss is minimized for the largest p such that (7) holds.

Note that δ_p/σ^2 is large for relatively strong factors. Therefore, according to Lemma 1 (iii), $\mathrm{d} \ln f(z) / \mathrm{dln} z \approx 1$ at $z = \delta_p/\sigma^2$, so that (7) is satisfied. In other words, including very strong factors to the model always leads to a decrease in the loss L_p. For very weak factors, $\mathrm{d}\ln f(z)/\mathrm{dln}z = 0$ at $z = \delta_p/\sigma^2$ because $f(\delta/\sigma^2)$ is constant for $\delta \leq \bar{\delta}$ by Lemma 1 (i), so that (7) is violated, and including such factors to the model leads to an increase in the loss. Inequality (7) tells us exactly how strong the factors should be so that including them to the model improves the prediction of the common component.

For special cases of $\mathcal{G}_A$ and $\mathcal{G}_B$, it is possible to obtain explicit formulae for $f(\cdot)$ in (5). For example, if $A_n = I_n$ and $B_T = I_T$ so that the idiosyncratic terms are i.i.d., we have (see Onatski (2006, Theorem 5)[1]),

$$f\left(\delta/\sigma^2\right) = \begin{cases} \left(1+\sqrt{c}\right)^2 & \text{for } 0 \leq \delta/\sigma^2 \leq \sqrt{c} \\ \left(\delta/\sigma^2+1\right)\left(\delta/\sigma^2+c\right)/\left(\delta/\sigma^2\right) & \text{for } \delta/\sigma^2 > \sqrt{c} \end{cases}. \tag{8}$$

Then, the asymptotic approximation (6) given in Proposition 2 simplifies.

Corollary 2 *Suppose that assumption A1w holds, and let the elements of* e *be i.i.d. with zero mean, variance* σ^2*, and finite fourth moment. Further, suppose that* $\delta_j \neq \sigma^2\sqrt{c}$ *for* $j = 1, ..., r$*, and let* q *be maximum* $p \in \{0, 1, ..., r\}$ *such that* $\delta_p > \sigma^2\sqrt{c}$*, where* $\delta_0 = \infty$*. Then,* $L_p = L_p^{(1)} + o_{\mathrm{P}}(1/T)$*, where*

$$L_p^{(1)} = \begin{cases} \sum_{j=p+1}^{r} \delta_{jn}/n + p\sigma^2\left(1/n+1/T\right) + 3\sum_{j=1}^{p}\sigma^4/\left(T\delta_{jn}\right) & \text{for } p \leq q \\ L_q^{(1)} + (p-q)\,\sigma^2\left(1/\sqrt{n}+1/\sqrt{T}\right)^2 & \text{for } p > q \end{cases}.$$

Note that, under the weak factors asymptotics, the minimum of $L_p^{(1)}$ is achieved not necessarily at $p = r$, as is the case under the strong factors asymptotics. In-

[1]In that paper, Theorem 5 is proven under the Gaussianity of ε. However, as follows from the proof of Lemma 1, $f(\cdot)$ does not depend on the Gaussianity assumption.

stead, it is achieved at the maximum of $p \in \{0, 1, 2, ..., q\}$ such that $\delta_{pn}/n > \sigma^2 (1/n + 1/T) + 3\sigma^4 / (T\delta_{pn})$. The optimal p trades off the bias introduced by not estimating all factors with the reduction in variance that comes from excluding factors that are too weak to be accurately estimated.

3 Loss estimation

In this section, we develop statistics $\hat{L}_p$ that approximate L_p. As mentioned in the introduction, although AIC is a natural candidate for $\hat{L}_p$, it fails in our setting. Let us explore this in more detail. In the simplest special case where the idiosyncratic terms are i.i.d. $N(0, 1)$, the log-likelihood equals

$$\ln L(X|\Lambda, F) = -\frac{nT}{2} \ln 2\pi - \frac{1}{2} \operatorname{tr}\left[(X - \Lambda F')(X - \Lambda F')'\right],$$

so that the Kullback-Leibler distance between the true model (1) and the model with parameters $\tilde{F}, \tilde{\Lambda}$ is

$$KL\left(F, \Lambda; \tilde{F}, \tilde{\Lambda}\right) = E \ln \frac{L(X|\Lambda, F)}{L\left(X|\tilde{\Lambda}, \tilde{F}\right)} = \frac{1}{2} \operatorname{tr}\left[\left(\Lambda F' - \tilde{\Lambda}\tilde{F}'\right)\left(\Lambda F' - \tilde{\Lambda}\tilde{F}'\right)'\right].$$

Hence, L_p equals $2KL\left(F, \Lambda; \tilde{F}, \tilde{\Lambda}\right)/(nT)$, evaluated at $\tilde{F} = \hat{F}_{1:p}$ and $\tilde{\Lambda} = \hat{\Lambda}_{1:p}$, which is the exact analog for the factor models of the loss used by Akaike (1973) to derive AIC (see also deLeeuw (1992), which explains Akaike's (1973) innovative ideas in much detail).

The loss L_p depends on p only through $-2E \ln L\left(X|\tilde{\Lambda}, \tilde{F}\right)/(nT)$, evaluated at

$\tilde{\Lambda} = \hat{\Lambda}_{1:p}$ and $\tilde{F} = \hat{F}_{1:p}$. In our setting, the Akaike's (1973) idea is to approximate this part of L_p by $-2 \ln L\left(X|\hat{\Lambda}_{1:p}, \hat{F}_{1:p}\right)/(nT)$ and correct for the bias. The correction term, at least for the over-specified models, should be two times the parameter dimensionality divided by the sample size. Unfortunately, this simple rule does not hold here.

For the sake of illustration, let there be no factors in the data ($r = 0$). Then,

$$\begin{aligned} -\frac{2}{nT} E \ln L\left(X|\tilde{\Lambda}, \tilde{F}\right)\Big|_{\tilde{\Lambda}, \tilde{F} = \hat{\Lambda}_{1:p}, \hat{F}_{1:p}} &= \ln 2\pi + 1 + \frac{1}{nT} \operatorname{tr}\left[\hat{\Lambda}_{1:p} \hat{F}'_{1:p} \hat{F}_{1:p} \hat{\Lambda}_{1:p}\right] \\ &= \ln 2\pi + 1 + \frac{1}{n} \sum_{j=1}^{p} \mu_j, \end{aligned}$$

where μ_j is a shorthand notation for $\mu_j(XX'/T)$. Furthermore,

$$-\frac{2}{nT} \ln L\left(X|\hat{\Lambda}_{1:p}, \hat{F}_{1:p}\right) = \ln 2\pi + \frac{1}{n} \sum_{j=1}^{n} \mu_j - \frac{1}{n} \sum_{j=1}^{p} \mu_j. \tag{9}$$

Combining these two equalities, we obtain

$$-\frac{2}{nT}\left[E \ln L\left(X|\tilde{\Lambda}, \tilde{F}\right)\Big|_{\tilde{\Lambda}, \tilde{F} = \hat{\Lambda}_{1:p}, \hat{F}_{1:p}} - \ln L\left(X|\hat{\Lambda}_{1:p}, \hat{F}_{1:p}\right)\right] = 1 - \frac{1}{n} \sum_{j=1}^{n} \mu_j + \frac{2}{n} \sum_{j=1}^{p} \mu_j. \tag{10}$$

As shown in Onatski et al (2013, Lemma 12), $n\left(1 - \frac{1}{n}\sum_{j=1}^{n} \mu_j\right) \xrightarrow{d} N(0, 2c)$ as $n, T \to_c \infty$. Hence, the term $1 - \frac{1}{n}\sum_{j=1}^{n} \mu_j$ in the latter equality does not contribute to the bias correction through the order $1/n$. Further, by Yin et al's (1988) result, $2\sum_{j=1}^{p} \mu_j \xrightarrow{a.s.} 2p\left(1 + \sqrt{c}\right)^2$. Replacing c by n/T, we see that $\frac{2}{n}\sum_{j=1}^{p} \mu_j$ can be approximated through the order $1/n$ by $\frac{2p}{nT}(\sqrt{n} + \sqrt{T})^2$. In the factor model setting, the sample size is nT. Thus, had the Akaike's (1973) rule for the bias correction worked, we would have had $p(\sqrt{n}+\sqrt{T})^2$ as the parameter dimensionality. However, to the order n, the number of free parameters in the p-factor model is $p\,(n+T) \neq p(\sqrt{n}+\sqrt{T})^2$.

Akaike's (1973) derivations of his bias correction rule is based on the quadratic approximations to the log-likelihood and on the standard properties of the maximum likelihood estimates. There are at least two reasons why this standard machinery does not work in the setting of large factor models. First, the number of parameters is increasing with the sample size. Second, parameters of an over-specified model are not identified (when the true loadings of a factor are identically zero, the factor may correspond to any point on the sphere of radius $\sqrt{T}$ in $\mathbb{R}^T$). These problems are related to the well-known incidental parameters problem (Lancaster, 2000) and the non-standard inference in cases where some parameters are not identified under the null (Hansen, 1996).

Although the AIC's bias correction rule does not work, equation (10) shows that the bias can be corrected simply by adding $2\sum_{j=1}^{p}\mu_j/n$ to $-2\ln L\left(X|\hat{\Lambda}_{1:p},\hat{F}_{1:p}\right)/(nT)$. Even though under the general assumptions A1-A3 or A1w-A3w and their non-Gaussian version, the Kullback-Leibler interpretation of L_p is lost, Propositions 1 and 2 suggest that, up to a quantity that does not depend on p, L_p is still well approximated by the right hand side of (9) after a bias correction which is a function of μ_j. Specifically, at least for $p>r$, adding $2\sum_{j=r+1}^{p}\mu_j/n$ to the right hand side of (9) will perfectly match $L_p^{(1)}$, after a shift by a quantity that does not depend on p.

Below we propose several estimators of the loss L_p, shifted by a quantity that does not depend on p. All our estimators utilize a preliminary estimator $\hat{r}$ of the number of factors. Under the strong factors asymptotics, there exist many $\hat{r}$ that are consistent for r (see, for example, Bai and Ng (2002) and Ahn and Horenstein (2013)). Under the weak factors asymptotics, consistent estimation of r may not be possible if the strength of some of the factors does not exceed the threshold $\bar{\delta}$. However, it is possible to consistently estimate q. Onatski (2012) calls q the "effective" number of factors and points out that estimator

$$\hat{r}(\varepsilon) = \max\left\{i \leq r_{\max} : \mu_j - \mu_{j+1} > \varepsilon\right\}, \tag{11}$$

where $r_{\max}$ is a fixed maximum possible number of factors and $\varepsilon > 0$ is a small tuning parameter[2], is consistent for q as long as $f(\delta_q/\sigma^2) - f\left(\bar{\delta}/\sigma^2\right) > \varepsilon/\sigma^2$. The latter condition is violated only if the strength δ_q of the q-th strongest factor is sufficiently close to the threshold $\bar{\delta}$ below which the consistent detection of factors is theoretically impossible.

Consider estimators of the shifted loss L_p that have form

$$\hat{L}_p = \begin{cases} \sum_{j=1}^{p}\left(1 - 2\hat{\rho}_j\right)\mu_j/n & \text{for } p \leq \hat{r} \\ \hat{L}_{\hat{r}} + \sum_{j=\hat{r}+1}^{p}\mu_j/n & \text{for } p > \hat{r} \end{cases}, \tag{12}$$

where $\hat{\rho}_j$ with $j = 1, ..., p$ are data dependent quantities. The following choices of $\hat{\rho}_j$

[2] Onatski (2010) proposes a data-dependent calibration procedure of ε, which we will use in the Monte Carlo section of this paper.

give us two special cases of $\hat{L}_p$:

$$\underline{L}_p = \hat{L}_p \text{ with } \hat{\rho}_j = 1, \text{ and} \tag{13}$$

$$\bar{L}_p = \hat{L}_p \text{ with } \hat{\rho}_j = 1/(\mu_j^2 \max\{\hat{m}'(\mu_j), \widetilde{m}'(\mu_j)\}). \tag{14}$$

Here

$$\hat{m}'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\hat{m}(x) \text{ with } \hat{m}(x) = (n-\hat{r})^{-1}\sum\nolimits_{i=\hat{r}+1}^{n} (\mu_i - x)^{-1},$$

and

$$\widetilde{m}'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\widetilde{m}(x) \text{ with } \widetilde{m}(x) = (T-\hat{r})^{-1}\sum\nolimits_{i=\hat{r}+1}^{T} (\mu_i - x)^{-1},$$

where, for $j > n$, μ_j is defined as zero.

Proposition 3 *(i) Let $\tilde{L}_p = \underline{L}_p$ or $\tilde{L}_p = \bar{L}_p$. Then, under assumptions A1-A3, if $\hat{r} \xrightarrow{p} r$ as $n, T \to_c \infty$, then*

$$\max_{0\le p<r} \left| L_p - \tilde{L}_p - (L_r - \tilde{L}_r) \right| = o_{\mathrm{P}}(1) \text{ and} \tag{15}$$

$$\max_{r\le p\le r_{\max}} \left| L_p - \tilde{L}_p - (L_r - \tilde{L}_r) \right| = o_{\mathrm{P}}(1/T). \tag{16}$$

(ii) Under assumptions A1w, A2w or A2w (nG), and A3w, if $\hat{r} \xrightarrow{p} q$ as $n, T \to_c \infty$, then for any $\epsilon > 0$,

$$\Pr[\min_{0\le p\le r_{\max}} \left((L_p - L_0) - \underline{L}_p\right) \ge -\epsilon/n] \to 1 \text{ and} \tag{17}$$

$$\Pr[\max_{0\le p\le r_{\max}} \left((L_p - L_0) - \bar{L}_p\right) \le \epsilon/n] \to 1. \tag{18}$$

Part (i) of Proposition 3 shows that both $\underline{L}_p$ and $\bar{L}_p$ can be thought of as asymptotic approximations to a shifted version of L_p. As follows from Proposition 1, $\min_{0\leq p<r} L_p$ is bounded away from zero with probability approaching one. Hence, the approximation error in (15) is asymptotically negligible relative to the size of the loss. The portion of the loss L_p that corresponds to $r \leq p \leq r_{\max}$ converges to zero. It can be shown (see the proof of Proposition 5) that, for $r < p \leq r_{\max}$, the rate of such convergence is $1/T$, and the approximation error in (16) is also negligible relative to the size of the loss.

The reason why both $\underline{L}_p$ and $\bar{L}_p$ approximate L_p well asymptotically is that, under the strong factors asymptotics, the difference $\underline{L}_p - \bar{L}_p$ converges to zero. Indeed, with probability approaching one, $\mu_j \to \infty$ for any $j \leq \hat{r}$, and $\mu_{\hat{r}+1} = O_{\mathrm{P}}(1)$. Therefore, $\hat{\rho}_j$ with $j \leq \hat{r}$ defined in (14) converge in probability to one, which coincides with the value of $\hat{\rho}_j$ in (13).

Part (ii) of Proposition 3 shows that, under the weak factors asymptotics, $\underline{L}_p$ and $\bar{L}_p$ can be thought of as asymptotic lower and upper bounds on a shifted version of L_p. According to Proposition 2, an estimator $\hat{L}_p$ that would approximate L_p well under the weak factors asymptotics must have $\operatorname{plim} \hat{\rho}_j = \mathrm{d} \ln f\left(\delta_j/\sigma^2\right)/\mathrm{dln}(\delta_j/\sigma^2)$. We were able to develop such $\hat{\rho}_j$ only in the special cases where either $A = I_n$ or $B = I_T$, that is where there is either no cross-sectional or no temporal correlation in the idiosyncratic terms.

Let $\hat{L}_p^{(A=I)}$ and $\hat{L}_p^{(B=I)}$ be the estimators $\hat{L}_p$ with $\hat{\rho}_j = \hat{\rho}_j^{(A=I)}$ and $\hat{\rho}_j = \hat{\rho}_j^{(B=I)}$, respectively, where

$$\hat{\rho}_j^{(A=I)} = -(1+\hat{m}(\mu_j)\hat{\sigma}^2 n/T)\hat{m}(\mu_j)/(\mu_j \hat{m}'(\mu_j)), \text{ and} \tag{19}$$

$$\hat{\rho}_j^{(B=I)} = -(1+\widetilde{m}(\mu_j)\widetilde{\sigma}^2 T/n)\widetilde{m}(\mu_j)/(\mu_j \widetilde{m}'(\mu_j)). \tag{20}$$

In the above expressions, $\hat{\sigma}^2 = (n-\hat{r})^{-1}\sum_{i=\hat{r}+1}^{n}\mu_i$, and $\widetilde{\sigma}^2 = (T-\hat{r})^{-1}\sum_{i=\hat{r}+1}^{T}\mu_i$.

Proposition 4 *(i) Let $\tilde{L}_p = \hat{L}_p^{(A=I)}$ or $\tilde{L}_p = \hat{L}_p^{(B=I)}$. Then, under assumptions A1-A3, if $\hat{r} \xrightarrow{p} r$ as $n, T \to_c \infty$, then*

$$\max_{0\le p<r}\left|L_p - \tilde{L}_p - (L_r - \tilde{L}_r)\right| = o_{\mathrm{P}}(1) \text{ and}$$
$$\max_{r\le p\le r_{\max}}\left|L_p - \tilde{L}_p - (L_r - \tilde{L}_r)\right| = o_{\mathrm{P}}(1/T).$$

(ii) Suppose that assumptions A1w, A2w or A2w (nG), and A3w hold. If $\hat{r} \xrightarrow{p} q$ as $n, T \to_c \infty$, and $A = I_n$ or $B = I_T$, respectively,

$$\max_{0\le p\le r_{\max}}\left|L_p - \hat{L}_p^{(A=I)} - (L_q - \hat{L}_q^{(A=I)})\right| = o_{\mathrm{P}}(1/T) \text{ or}$$
$$\max_{0\le p\le r_{\max}}\left|L_p - \hat{L}_p^{(B=I)} - (L_q - \hat{L}_q^{(B=I)})\right| = o_{\mathrm{P}}(1/T).$$

As can be seen from part (i) of Proposition 4, both $\hat{L}_p^{(A=I)}$ and $\hat{L}_p^{(B=I)}$ approximate a shifted loss L_p under the strong factors asymptotics. This result is similar to part (i) of Proposition 3. It holds because both $\hat{\rho}_j^{(A=I)}$ and $\hat{\rho}_j^{(B=I)}$ converge in probability to one under the strong factors asymptotics. For the weak factors asymptotics, $\hat{L}_p^{(A=I)}$ approximates a shifted loss L_p when $A = I_n$, whereas $\hat{L}_p^{(B=I)}$ approximates a shifted loss L_p when $B = I_T$. These approximations improve upon the asymptotic bounds

$\underline{L}_p$ and $\bar{L}_p$ from part (ii) of Proposition 3.

The approximations to the shifted loss given in Propositions 3 and 4 can be used to assess changes in the loss that result from different specifications of the number of factors. Alternatively, they can be used to select an *asymptotically loss efficient* number of factors. The concept of *asymptotic loss efficiency* of model selection procedures was studied in detail by Shibata (1980), Li (1987), and Shao (1997), among others. In the context of factor models and loss L_p, it can be described as follows. Let $\hat{p}$ be an estimator of the number of factors that may or may not coincide with the preliminary estimator $\hat{r}$. Estimator $\hat{p}$ is called *asymptotically loss efficient* if

$$\frac{L_{\hat{p}}}{\min_{0\leq p\leq r_{\max}} L_p} \xrightarrow{p} 1. \tag{21}$$

Shao (1997) points out that a sufficient but not necessary condition for the *asymptotic loss efficiency* is

$$\Pr\left(\hat{p} = p^*\right) \rightarrow 1, \tag{22}$$

where $p^* = \arg\min_{0\leq p\leq r_{\max}} L_p$. He calls this stronger property of $\hat{p}$ *consistency.* Since the minimizer p^* of the loss L_p does not necessarily coincide with the true number of factors r, we will call the property (22) *optimal loss consistency* instead.

Proposition 5 *Let* $\underline{p}$, $\bar{p}$, $\hat{p}^{(A=I)}$, *and* $\hat{p}^{(B=I)}$ *be the minimizers of* $\underline{L}_p$, $\bar{L}_p$, $\hat{L}_p^{(A=I)}$, *and* $\hat{L}_p^{(B=I)}$ *on* $0 \leq p \leq r_{\max}$, *respectively.*

(i) Suppose that assumptions A1-A3 hold. Then any estimator consistent for r *is optimal loss consistent. Furthermore, if the preliminary estimator* $\hat{r}$ *is consistent for* r, *then estimators* $\underline{p}$, $\bar{p}$, $\hat{p}^{(A=I)}$, *and* $\hat{p}^{(B=I)}$ *are also consistent for* r, *and thus, are*

also optimal loss consistent.

(ii) Suppose that assumptions A1w, A2w or A2w (nG), and A3w hold, and let $\hat{r} \xrightarrow{p} q$ *as* $n, T \to_c \infty$*. Then,* $\hat{r}$ *is not, in general, optimal loss consistent. For any optimal loss consistent estimator* $\hat{p}$*,* $\Pr\left(\underline{p} \leq \hat{p} \leq \bar{p}\right) \to 1$*. Moreover,* $L_{\bar{p}} - L_{p^*} \leq \bar{L}_{\bar{p}} - \underline{L}_{\underline{p}} + o_{\mathrm{P}}(1/T)$ *and* $L_{\underline{p}} - L_{p^*} \leq \bar{L}_{\underline{p}} - \bar{L}_{\bar{p}} + o_{\mathrm{P}}(1/T)$*.*

(iii) If, in addition to the assumptions of (ii), we have $A = I_n$ *or* $B = I_T$*, then estimators* $\hat{p}^{(A=I)}$ *or* $\hat{p}^{(B=I)}$*, respectively, are optimal loss consistent.*

Parts (ii) and (iii) of Proposition 5 imply that, when factors are weak, the under-estimation of the number of factors r may lead to improvements in the loss, even in large samples. The optimal loss consistent estimator will tend to be smaller than the preliminary estimator $\hat{r}$. Such an optimal estimator is asymptotically bracketed by the estimators $\underline{p}$ and $\bar{p}$.

Note that the quality of estimator $\hat{L}_p$ of the shifted loss depends on the quality of the corresponding preliminary estimator $\hat{r}$ under particular asymptotic regime. For example, choosing $\hat{r}$ equal to an IC or a PC estimator of Bai and Ng (2002) will insure the asymptotic accuracy of $\hat{L}_p$ under strong but not under weak factors asymptotics, because all estimators that satisfy conditions of Theorem 2 of Bai and Ng (2002), although consistent for r under strong factors asymptotics, must converge to zero when factors are weak. Similarly, choosing $\hat{r}$ equal to Onatski's (2010) estimator (11) may result in poor asymptotic behavior of $\hat{L}_p$ under the strong factors asymptotics, because this estimator is, in general, not consistent for r under assumptions A1-A3. To guarantee the consistency of this estimator under the strong factors asymptotics we need to require that $\mu_1(ee'/T) - \mu_{r_{\max}+1}(ee'/T)$ converges to zero in probability as $n, T \to_c \infty$. Assumptions A2w (nG) (i) and A3w would be sufficient, but not necessary, for such a convergence.

In this paper, we do not address the problem of finding standard errors of our loss estimates. Since the estimates can be interpreted as sample analogs of the asymptotic loss approximation $L_p^{(1)}$, it would be relatively straightforward to analyze statistical properties of the differences between the estimates and $L_p^{(1)}$ (as opposed to L_p). However, the order of these differences will be the same as that of $L_p^{(1)} - L_p$. Hence, to derive the standard errors, we need to engage in a higher-order asymptotic analysis. Onatski (2012) does develop some higher-order asymptotic results. However, these results are insufficient to obtain the standard errors of our loss estimates, unless the idiosyncratic terms are i.i.d. Gaussian random variables. We, therefore, leave the standard error analysis for future research.

4 Monte Carlo Experiments

In this section, we use Monte Carlo experiments to assess the finite sample quality of the asymptotic loss approximations $L_p^{(1)}$ and estimators $\underline{L}_p$, $\bar{L}_p$, $\hat{L}_p^{(A=I)}$, and $\hat{L}_p^{(B=I)}$. Our simulation setting is similar to that used in many previous studies, including Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). The data are generated from

$$X_{it} = C_{it} + \sqrt{\theta} e_{it},$$

where the common component C_{it} is independent from the the idiosyncratic component e_{it}, both components are normalized to have variance one, and parameter θ

measures the inverse of the signal-to-noise ratio.

The common component is generated by process $C_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj}/\sqrt{r}$, where $r = 3$, and λ_{ij} and F_{tj} are i.i.d. $N(0,1)$. Dividing by $\sqrt{r}$ insures that the variance of C_{it} equals one. The idiosyncratic component e_{it} is generated by the process

$$e_{it} = \rho e_{i,t-1} + v_{it} + \sum\nolimits_{j\neq 0, j=-J}^{J} \beta v_{i-j,t},$$

where v_{it}, $i, t \in \mathbb{Z}$ are i.i.d. random variables with mean zero and variance $\sigma_v^2 = (1-\rho^2)/(1+2J\beta^2)$ with $J = \min(n/20, 10)$. We consider two distributions for v_{it} : the Gaussian and Student's t(5), the latter having only four finite moments. Following Ahn and Horenstein (2013), we consider $(\rho, \beta) = (0,0)\,, (0.7, 0)\,, (0, 0.5)\,,$ or $(0.5, 0.2)$. The signal-to-noise ratio θ^{-1} takes on five possible values: 4, 2, 1, 1/2, or 1/4. The sample sizes are $(n, T) = (50, 200)\,, (100, 100)$ and $(200, 50)$. The maximum possible number of factors $r_{\max}$ is set to 8.

To give the reader an idea on how the loss function in our MC experiments looks like, Figure 1 shows two particular realizations of L_p for $(n, T) = (100, 100)$, $(\rho, \beta) = (0.7, 0)$, and the Gaussian distribution for the idiosyncratic terms. These realizations are superimposed with the strong factors (dashed lines) and weak factors (dotted lines) asymptotic approximations $L_p^{(1)}$, derived in Propositions 1 and 2. Function $f(\cdot)$ that appears in the weak factors asymptotic approximation is computed numerically using the MATLAB code developed in Onatski (2012, p. 248).

The left and right panels of the figure correspond to the lowest and the highest signal-to-noise ratio, respectively. In practice, the relative strength of the signal is often assessed using the scree plot. Hence, we also provide graphs (dots with abscissa

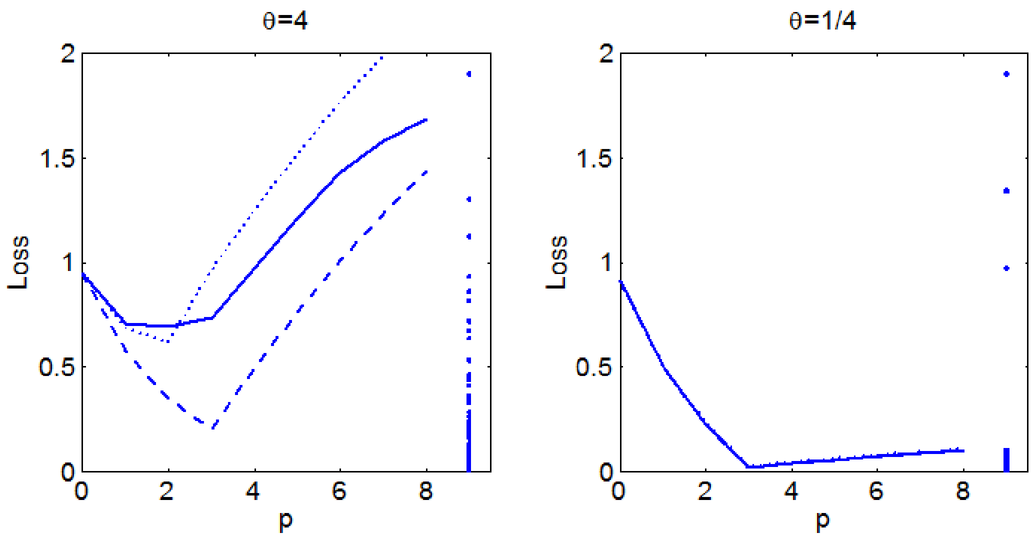

Figure 1: Particular realizations of the loss L_p (solid lines) and the corresponding asymptotic approximations $L_p^{(1)}$ (dotted lines – weak factors approximation, dashed lines – strong factors approximation). $n = T = 100$, $(\rho, \beta) = (0.7, 0)$.

9) of the sorted eigenvalues of the sample covariance matrix, scaled to fit the picture. For the low signal-to-noise ratio ($\theta = 4$), the first three eigenvalues do not clearly separate from the smaller eigenvalues, whereas for the large signal-to-noise ratio ($\theta = 1/4$) the separation is obvious.

We see that when $\theta = 1/4$, so that the factors are relatively strong, L_p is minimized at the true number of factors $p = r = 3$, and both approximations to the loss are very close to the actual realization.

When $\theta = 4$, so that the factors are relatively weak, L_p is no longer necessarily minimized at $p = 3$. For the particular realization shown at the picture, the loss is minimized at $p = 2$, but the minimum is relatively large. Its value is 0.69, which means, roughly, that 69% of the variation of the PC estimator of the common component that uses the optimal number of factors is due to the error (recall that the common component is normalized to have variance one). In this difficult case, the weak factors asymptotic approximation is better than the strong factors one for $p \leq r$.

Figure 2 shows the root mean squared errors (RMSE) of the strong factors (dashed lines) and weak factors (dotted lines) asymptotic approximations of L_p, the mean being taken over 1000 MC replications. The figure corresponds to $(\rho, \beta) = (0.7, 0)$ and Student's t(5) distribution for the idiosyncratic innovations. The other cases provide qualitatively similar information. The corresponding results are reported in the Supplementary Appendix.[3] For relatively strong factors, the quality of the strong factors asymptotic approximation is uniformly better than that of the weak factors approximation. However, the scale of the difference between the qualities of the two approximations is very small (both approximations work very well). For relatively weak factors, the scale of the difference between the qualities of the approximations increases, and the weak factors asymptotic approximation become preferable, especially for $p \leq r$.

We now turn to the analysis of the proposed loss estimators $\hat{L}_p$. Since $\hat{L}_p$ estimate L_p only up to a shift that does not depend on p, it is natural to compare $\hat{L}_p - \hat{L}_{p-1}$ and $L_p - L_{p-1}$ rather than $\hat{L}_p$ and L_p. Figure 3 establishes a benchmark for such a comparison, by showing RMSE of $L_p^{(1)} - L_{p-1}^{(1)}$, where $L_p^{(1)}$ correspond to the strong factors (solid line) and weak factors (dotted line) asymptotic approximations to the loss. The MC setting is the same as that of Figure 2. The solid and dotted curves

[3] For the remaining MC experiments, the results for all the considered settings were qualitatively similar. Therefore, we report only the results for t(5) distribution and $(\rho, \beta) = (0.7, 0)$. The results for all other cases are reported in the Supplementary Appendix.

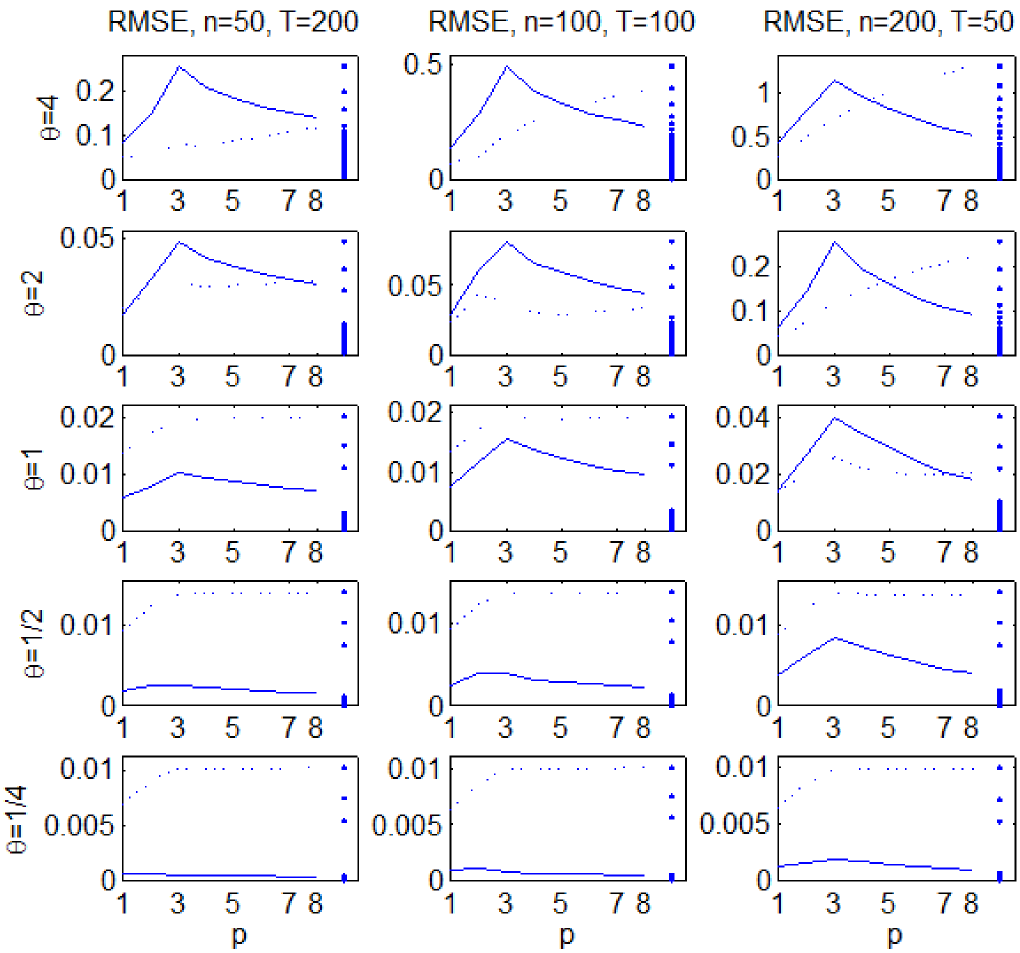

Figure 2: The root mean (over 1000 MC replications) squared errors of the strong (solid lines) and weak (dotted lines) factors asymptotic approximations of L_p. The dots with abscissa p=9 show the MC average values of the sorted eigenvalues of the sample covariance matrix. The idiosyncratic innovations have Student's t(5) distribution. Parameters ρ, β are set to $(\rho, \beta) = (0.7, 0)$. The left, central, and right panels correspond to, respectively, $(n, T) = (50, 200)$, $(100, 100)$ and $(200, 50)$. The strength of the factors increases from top panel (the weakest factors) to bottom panel (the strongest factors).

coincide for $p \geq 4$ because the strong and weak factors asymptotic approximations have the same increments $\mu_p(XX')/(nT)$ for $p > r = 3$.

Since our estimators $\hat{L}_p$ are constructed as sample analogs of the weak factors asymptotic approximation $L_p^{(1)}$, we expect the RMSE of $\hat{L}_p - \hat{L}_{p-1}$ to be, at best, of the same magnitude as the levels of the dotted lines on Figure 3. Figure 4 shows these RMSE for $\hat{L}_p = \bar{L}_p$ (solid lines), $\hat{L}_p = \underline{L}_p$ (dotted lines), $\hat{L}_p = \hat{L}_p^{(A=I)}$ (dashed lines), and $\hat{L}_p = \hat{L}_p^{(B=I)}$ (dash-dotted lines). The MC setting is, again, the same as that of Figure 2. As a preliminary estimator $\hat{r}$, we use Onatski's (2010) ED estimator, given by (11) with the tuning parameter ε calibrated as in Onatski (2010, p. 1008).

Our simplest estimator, $\underline{L}_p$, is dominated by $\bar{L}_p$, $\hat{L}_p^{(A=I)}$, and $\hat{L}_p^{(B=I)}$. The performances of the latter three estimators are virtually the same for $n = T = 100$. For $n \neq T$, $\hat{L}_p^{(A=I)}$ and $\hat{L}_p^{(B=I)}$ perform very similarly, and better than $\bar{L}_p$. For relatively strong factors ($\theta = 1/4$ and $\theta = 1/2$), the accuracy of $\hat{L}_p^{(A=I)} - \hat{L}_{p-1}^{(A=I)}$ and $\hat{L}_p^{(B=I)} - \hat{L}_{p-1}^{(B=I)}$ as estimators of $L_p - L_{p-1}$ is comparable to that of the infeasible estimator $L_p^{(1)} - L_{p-1}^{(1)}$, represented by dotted lines on Figure 3. For relatively weak factors ($\theta = 2$ and $\theta = 4$), the accuracy of $\hat{L}_p^{(A=I)} - \hat{L}_{p-1}^{(A=I)}$ and $\hat{L}_p^{(B=I)} - \hat{L}_{p-1}^{(B=I)}$ is substantially worse than that of $L_p^{(1)} - L_{p-1}^{(1)}$, at least for $p \leq r$.

It turns out that the main reason behind this quality deterioration is the inability of our preliminary estimator $\hat{r}$ to accurately estimate q when factors are relatively weak. Figure 5 illustrates this finding. It shows the same realization of L_p as on the left panel of Figure 1 (solid line) superimposed with $\hat{L}_p^{(A=I)}$ (dashed line), shifted to match the value of L_p at $p = 0$. For the corresponding data replication, we have $q = 2$. However, our preliminary estimator $\hat{r} = 1 < q$. The dotted line shows what

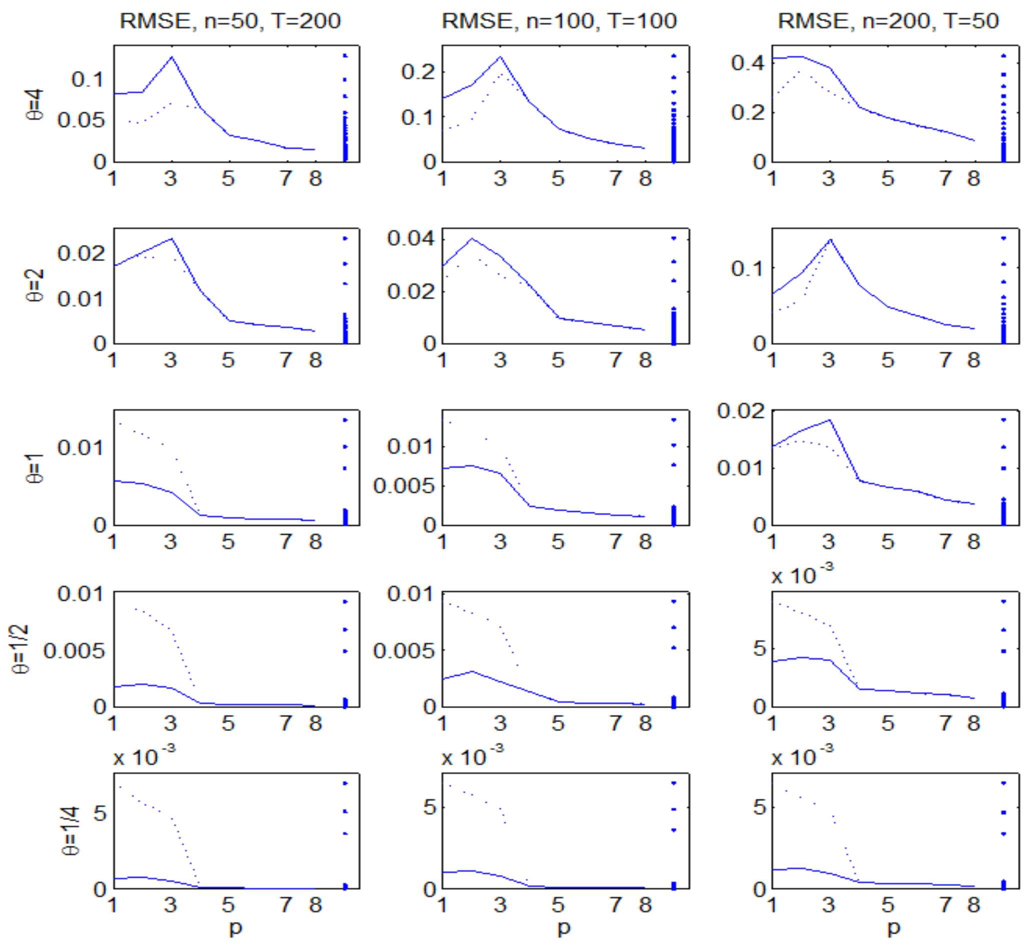

Figure 3: The root mean squared errors of $L_p^{(1)} - L_{p-1}^{(1)}$ (as approximations of $L_p - L_{p-1}$). Solid and dotted lines correspond, respectively, to the strong and weak factors asymptotic approximations $L_p^{(1)}$. The idiosyncratic innovations have Student's t(5) distribution. Parameters ρ, β are set to $(\rho, \beta) = (0.7, 0)$.

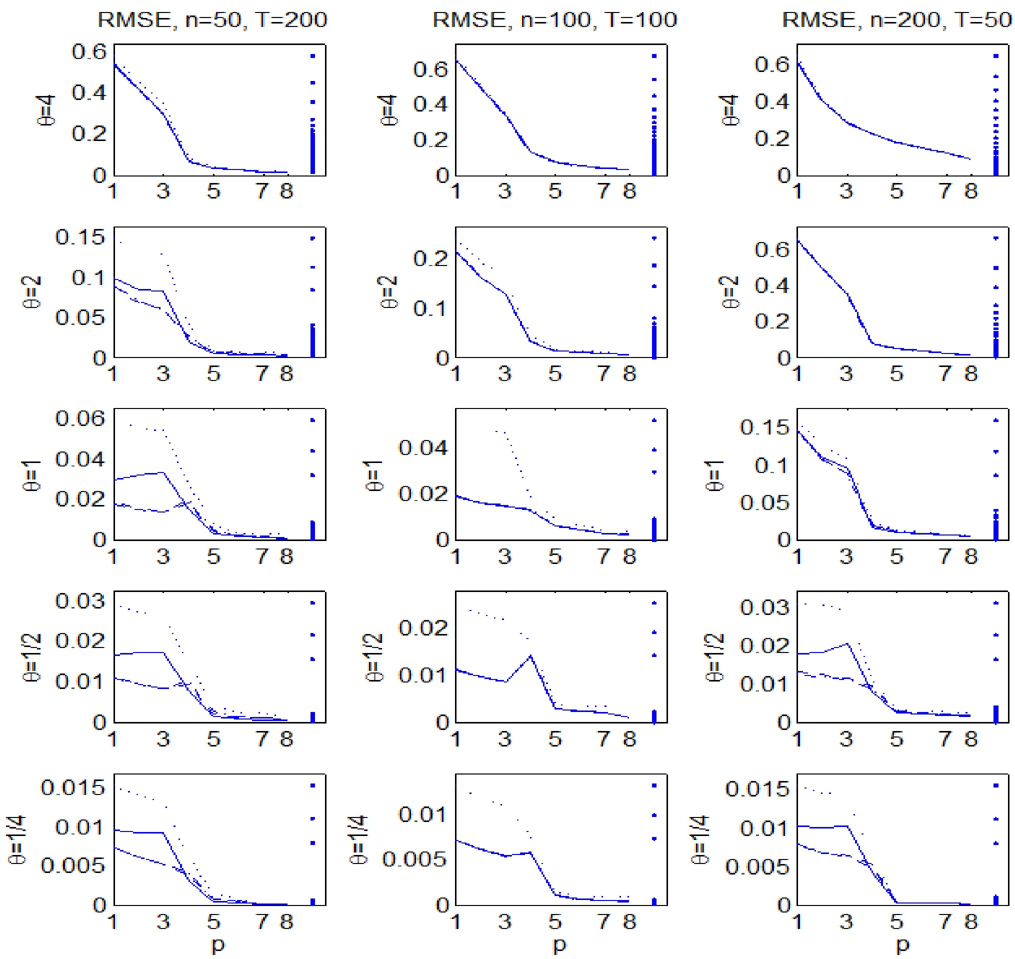

Figure 4: The root mean squared errors of $\hat{L}_p - \hat{L}_{p-1}$ (as approximations of $L_p - L_{p-1}$). Solid, dotted, dashed and dash-dotted lines correspond, respectively, to $\hat{L}_p = \bar{L}_p$, $\hat{L}_p = \underline{L}_p$, $\hat{L}_p = \hat{L}_p^{(A=I)}$, and $\hat{L}_p = \hat{L}_p^{(B=I)}$. The idiosyncratic innovations have Student's t(5) distribution. Parameters ρ, β are set to $(\rho, \beta) = (0.7, 0)$.

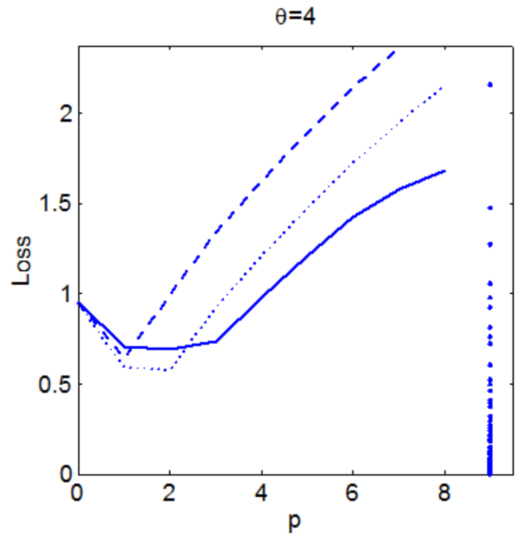

Figure 5: A realization of L_p (solid line) and $\hat{L}_p^{(A=I)}$ (dashed line). Dotted line corresponds to $\hat{L}_p^{(A=I)}$ based on the counterfactual $\hat{r} = q$. $(\rho, \beta) = (0.7, 0)$, $n = T = 100$.

the value of $\hat{L}_p^{(A=I)}$ would have been, had $\hat{r}$ been equal to $q = 2$. Clearly, the accurate estimation of q leads to much more accurate estimation of L_p.

Accurate estimation of q is difficult when factors are weak. The difficulty is well illustrated by the relative position of the eigenvalues shown on Figure 5. All methods of the number of factors estimation explicitly or implicitly look for a separation between r largest eigenvalues and the rest of the eigenvalues. For weak factors, the separation theoretically cannot occur in large samples for more than q eigenvalues, hence the focus on the estimation of q when the factors are weak. The eigenvalues reported in Figure 5 do not show any visible separation, except, perhaps, between the first eigenvalue and the rest, which is captured by the fact that $\hat{r} = 1 < q = 2$.

Under the weak factors asymptotics, only Onatski's (2010) ED estimator has been formally shown to be consistent for q (under the additional assumption that

$f(\delta_q/\sigma^2) - f(\bar{\delta}/\sigma^2) > \varepsilon/\sigma^2)$. However, in principle, other estimators may accurately estimate q in finite samples. Therefore, below, we compare the quality of the loss estimates based on various preliminary estimators $\hat{r}$. In addition to ED, we consider Bai and Ng's (2002) estimators based on their criteria BIC_3, PC_{pj}, and IC_{pj} with $j = 1, 2$, estimators ER and GR developed by Ahn and Horenstein (2013), and Alessi et al's (2010) ABC estimator.

Table 1 reports average (over $p = 1, ..., r_{\max}$) RMSE of $\hat{L}_p^{(A=I)} - \hat{L}_{p-1}^{(A=I)}$, for $\hat{L}_p^{(A=I)}$ based on different versions of $\hat{r}$. Precisely, we compute

$$\frac{100}{r_{\max}} \sum\nolimits_{j=1}^{r_{\max}} \sqrt{E_{MC}\left(\hat{L}_j^{(A=I)} - \hat{L}_{j-1}^{(A=I)} - (L_j - L_{j-1})\right)^2},$$

where E_{MC} denotes the operator of taking mean over MC replications. The MC setting is the same as that for Figures 2-4. We focus on the performance of $\hat{L}_j^{(A=I)}$ because, as shown in Figure 4, it is similar to that of $\hat{L}_j^{(B=I)}$ and better than the performance of the other estimators, at least, in our MC setting.

Since the common component has variance one in all MC experiments, the units of the quality measure reported in Table 1 can be interpreted, roughly, as percents of the standard deviation of the common component. We see that for relatively strong factors all estimators fare very well. For weak factors, the quality substantially deteriorates, especially for relatively small T. Overall, ED, ER, GR, ABC, and BIC show more robust performance. However, none of these estimators clearly dominates.

Another basis for comparison of different estimators of the loss is related to the quality of the corresponding *loss efficient* estimators of the number of factors. Let

n	T	θ	ED	ER	GR	ABC	BIC	PC1	PC2	IC1	IC2
		4	17.6	25.3	25.0	16.4	22.9	7.6	6.6	8.2	10.8
		2	3.3	6.9	5.0	3.8	7.0	3.5	2.9	1.7	1.6
50	200	1	0.9	0.9	0.8	1.0	0.8	1.8	1.5	0.9	0.9
		1/2	0.6	0.4	0.4	0.6	0.4	1.0	0.8	0.5	0.5
		1/4	0.3	0.2	0.2	0.3	0.2	0.5	0.5	0.3	0.3
		4	22.9	25.3	25.2	20.2	17.8	23.1	17.2	23.3	11.4
		2	7.3	7.2	5.5	5.7	3.1	11.1	8.0	11.4	4.4
100	100	1	1.0	0.9	0.8	1.5	0.7	5.4	3.9	5.6	2.1
		1/2	0.7	0.6	0.5	1.3	0.5	2.8	2.0	2.9	1.1
		1/4	0.3	0.2	0.2	0.4	0.3	1.4	1.1	1.5	0.6
		4	25.9	26.2	26.2	27.5	23.2	45.6	45.4	45.6	45.6
		2	21.4	19.8	18.8	15.5	7.2	21.5	21.4	21.5	21.5
200	50	1	5.0	3.6	2.9	4.7	2.5	10.5	10.5	10.5	10.5
		1/2	0.7	0.8	0.6	2.2	1.1	5.2	5.2	5.2	5.2
		1/4	0.3	0.3	0.3	1.1	0.6	2.6	2.6	2.6	2.6

Table 1: Values of $\frac{100}{r_{\max}} \sum_{j=1}^{r_{\max}} \sqrt{E_{MC} \left(\hat{L}_j^{(A=I)} - \hat{L}_{j-1}^{(A=I)} - (L_j - L_{j-1}) \right)^2}$ corresponding to different preliminary estimators of q. The idiosyncratic innovations have Student's t(5) distribution. Parameters ρ, β are set to $(\rho, \beta) = (0.7, 0)$.

$\hat{p} = \arg\min_{0 \le p \le r_{\max}} \hat{L}_p^{(A=I)}$ and $p^* = \arg\min_{0 \le p \le r_{\max}} L_p$. Then, a natural measure of quality of $\hat{L}_p^{(A=I)}$ is $E_{MC} \left(L_{\hat{p}} / L_{p^*} \right)$. Table 2 reports this quality measure for $\hat{L}_p^{(A=I)}$ based on different versions of $\hat{r}$, when $(\rho, \beta) = (0.7, 0)$ and the idiosyncratic innovations are Student's t(5). Loss estimates $\hat{L}_p^{(A=I)}$ based on ED, ER, GR, ABC, and BIC work reasonably well, and better than those based on the other preliminary estimators. However, the choice between ED, ER, GR, ABC, and BIC is not obvious.

n	T	θ	ED	ER	GR	ABC	BIC	PC1	PC2	IC1	IC2
		4	1.44	2.01	1.99	1.36	1.81	1.11	1.07	1.07	1.14
		2	1.06	1.32	1.15	1.09	1.59	1.15	1.08	1.02	1.01
50	200	1	1.02	1.01	1.01	1.04	1.00	1.19	1.11	1.02	1.01
		1/2	1.03	1.00	1.00	1.04	1.00	1.22	1.13	1.03	1.02
		1/4	1.02	1.00	1.00	1.04	1.00	1.22	1.12	1.02	1.01
		4	1.50	1.66	1.65	1.36	1.26	2.20	1.69	2.20	1.14
		2	1.29	1.33	1.18	1.23	1.06	3.27	2.40	3.35	1.38
100	100	1	1.02	1.01	1.00	1.18	1.00	4.05	2.86	4.13	1.50
		1/2	1.03	1.00	1.00	1.23	1.01	4.49	3.10	4.61	1.60
		1/4	1.04	1.00	1.00	1.19	1.01	4.68	3.30	4.81	1.69
		4	1.10	1.10	1.10	1.24	1.27	2.54	2.54	2.55	2.55
		2	2.23	2.02	1.85	1.57	1.10	3.04	3.04	3.04	3.04
200	50	1	1.27	1.21	1.12	1.61	1.23	4.73	4.73	4.73	4.73
		1/2	1.01	1.02	1.00	1.81	1.31	5.83	5.83	5.83	5.83
		1/4	1.01	1.00	1.00	1.88	1.38	6.40	6.40	6.40	6.40

Table 2: Values of $E_{MC}L_{\hat{p}}/L_{p^*}$ corresponding to different preliminary estimators of q. The idiosyncratic innovations have Student's t(5) distribution. Parameters ρ, β are set to $(\rho, \beta) = (0.7, 0)$.

5 Empirical illustration

In this section we illustrate our loss estimation methodology by an analysis of excess return data. A fundamental assumption of the Arbitrage Pricing Theory is that excess returns admit an approximate factor structure. Statistical and fundamental factor models (Connor, 1995) use factors estimated from the excess return data itself or constructed using additional information, such as the book-to-price and market value, respectively. The popular Fama-French three factor model is an example of a fundamental model.

Many studies of statistical factor models, including Connor and Korajczyk (1993), Huang and Jo (1995), Bai and Ng (2002), and Onatski (2010) find only two factors in the excess returns data. If both the Fama-French model and the statistical factor model are correct, the number of factors in the two models must be the same. Assuming that there are indeed three factors in the data, as the Fama-French model postulates, what is the loss from not estimating the third factor in the statistical factor model? This is a question that we can answer using our loss estimator.

We use monthly excess return data constructed from the stock price CRSP data and the historical data on the 3-month T-bill rate. Our data set consists of 284 stocks listed on NYSE selected as follows. First, we selected all stocks for which the price data were available for the entire period from Jan2001 to Dec2012. For each of these stocks, we computed the transaction volume (the product of the share price and the share volume), and sorted the stocks according to the value of the cumulative transaction volume for the entire period. We selected the relatively more actively traded stocks that together constituted 90% of the entire transaction volume. Finally, we eliminated all remaining stocks with standard deviations above three times the median standard deviation. This left us with 284 stocks.

Assuming that these data have three factors, and that the PC method does not break down for the third factor so that $q = 3$, we can estimate the loss function L_p by $\bar{L}_p$, $\underline{L}_p$, $\hat{L}_p^{(A=I)}$, and $\hat{L}_p^{(B=I)}$ with the preliminary estimator $\hat{r}$ equal to the postulated $q = 3$. Since the stock return data are poorly predictable, but have non-trivial idiosyncratic cross-sectional correlation, the assumption $B = I_T$ is plausible, whereas $A = I_n$ is not. Further, since $\underline{L}_p$ does not perform well in our MC exercises, we restrict attention to estimators $\bar{L}_p$ and $\hat{L}_p^{(B=I)}$.

The left panel of Figure 6 reports $\bar{L}_p$ (solid line) and $\hat{L}_p^{(B=I)}$ (dotted line) nor-

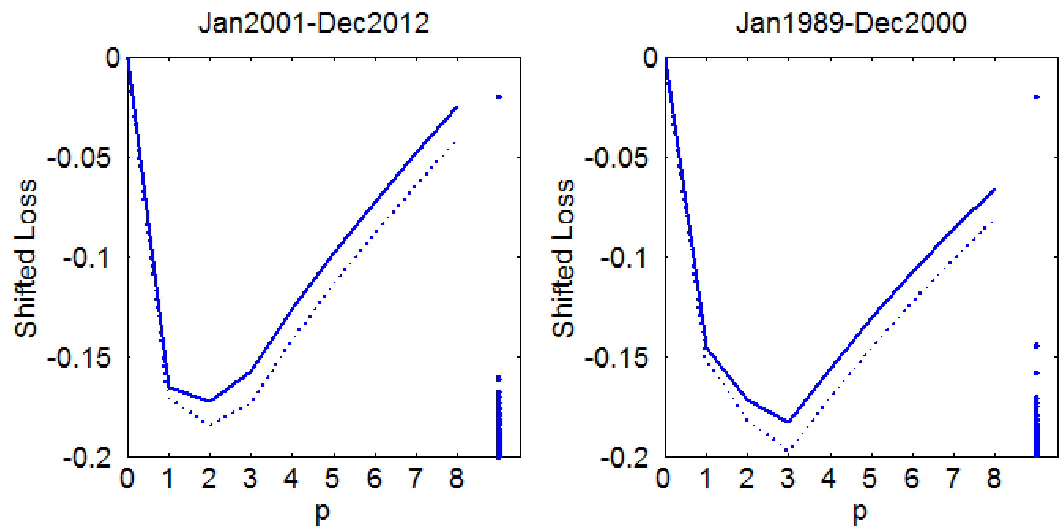

Figure 6: Estimated loss of the PC estimator of a factor model of excess returns. $\bar{L}_p$ – solid lines, $\hat{L}_p^{(B=I)}$ – dotted lines. $\hat{r}$ is set to 3.

malized to the units of the sample variance of the pooled excess return data, and shifted so that $\bar{L}_0 = \hat{L}_0^{(B=I)} = 0$. Both estimates of the loss function are minimized at $p = 2$, despite our forcing $\hat{r} = 3$. Moreover, estimating three instead of two factors wipes out all the benefit obtained from estimating two rather than one factor. In fact, according to $\bar{L}_p$ estimate, the marginal gain from estimating two rather than one factors is less than half the marginal loss from estimating three rather than two factors. Of course, the reason why estimating three factors is undesirable in these data, even after assuming that $q = 3$, is that the PC estimator of the third factor is too noisy to be useful. Note, however, that since we do not have standard errors of our loss estimates, these conclusions should be taken with a grain of salt.

Interestingly, the entire exercise repeated for the Jan1989-Dec2000 data yields different results. As the right panel of Figure 6 shows, for that time period, estimating three factors would have been beneficial from the point of view of minimizing the loss. Note that the gap between the first three and the fourth sample covariance eigenvalues (shown as dots with abscissas 9) in the Jan1989-Dec2000 period was much larger than that in the Jan2001-Dec2012 period. This can be interpreted as lower signal-to-noise ratio in the more recent period, which hurts the precision of the PC estimator and makes estimating the third factor useless.

6 Conclusion

In this paper, we study the effect of misspecification of the number of factors in approximate factor models on the quadratic loss from the estimation of the common component. We derive asymptotic approximations for the quadratic loss through the terms of order $O_{\mathrm{P}}(1/T) \sim O_{\mathrm{P}}(1/n)$ under both weak and strong factors asymptotics.

We develop several estimators of the loss, all of which are consistent under the strong factors asymptotics. The consistency under the weak factors asymptotics requires either no cross-sectional or no temporal correlation in the idiosyncratic terms. The estimators of the number of factors that minimize the proposed estimators are shown to be asymptotically loss efficient. When the idiosyncratic terms exhibit both cross-sectional and temporal correlation and factors are weak, we derive upper and lower bounds on the loss. The minimizers of these bounds bracket the loss-minimizing number of factors with asymptotic probability one.

Many important issues are not considered in this paper. As explained in Bai and Ng (2008, p. 95), static factor models are sufficiently flexible to accommodate dynamic factor models with loadings represented by lag polynomials of fixed finite order. However, the generalized dynamic factor models introduced by Forni et al (2000) cannot be represented in the form (1), and their study is left for future research.

Further, this paper does not derive the standard errors of the proposed loss estimators. Finding these standard errors and proposing methods of their estimation remains an important task for future studies.

Finally, the quadratic loss from the estimation of the common component is not the only interesting loss that can be considered in the factor models context. Many applications of factor models are related to diffusion index forecasts (Stock and Watson, 2006). From the point of view of these applications, a natural and interesting loss to consider is the squared forecast error.

Another interesting loss to consider would be the mean squared error loss of an IV estimator based on factors selected from a large dataset, as in Bai and Ng (2010). We speculate that weak factors asymptotic techniques developed in this paper may be useful for optimal selection of potentially weak instruments from a large number of possibly endogenous variables. Recently, Belloni et al (2012) showed that Lasso is successful in this context when instruments are strong. This raises a broader question of whether Lasso or similar sparsity-based methods can be useful to select factors in weak factor models.

Under strong factors, Caner and Han (2014) has shown how to use group bridge estimator to consistently estimate factor loadings and the true number of factors. Note that the strong factor assumptions imply that the model is approximately sparse in the principal components space. Precisely, the average explanatory power of the first r principal components, measured by $\mu_i (XX') / (nT)$, $i = 1, ..., r$, is bounded away from zero, whereas that of the further principal components, measured by $\mu_i (XX') / (nT)$, $i = r + 1, ..., n$, converges to zero asymptotically. Under weak factors, this approximate sparsity does not hold. The identification of the weak factors is based not on the relative negligibility of "idiosyncratic" eigenvalues, but on the clustering of this eigenvalues in a tightly packed group, so that $\mu_{r+1} (XX'/T) - \mu_{r+k} (XX'/T) \xrightarrow{p} 0$ for any fixed k. Hence, the "sparsity" in the weak factor models is related to the negligibility of the *gaps* between the adjacent "idiosyncratic" eigenvalues rather than to the negligibility of the normalized levels of these eigenvalues. In the future research it would be interesting to see whether and how the eigenvalue gap sparsity of weak factor models can be utilized by Lasso-type techniques.

7 Appendix

For any matrix A, let $\|A\|$ denote the spectral norm of A, that is $\|A\|$ equals the maximum singular value of A.

7.1 Primitive conditions for A3 (iii)

Proposition 6 *Let* $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{nt})'$, *where* ε_{it} *with* $i \in \mathbb{N}$ *and* $t \in \mathbb{Z}$ *are independent zero mean random variables with uniformly bounded fourth moments. Assump-*

tion A3 (iii) is satisfied for $e = [e_1, ..., e_T]$ *with*

$$e_t = \sum\nolimits_{j=0}^{\infty} \Psi_{nj} \varepsilon_{t-j},$$

where Ψ_{nj} *are* $n \times n$ *matrices such that* $\sum_{j=0}^{\infty} j \left\| \Psi_{nj} \right\|^2 < M$, *and* $\sum_{j=0}^{\infty} \left\| \Psi_{nj} \right\| < M$ *for an* $M < \infty$ *that does not depend on* n.

Proof: Our proof is similar to Moon and Weidner's (2010a) proof of their example (ii). We have

$$\left(\mu_1 \left(ee'/T\right)\right)^{1/2} = \|e\| / \sqrt{T} \leq \sum\nolimits_{j=0}^{T} \left\| \Psi_{nj} \right\| \left\| \varepsilon_{-j} \right\| / \sqrt{T} + \left\| r_{n,T} \right\|,$$

where $\varepsilon_{-j} = [\varepsilon_{1-j}, ..., \varepsilon_{T-j}]$ and $r_{n,T} = \sum_{j=T+1}^{\infty} \Psi_{nj} \varepsilon_{-j} / \sqrt{T}$. Obviously, for any $j = 0, ..., T$, $\left\| \varepsilon_{-j} \right\| \leq \|\varepsilon\|$, where $\varepsilon = [\varepsilon_{1-T}, ..., \varepsilon_T]$. As explained by Moon and Weidner (2010a), $\|\varepsilon\| / \sqrt{T} = O_{\mathrm{P}}(1)$. Therefore,

$$\left(\mu_1 \left(ee'/T\right)\right)^{1/2} \leq O_{\mathrm{P}}(1) \sum\nolimits_{j=0}^{T} \left\| \Psi_{nj} \right\| + \left\| r_{n,T} \right\| = O_{\mathrm{P}}(1) + \left\| r_{n,T} \right\|. \qquad (23)$$

Next, since the fourth moments of ε_{it} are uniformly bounded, and $E\varepsilon_{it}^2 \leq \left(E\varepsilon_{it}^4\right)^{1/2}$, the second moments of ε_{it} are uniformly bounded too. Let us denote the uniform bound on the second moments of ε_{it} as B. We have

$$\begin{aligned} E \left\| r_{n,T} \right\|^2 &\leq \sum_{i=1}^{n} \sum_{t=1}^{T} E\left(\left(r_{n,T}\right)_{it}^2\right) = \frac{1}{T} \sum_{i=1}^{n} \sum_{t=1}^{T} E \left(\sum_{j=T+1}^{\infty} \sum_{s=1}^{n} \left(\Psi_{nj}\right)_{is} \varepsilon_{s,t-j} \right)^2 \\ &\leq B \sum_{j=T+1}^{\infty} \left\| \Psi_{nj} \right\|_F^2 \leq \frac{B}{T} \sum_{j=T+1}^{\infty} j \left\| \Psi_{nj} \right\|_F^2, \end{aligned}$$

where $\|M\|_F$ denotes the Frobenius norm of matrix M. Since $\|\Psi_{nj}\|_F^2 \leq n\|\Psi_{nj}\|^2$ (see Horn and Johnson (1985), p. 314), we have

$$E\|r_{n,T}\|^2 \leq \frac{Bn}{T}\sum_{j=T+1}^{\infty} j\|\Psi_{nj}\|^2 = o(1).$$

Hence, $\|r_{n,T}\|^2 = o_{\mathrm{P}}(1)$, and $\|r_{n,T}\| = o_{\mathrm{P}}(1)$ too. Combining this with (23), we obtain $\mu_1(ee'/T) = O_{\mathrm{P}}(1)$.□

7.2 Two auxiliary Lemmas

In this subsection, we state and prove two auxiliary lemmas that will be used below. Let $e^{(j,k)} = \Lambda'_{\cdot j} eF_{\cdot k}/\sqrt{d_{jn}nT}$.

Lemma 2 *Under assumptions A1-A3, for $j \leq r$, as $n, T \rightarrow_c \infty$,*

$$\begin{aligned}\mu_j(X'X)/(nT) &= d_{jn} + 2\sqrt{d_{jn}/(nT)}e^{(j,j)} + F'_{\cdot j}e'eF_{\cdot j}/\left(nT^2\right) \\ &\quad + \Lambda'_{\cdot j}ee'\Lambda_{\cdot j}/\left(d_{jn}n^2T\right) + o_{\mathrm{P}}(1/T).\end{aligned}$$

Proof: Consider a decomposition

$$X'X/(nT) = M + M^{(1)}/\sqrt{T} + M^{(2)}/T, \tag{24}$$

where

$$M = FD_nF'/T,\ M^{(1)} = (F\Lambda'e + e'\Lambda F')/(n\sqrt{T}),\ \text{and}\ M^{(2)} = e'e/n.$$

We use Kato's (1980) theory to characterize the eigenvalues and eigenprojections of $X'X/(nT)$ as perturbations of those of M. Similar techniques were recently used in the analysis of the quasi maximum likelihood estimator in panel data models with interactive fixed effects by Moon and Weidner (2010). Let $R(z) = (M - zI_T)^{-1}$, $z \in \mathbb{C}$. Then, according to Kato (1980, p.78-79), for $1 \leq j \leq r$,

$$\mu_j(X'X)/(nT) = \mu_j(M) + \sum\nolimits_{s=1}^{\infty} \mu_j^{(s)} T^{-s/2} = d_{jn} + \sum\nolimits_{s=1}^{\infty} \mu_j^{(s)} T^{-s/2}, \qquad (25)$$

where

$$\mu_j^{(s)} = \sum\nolimits_{\nu_1+\ldots+\nu_p=s} \frac{(-1)^p}{2\pi p \mathrm{i}} \operatorname{tr} \int_\Gamma M^{(\nu_1)} R(z) \ldots M^{(\nu_p)} R(z) \, \mathrm{d}z$$

with ν_k, $k = 1, ..., p$, taking on only values one or two; $\mathrm{i} \in \mathbb{C}$ being the imaginary unit; and Γ being the circle in $\mathbb{C}$ with center at d_{jn} and radius $r_j = \min_{i=0,1} \{d_{j+i-1,n} - d_{j+i,n}\}/2$. Here, we define $d_{0,n}$ as $+\infty$ and $d_{r+1,n}$ as 0.

As explained by Kato (1980, p.88), the series in (25) are absolutely converging as long as $\sup_{z\in\Gamma} \sum_{i=1}^{2} T^{-i/2} \left\| M^{(i)} R(z) \right\| < 1$. By definition of Γ and $R(z)$, $\sup_{z\in\Gamma} \|R(z)\| = r_j^{-1}$. Therefore, a sufficient condition for the convergence is $\max_{i=1,2} \left\{ \left\| M^{(i)} \right\| \right\} < Tr_j/2$. We have

$$\begin{aligned} \left\| F\Lambda' e/(n\sqrt{T}) \right\| &\leq \|\Lambda\| \|F\| \|e\| / (n\sqrt{T}) \\ &= \|\Lambda'\Lambda/n\|^{1/2} \|F'F/T\|^{1/2} \|e'e/n\|^{1/2} = d_{1n}^{1/2} \|e'e/n\|^{1/2}. \end{aligned}$$

Therefore, by A3 (iii), $\left\| M^{(1)} \right\| = O_{\mathrm{P}}(1)$ and $\left\| M^{(2)} \right\| = O_{\mathrm{P}}(1)$. In particular, for any $\varepsilon > 0$ and any sequence $\{n, T\}$ such that $n, T \to_c \infty$, there exists $\bar{T} > 0$ such that $\Pr\left(\max_{i=1,2} \left\{ \left\| M^{(i)} \right\| \right\} < \bar{T} r_j/2 \right) > 1 - \varepsilon$ for all $T > \bar{T}$. That is, with probability larger than $1 - \varepsilon$, the convergence in (25) takes place for all $T > \bar{T}$. Furthermore, by Kato's (1980, p.89) formula (3.6), with the same probability, for all $T > \bar{T}$,

$$\left|\mu_j\left(X'X/\left(nT\right)\right)-d_{jn}-\mu_j^{(1)}/T^{1/2}-\mu_j^{(2)}/T\right|\leq\frac{r_jT^{-3/2}}{\bar{T}^{-1}\left(\bar{T}^{-1/2}-T^{-1/2}\right)},$$

which implies that

$$\mu_j\left(X'X/\left(nT\right)\right)=d_{jn}+\mu_j^{(1)}/T^{1/2}+\mu_j^{(2)}/T+o_{\mathrm{P}}\left(1/T\right). \tag{26}$$

Let $P_j = F_{\cdot j}\left(F_{\cdot j}'F_{\cdot j}\right)^{-1}F_{\cdot j}' = F_{\cdot j}F_{\cdot j}'/T$ be the eigenprojection corresponding to the j-th eigenvalue of M, $\mu_j\left(M\right)=d_{jn}$, and let P_0 be the projection on the subspace of $\mathbb{R}^T$ orthogonal to all columns of F. Kato (1980, p. 79) gives the following explicit formulae for $\mu_j^{(1)}$ and $\mu_j^{(2)}$:

$$\mu_j^{(1)}=\operatorname{tr}\left[M^{(1)}P_j\right]\text{ and }\mu_j^{(2)}=\operatorname{tr}\left[M^{(2)}P_j-M^{(1)}S_jM^{(1)}P_j\right], \tag{27}$$

where

$$S_j=\sum\nolimits_{k\neq j,k=1}^{r}P_k/\left(d_{kn}-d_{jn}\right)-P_0/d_{jn}. \tag{28}$$

Using (27) and the definition of $M^{(1)}$, we have

$$\mu_j^{(1)}=2\operatorname{tr}\left[F\Lambda'eF_{\cdot j}F_{\cdot j}'\right]/\left(nT^{3/2}\right)=2\Lambda_{\cdot j}'eF_{\cdot j}/\left(nT^{1/2}\right)=2\sqrt{d_{jn}}e^{(j,j)}/\sqrt{n}. \tag{29}$$

Further, straightforward algebra that employs (28), shows that

$$\begin{aligned}\operatorname{tr}\left[M^{(1)}S_jM^{(1)}P_j\right] &= \sum_{k\neq j,k=1}^{r}\frac{\left(\sqrt{d_{jn}}e^{(j,k)}+\sqrt{d_{kn}}e^{(k,j)}\right)^2}{n\left(d_{kn}-d_{jn}\right)}-\Lambda_{\cdot j}'eP_0e'\Lambda_{\cdot j}/\left(d_{jn}n^2\right)\\ &= \sum_{k\neq j,k=1}^{r}\frac{\left(\sqrt{d_{jn}}e^{(j,k)}+\sqrt{d_{kn}}e^{(k,j)}\right)^2}{n\left(d_{kn}-d_{jn}\right)}-\Lambda_{\cdot j}'ee'\Lambda_{\cdot j}/\left(d_{jn}n^2\right)\\ &\quad+\sum_{k=1}^{r}\left(e^{(j,k)}\right)^2/n.\end{aligned}$$

By A3 (ii), $e^{(j,k)} = O_{\mathrm{P}}(1)$ and $e^{(k,j)} = O_{\mathrm{P}}(1)$. Therefore, recalling that n and T are of the same order when $n, T \to_c \infty$, we get

$$\mathrm{tr}\left[M^{(1)} S_j M^{(1)} P_j\right] = -\Lambda'_{\cdot j} e e' \Lambda_{\cdot j} / \left(d_{jn} n^2\right) + O_{\mathrm{P}}\left(1/T\right).$$

Since $\mathrm{tr}\left[M^{(2)} P_j\right] = F'_{\cdot j} e' e F_{\cdot j} / nT$, we obtain

$$\mu_j^{(2)} = F'_{\cdot j} e' e F_{\cdot j} / nT + \Lambda'_{\cdot j} e e' \Lambda_{\cdot j} / \left(d_{jn} n^2\right) + O_{\mathrm{P}}\left(1/T\right). \tag{30}$$

Equalities (26), (29), and (30) imply the lemma.□

Similarly to P_j and P_0, defined in the above proof, let $Q_j = \Lambda_{\cdot j}\left(\Lambda'_{\cdot j}\Lambda_{\cdot j}\right)^{-1}\Lambda'_{\cdot j}$ be the projection on the space spanned by $\Lambda_{\cdot j}$, and let Q_0 be the projection on the subspace of $\mathbb{R}^n$ orthogonal to all columns of Λ. Further, let $\hat{Q}_j = \hat{\Lambda}_{\cdot j}\left(\hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j}\right)^{-1}\hat{\Lambda}'_{\cdot j}$ and let $\hat{P}_j = \hat{F}_{\cdot j}\left(\hat{F}'_{\cdot j}\hat{F}_{\cdot j}\right)^{-1}\hat{F}'_{\cdot j}$.

Lemma 3 *Let k and j be integers such that $0 < k, j \leq r$. Then, under assumptions A1-A3, as $n, T \to_c \infty$,*

$$\textit{(i)}\ \ \mathrm{tr}\left[P_k \hat{P}_j\right] = \begin{cases} 1 - \Lambda'_{\cdot j} e e' \Lambda_{\cdot j} / \left(d_{jn}^2 T n^2\right) + o_{\mathrm{P}}\left(1/T\right) & \text{if } k = j \\ o_{\mathrm{P}}\left(1/T\right) & \text{if } k \neq j \end{cases}, \ \textit{and}$$

$$\textit{(ii)}\ \ \mathrm{tr}\left[Q_k \hat{Q}_j\right] = \begin{cases} 1 - F'_{\cdot j} e' e F_{\cdot j} / \left(d_{jn} n T^2\right) + o_{\mathrm{P}}\left(1/T\right) & \text{if } k = j \\ o_{\mathrm{P}}\left(1/T\right) & \text{if } k \neq j \end{cases}.$$

Proof: Consider decomposition (24). According to Kato (1980, p.68),

$$\hat{P}_j = P_j + \sum\nolimits_{s=1}^{\infty} P_j^{(s)} / T^{s/2}, \tag{31}$$

where

$$P_j^{(s)} = -\sum\nolimits_{\nu_1+\ldots+\nu_p=s} \frac{(-1)^p}{2\pi i} \int_\Gamma R(z)\, M^{(\nu_1)} R(z)\, M^{(\nu_2)} \ldots M^{(\nu_p)} R(z)\, \mathrm{d}z \tag{32}$$

with ν_k, $k = 1, \ldots, p$, $R(z)$, and Γ defined as in the proof of Lemma 2. As in that proof, for any $\varepsilon > 0$ and any sequence $\{n, T\}$ such that $n, T \to_c \infty$, let $\bar{T}$ be such that $\Pr\left(\max_{i=1,2}\left\{\left\|M^{(i)}\right\|\right\} < \bar{T} r_j/2\right) > 1 - \varepsilon$ for all $T > \bar{T}$. Then, since $\sup_{z\in\Gamma} |R(z)| = 1/r_j$, with probability larger than $1-\varepsilon$,

$$\left\|P_j^{(s)}\right\| \leq \sum\nolimits_{\nu_1+\ldots+\nu_p=s} \frac{1}{2\pi} \int_\Gamma (1/r_j)^{p+1} \left(\bar{T} r_j/2\right)^p |\mathrm{d}z| = \sum\nolimits_{\nu_1+\ldots+\nu_p=s} \left(\bar{T}/2\right)^p.$$

Since ν_i may only be equal to one or two, there are no more than 2^s summands in the latter sum. Therefore, with probability larger than $1-\varepsilon$, $\left\|P_j^{(s)}\right\| \leq \left(2\bar{T}\right)^s$ for all $s = 1, 2, \ldots$ and all $T > \bar{T}$. Hence, by (31), with probability larger than $1-\varepsilon$, for all $T > \left(2\bar{T}\right)^2$,

$$\left\|\hat{P}_j - P_j - P_j^{(1)}/\sqrt{T} - P_j^{(2)}/T\right\| \leq \left(2\bar{T}/\sqrt{T}\right)^3 / \left(1 - 2\bar{T}/\sqrt{T}\right),$$

which implies that

$$\hat{P}_j = P_j + P_j^{(1)}/\sqrt{T} + P_j^{(2)}/T + o_{\mathrm{P}}(1/T), \tag{33}$$

where $T o_{\mathrm{P}}(1/T)$ converges to zero in probability in spectral norm.

Kato (1980, p.77) gives the following explicit formulae for $P_j^{(1)}$ and $P_j^{(2)}$:

$$P_j^{(1)} = -P_j M^{(1)} S_j - S_j M^{(1)} P_j, \text{ and} \tag{34}$$

$$P_j^{(2)} = -P_j M^{(2)} S_j - S_j M^{(2)} P_j + P_j M^{(1)} S_j M^{(1)} S_j \tag{35}$$
$$+ S_j M^{(1)} P_j M^{(1)} S_j + S_j M^{(1)} S_j M^{(1)} P_j - P_j M^{(1)} P_j M^{(1)} S_j^2$$
$$- P_j M^{(1)} S_j^2 M^{(1)} P_j - S_j^2 M^{(1)} P_j M^{(1)} P_j.$$

Using (33)-(35) and the fact that $P_j S_j = 0$, we obtain, for $j \leq r$,

$$\operatorname{tr}\left[P_j \hat{P}_j\right] = 1 - \operatorname{tr}\left[P_j M^{(1)} S_j^2 M^{(1)} P_j\right]/T + o_{\mathrm{P}}(1/T).$$

From the latter formula and definition (28) of S_j, we have

$$\begin{aligned} \operatorname{tr}\left[P_j \hat{P}_j\right] &= 1 - \sum_{k\neq j, k=1}^{r} \frac{\left(\sqrt{d_{kn}} e^{(k,j)} + \sqrt{d_{jn}} e^{(j,k)}\right)^2}{\left(d_{kn} - d_{jn}\right)^2 nT} \\ &\quad - \Lambda_{\cdot j}' e P_0 e' \Lambda_{\cdot j} / (d_{jn}^2 n^2 T) + o_{\mathrm{P}}(1/T). \end{aligned}$$

Assumption A3 (ii) implies that the second summand on the right hand side of the above equation is $o_{\mathrm{P}}(1/T)$, and hence,

$$\operatorname{tr}\left[P_j \hat{P}_j\right] = 1 - \Lambda_{\cdot j}' e P_0 e' \Lambda_{\cdot j} / (d_{jn}^2 n^2 T) + o_{\mathrm{P}}(1/T).$$

Since $P_0 = I_T - \sum_{k=1}^{r} P_k$, we have

$$\operatorname{tr}\left[P_j \hat{P}_j\right] = 1 - \Lambda_{\cdot j}' e e' \Lambda_{\cdot j} / (d_{jn}^2 n^2 T) + \sum_{k=1}^{r} \Lambda_{\cdot j}' e P_k e' \Lambda_{\cdot j} / (d_{jn}^2 T n^2) + o_{\mathrm{P}}(1/T).$$

Noting that $\Lambda_{\cdot j}' e P_k e' \Lambda_{\cdot j} = d_{jn} n \left(e^{(j,k)}\right)^2$ and using A3 (ii) one more time, we get

$$\operatorname{tr}\left[P_j \hat{P}_j\right] = 1 - \Lambda_{\cdot j}' e e' \Lambda_{\cdot j} / (d_{jn}^2 n^2 T) + o_{\mathrm{P}}(1/T). \tag{A13}$$

For $k \neq j$, using (33)-(35) and (28), we have

$$\operatorname{tr}\left[P_k\hat{P}_j\right] = \frac{1}{nT}\left(\sqrt{d_{jn}}e^{(j,k)} + \sqrt{d_{kn}}e^{(k,j)}\right)^2 / \left(d_{kn} - d_{jn}\right)^2 + o_{\mathrm{P}}\left(1/T\right).$$

By A3 (ii), the first term in the above sum is $o_{\mathrm{P}}\left(T^{-1}\right)$, and thus, for $k \neq j$,

$$\operatorname{tr}\left[P_k\hat{P}_j\right] = o_{\mathrm{P}}\left(T^{-1}\right). \tag{A14}$$

Lemma 3 (ii) follows from the symmetry of our model with respect to interchanging temporal and cross-sectional dimensions. The symmetry holds up to different normalizations of $\Lambda'\Lambda$ and $F'F$, which explains the "extra d_{jn}" in the denominator of the formula for $\operatorname{tr}\left[P_k\hat{P}_j\right]$ relative to that for $\operatorname{tr}\left[Q_k\hat{Q}_j\right]$.□

7.3 Proof of Proposition 1.

Opening brackets in (2) and using the definition of $\hat{\Lambda}_{1:p}$ and $\hat{F}_{1:p}$ and assumption A1, we obtain

$$\begin{aligned} L_p &= \operatorname{tr}[\hat{\Lambda}_{1:p}'\hat{\Lambda}_{1:p}]/n + \operatorname{tr}\left[\Lambda'\Lambda\right]/n - 2\operatorname{tr}[\hat{\Lambda}_{1:p}\hat{F}_{1:p}'F\Lambda']/\left(nT\right) \\ &= \sum_{j=1}^{p}\mu_j\left(X'X/\left(nT\right)\right) + \sum_{j=1}^{r}d_{jn} - 2\sum_{k=1}^{r}\sum_{j=1}^{p}(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j})(F_{\cdot k}'\hat{F}_{\cdot j})/\left(nT\right). \end{aligned} \tag{36}$$

Let us consider the last term of (36). Since

$$\begin{aligned} \operatorname{tr}[Q_k\hat{Q}_j] &= \operatorname{tr}[\Lambda_{\cdot k}\left(\Lambda_{\cdot k}'\Lambda_{\cdot k}\right)^{-1}\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j}\left(\hat{\Lambda}_{\cdot j}'\hat{\Lambda}_{\cdot j}\right)^{-1}\hat{\Lambda}_{\cdot j}'] \\ &= \left(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j}\right)^2/\left[\left(\Lambda_{\cdot k}'\Lambda_{\cdot k}\right)\left(\hat{\Lambda}_{\cdot j}'\hat{\Lambda}_{\cdot j}\right)\right] = \left(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j}\right)^2/\left[\|\Lambda_{\cdot k}\|^2\,\|\hat{\Lambda}_{\cdot j}\|^2\right], \end{aligned}$$

and since

$$\begin{aligned}\operatorname{tr}[P_k\hat{P}_j] &= \operatorname{tr}[F_{\cdot k}\left(F_{\cdot k}'F_{\cdot k}\right)^{-1}F_{\cdot k}'\hat{F}_{\cdot j}\left(\hat{F}_{\cdot j}'\hat{F}_{\cdot j}\right)^{-1}\hat{F}_{\cdot j}'] \\ &= \left(F_{\cdot k}'\hat{F}_{\cdot j}\right)^2/\left[\left(F_{\cdot k}'F_{\cdot k}\right)\left(\hat{F}_{\cdot j}'\hat{F}_{\cdot j}\right)\right] = \left(F_{\cdot k}'\hat{F}_{\cdot j}\right)^2/T^2,\end{aligned}$$

we have

$$\begin{aligned}\left|(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j})(F_{\cdot k}'\hat{F}_{\cdot j})\right|^2 &= \|\Lambda_{\cdot k}\|^2\,||\hat{\Lambda}_{\cdot j}||^2T^2\operatorname{tr}[Q_k\hat{Q}_j]\operatorname{tr}[P_k\hat{P}_j] \\ &= d_{kn}nT^2\mu_j\left(X'X/T\right)\operatorname{tr}[Q_k\hat{Q}_j]\operatorname{tr}[P_k\hat{P}_j]. \end{aligned}\qquad (37)$$

The latter equality holds because $\|\Lambda_{\cdot k}\|^2 = d_{kn}n$ by assumption, and

$$||\hat{\Lambda}_{\cdot j}||^2 = \hat{F}_{\cdot j}'X'X\hat{F}_{\cdot j}/T^2 = \left(\hat{F}_{\cdot j}'/\sqrt{T}\right)\left(X'X/T\right)\left(\hat{F}_{\cdot j}/\sqrt{T}\right) = \mu_j\left(X'X/T\right)$$

by definition of the principal components estimators $\hat{F}_{1:p}$ and $\hat{\Lambda}_{1:p}$ given in the introduction. For $j \leq r$ and $j \neq k$, by Lemmas 2 and 3, $\mu_j\left(X'X/T\right)\operatorname{tr}[Q_k\hat{Q}_j]\operatorname{tr}[P_k\hat{P}_j] = o_{\mathrm{P}}\left(1/T\right)$. Therefore

$$(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j})(F_{\cdot k}'\hat{F}_{\cdot j})/\left(nT\right) = o_{\mathrm{P}}\left(1/T\right). \qquad (38)$$

This equality holds also for $j > r$. Indeed, according to a singular value analog of Weyl's eigenvalue inequalities (see Theorem 3.3.16 of Horn and Johnson (1991)), for any $n \times T$ matrices A and B,

$$\mu_{i+s-1}^{1/2}\left(\left(A+B\right)\left(A+B\right)'\right) \leq \mu_i^{1/2}\left(AA'\right) + \mu_s^{1/2}\left(BB'\right), \qquad (39)$$

where $1 \leq i, s \leq \min\{n, T\}$. Setting $A = F\Lambda'/\sqrt{nT}$, $B = e'/\sqrt{nT}$, $i = r+1$, and $s = j - r$, and noting that $\mu_{r+1}(F\Lambda'\Lambda F'/(nT)) = 0$, we get

$$\mu_j(X'X)/(nT) \leq \mu_{j-r}(e'e)/(nT). \tag{40}$$

Similarly, setting $A = -F\Lambda'/\sqrt{nT}$, $B = X'/\sqrt{nT}$, $i = r+1$, and $s = j$, we get

$$\mu_j(X'X)/(nT) \geq \mu_{j+r}(e'e)/(nT). \tag{41}$$

Further, for $j > r$, we have $0 \leq \operatorname{tr}[P_k\hat{P}_j] \leq \operatorname{tr}[P_k\hat{P}_j] + \operatorname{tr}[P_k(I_T - \hat{P}_j - \hat{P}_k)] = 1 - \operatorname{tr}[P_k\hat{P}_k]$. Hence, by Lemma 3, $\operatorname{tr}[P_k\hat{P}_j] = O_{\mathrm{P}}(1/T)$. Similarly, $\operatorname{tr}[Q_k\hat{Q}_j] = O_{\mathrm{P}}(1/T)$. The latter two equalities together with (37) and the fact that, by (40), $\mu_j(X'X/T) \leq \mu_1(e'e/T) = O_{\mathrm{P}}(1)$ imply (38).

Using (38) together with (36), we obtain

$$L_p = \sum_{j=1}^{p} \mu_j(X'X/(nT)) + \sum_{j=1}^{r} d_{jn} - 2\sum_{k=1}^{\min\{p,r\}} (\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot k})(F'_{\cdot k}\hat{F}_{\cdot k})/(nT) + o_{\mathrm{P}}(1/T). \tag{42}$$

Now, by (37), $(\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot k})(F'_{\cdot k}\hat{F}_{\cdot k}) = \pm(d_{kn}nT^2\mu_k(X'X/T)\operatorname{tr}[Q_k\hat{Q}_k]\operatorname{tr}[P_k\hat{P}_k])^{1/2}$. On the other hand,

$$\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot k} = \Lambda'_{\cdot k}X\hat{F}_{\cdot k}/T = d_{kn}nF'_{\cdot k}\hat{F}_{\cdot k}/T + \Lambda'_{\cdot k}e\hat{F}_{\cdot k}/T = d_{kn}nF'_{\cdot k}\hat{F}_{\cdot k}/T + O_{\mathrm{P}}(1). \tag{43}$$

To see that the latter equality holds, note that

$$\left\|\hat{F}_{\cdot k} - F_{\cdot k}(F'_{\cdot k}\hat{F}_{\cdot k}/T)\right\|^2 = T\left(1 - \operatorname{tr}[P_k\hat{P}_k]\right) = O_{\mathrm{P}}(1),$$

by Lemma 3. Therefore, $\Lambda_{\cdot k}' e\hat{F}_{\cdot k}/T = \Lambda_{\cdot k}' eF_{\cdot k}(F_{\cdot k}'\hat{F}_{\cdot k}/T)/T + O_{\mathrm{P}}(1) = O_{\mathrm{P}}(1)$, where the last equality follows from A3 (ii) and the fact that $\left|F_{\cdot k}'\hat{F}_{\cdot k}/T\right| = (\mathrm{tr}[P_k\hat{P}_k])^{1/2} = O_{\mathrm{P}}(1)$, by Lemma 3. Equality (43) implies that $(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot k})(F_{\cdot k}'\hat{F}_{\cdot k})$ is positive with probability approaching one as $n, T \to_c \infty$. Hence,

$$(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot k})(F_{\cdot k}'\hat{F}_{\cdot k}) = (d_{kn}nT^2\mu_k\left(X'X/T\right)\mathrm{tr}[Q_k\hat{Q}_k]\,\mathrm{tr}[P_k\hat{P}_k])^{1/2}. \tag{44}$$

Using (42), (44), and Lemmas 2 and 3, we obtain

$$\begin{aligned} L_p &= \sum_{j=1}^{p}\mu_j\left(X'X/\left(nT\right)\right) + \sum_{j=1}^{r} d_{jn} - \\ &\quad 2\sum_{k=1}^{\min\{p,r\}} d_{kn}\left(1 + \frac{e^{(k,k)}}{\sqrt{d_{kn}nT}} + \frac{F_{\cdot k}'e'eF_{\cdot k}}{2d_{kn}nT^2} + \frac{\Lambda_{\cdot k}'ee'\Lambda_{\cdot k}}{2d_{kn}^2n^2T}\right)\times \\ &\quad \left(1 - F_{\cdot k}'e'eF_{\cdot k}/\left(2d_{kn}nT^2\right)\right)\left(1 - \Lambda_{\cdot k}'ee'\Lambda_{\cdot k}/\left(2d_{kn}^2Tn^2\right)\right) + o_{\mathrm{P}}\left(1/T\right) \\ &= \sum_{j=1}^{p}\mu_j\left(X'X/\left(nT\right)\right) + \sum_{j=1}^{r} d_{jn} - 2\sum_{k=1}^{\min\{p,r\}}\left(d_{kn} + \sqrt{\frac{d_{kn}}{nT}}e^{(k,k)}\right) + o_{\mathrm{P}}\left(1/T\right) \end{aligned}$$

From this and Lemma 2, we conclude that $L_p = L_p^{(1)} + o_{\mathrm{P}}\left(1/T\right)$, where

$$L_p^{(1)} = \begin{cases} \sum_{j=p+1}^{r} d_{jn} + \sum_{j=1}^{p}\left(F_{\cdot j}'e'eF_{\cdot j}/\left(nT^2\right) + \Lambda_{\cdot j}'ee'\Lambda_{\cdot j}/\left(d_{jn}n^2T\right)\right) & \text{if } p \le r \\ L_r^{(1)} + \sum_{j=r+1}^{p}\mu_j\left(X'X/\left(nT\right)\right) & \text{if } p > r \end{cases}.$$

The statement of Proposition 1 follows from the latter equality and the observation that

$$\sum_{j=1}^{p}\left(F_{\cdot j}'e'eF_{\cdot j}/T + \Lambda_{\cdot j}'ee'\Lambda_{\cdot j}/\left(d_{jn}n\right)\right) = \mathrm{tr}\left[eP_{1:p}e' + e'Q_{1:p}e\right].\square$$

7.4 Proof of Corollary 1.

As shown by Yin et al (1988), the assumption that the elements of e are i.i.d. zero mean random variables with variance σ^2 and a finite fourth moment implies that $\operatorname{plim} \mu_1(ee')/T \overset{a.s.}{\to} \sigma^2(1+\sqrt{c})^2$ as $n, T \to_c \infty$. On the other hand, the empirical distribution of the eigenvalues of ee'/T almost surely weakly converges to the Marchenko-Pastur distribution (see Bai, 1999, Theorem 2.5), which has $\sigma^2(1+\sqrt{c})^2$ as the upper boundary of its support. These two facts imply that, for any fixed j, $\operatorname{plim} \mu_j(ee')/T \overset{a.s.}{\to} \sigma^2(1+\sqrt{c})^2$ as $n, T \to_c \infty$. Therefore, inequalities (40) and (41) allow us to conclude that, for any fixed $j > r$, $\mu_j(X'X)/T = \sigma^2\left(1+\sqrt{n/T}\right)^2 + o_{\mathrm{P}}(1)$. The rest of the proof is elementary, and we omit it to save space.□

7.5 Proof of Lemma 1.

Our proof of Lemma 1 relies on Theorem 1 of Onatski (2012), which is established in Onatski (2012a) under assumptions A1w-A3w. In Onatski's (2012a) proof, the Gaussianity of ε is used solely to show that $\tilde{X} = U_A' X V_B'$ has the form

$$\tilde{X} = \sum\nolimits_{i=1}^{r} \tilde{\Lambda}_{\cdot i}\tilde{F}_{\cdot i} + \mathcal{A}_0 \eta \mathcal{B}_0, \tag{45}$$

where

$$\tilde{\Lambda}_{\cdot i} = e_i^{(n)} \delta_{in}^{1/2} \text{ and } \tilde{F}_{\cdot i} = e_i^{(T)} \sqrt{T} \tag{46}$$

with $e_i^{(n)}$ and $e_i^{(T)}$ being the i-th columns of I_n and I_T, respectively, and η being an $n \times T$ matrix with i.i.d. $N(0, \sigma^2)$ elements. However, the rest of that proof remains valid as long as η has i.i.d. elements η_{it} such that $E\eta_{it} = 0$, $E\eta_{it}^2 = \sigma^2$, and $E\eta_{it}^4 < \infty$.

But (45) and (46) are automatically satisfied under A1w, A2w (nG), and A3w with $\eta = \varepsilon$. Hence, Onatski's (2012) Theorem 1 valid not only under A1w-A3w but also under A1w, A2w (nG), and A3w.

For any $x \geq 0$, consider a system of equations in $u > \bar{x}_A$ and $v > \bar{x}_B$

$$\begin{cases} v = xg_1(u) \\ u = xg_2(v) \end{cases}, \tag{47}$$

where $g_1(u) = \left(c \int \lambda u / (u - \lambda) \, \mathrm{d}\mathcal{G}_A(\lambda)\right)^{-1}$ and $g_2(v) = \left(\int \lambda v / (v - \lambda) \, \mathrm{d}\mathcal{G}_B(\lambda)\right)^{-1}$. Direct differentiation shows that $g_1(u)$ is strictly increasing and concave on $u > \bar{x}_A$. Moreover, by assumption A3w, $\lim_{u \downarrow \bar{x}_A} g_1(u) = 0$ and $\lim_{u \to \infty} g_1(u) = 1/c$ (here, notation $u \downarrow \bar{x}_A$ means that u converge to $\bar{x}_A$ from above). Similarly, $g_2(v)$ is strictly increasing and concave on $v > \bar{x}_B$ with $\lim_{v \downarrow \bar{x}_B} g_2(v) = 0$ and $\lim_{v \to \infty} g_2(v) = 1$. These facts imply that there exists $\bar{x} > 0$ such that the curves defined by the equations of (47) do not intersect in the domain $\{u > \bar{x}_A, v > \bar{x}_B\}$ for any $x < \bar{x}$, the curves touch each other at one point $(\bar{u}, \bar{v})$ when $x = \bar{x}$, and intersect at two points (u_{1x}, v_{1x}) and (u_{2x}, v_{2x}), where $u_{2x} > u_{1x}$ and $v_{2x} > v_{1x}$, when $x > \bar{x}$. As $x \downarrow \bar{x}$, $(u_{2x}, v_{2x}) \to (\bar{u}, \bar{v})$, and as $x \to \infty$, u_{2x} and v_{2x} diverge to ∞.

Theorem 1 (iii) of Onatski (2012) links solutions of system (47) to function $f(z)$, defined by $f(\delta_j/\sigma^2) = \operatorname{plim} \mu_j(X'X)/(\sigma^2 T) = \operatorname{plim} \hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j}/\sigma^2$, as follows. For $z > \bar{z}$, where $\bar{z} = \bar{x}(1 - \bar{u}^{-1})(1 - \bar{v}^{-1})$, $f(z)$ equals the unique $x > \bar{x}$ such that $z = x\left(1 - u_{2x}^{-1}\right)\left(1 - v_{2x}^{-1}\right)$. Further, for $0 \leq z \leq \bar{z}$, $f(z)$ is fixed at $\bar{x} = \operatorname{plim} \mu_1(ee')/T$, and the latter probability limit is well defined. Statement (i) of Lemma 1, where $\bar{\delta} = \bar{z}\sigma^2$, follows immediately.

To establish the rest of Lemma 1, we study the function

$$g(x) = x\left(1 - u_{2x}^{-1}\right)\left(1 - v_{2x}^{-1}\right), \text{ for } x > \bar{x}.$$

Since (u_{2x}, v_{2x}) is the "larger" of the two intersection points of the graphs of concave functions $xg_1(u)$ and $xg_2(v)$ (in the coordinate plane (u,v)), we must have $\frac{\partial}{\partial u}[xg_1(u)] < 1/\frac{\partial}{\partial v}[xg_2(v)]$ at $(u,v) = (u_{2x}, v_{2x})$. This condition implies that

$$\det\begin{pmatrix} \frac{\partial}{\partial u}(v - xg_1(u)) & \frac{\partial}{\partial v}(v - xg_1(u)) \\ \frac{\partial}{\partial u}(u - xg_2(v)) & \frac{\partial}{\partial v}(u - xg_2(v)) \end{pmatrix} = \Delta < 0$$

at $(u,v) = (u_{2x}, v_{2x})$. Therefore, the implicit function theorem (see Krantz (1992), Theorem 1.4.11) applies, and u_{2x} and v_{2x} are analytic functions of x on $x > \bar{x}$.

Differentiating both sides of the identities $v_{2x} - xg_1(u_{2x}) = 0$ and $u_{2x} - xg_2(v_{2x}) = 0$ with respect to x and solving for $\mathrm{d}u_{2x}/\mathrm{d}x$ and $\mathrm{d}v_{2x}/\mathrm{d}x$, we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}u_{2x} &= (-\Delta)^{-1}\left(g_2(u_{2x}) + xg_1(v_{2x})g_2'(v_{2x})\right) > 0, \\ \frac{\mathrm{d}}{\mathrm{d}x}v_{2x} &= (-\Delta)^{-1}\left(g_1(v_{2x}) + xg_2(u_{2x})g_1'(u_{2x})\right) > 0. \end{aligned}$$

Therefore, $g(x)$ is strictly increasing and differentiable on $x > \bar{x}$. Since $f(z)$, $z > \bar{z}$, is the inverse function of $g(x)$, $x > \bar{x}$, we conclude that $f(z)$ is strictly increasing and differentiable on $z > \bar{z}$. Statement (ii) of Lemma 1 follows because $g(x) < x$, $\lim_{x \downarrow \bar{x}} g(x) = \bar{z}$, and $\lim_{x \to \infty} g(x)/x = 1$.

Note that $\mathrm{d}\ln f(z)/\mathrm{dln}z = 1/\left[\mathrm{d}\ln g(x)/\mathrm{d}\ln x\right]$, where $z = x\left(1 - u_{2x}^{-1}\right)\left(1 - v_{2x}^{-1}\right)$. Therefore, to establish (iii), it is enough to prove that $\mathrm{d}\ln g(x)/\mathrm{dln}x$ is decreasing on $x > \bar{x}$ and $\mathrm{d}\ln g(x)/\mathrm{d}\ln x \to 1$ as $x \to \infty$. From the definition of $g(x)$, we get

$$\mathrm{d}\ln g(x)/\mathrm{d}\ln x = 1 + \frac{1}{u_{2x}-1}\mathrm{d}\ln u_{2x}/\mathrm{d}\ln x + \frac{1}{v_{2x}-1}\mathrm{d}\ln v_{2x}/\mathrm{d}\ln x. \tag{48}$$

Since u_{2x} and v_{2x} are increasing functions of x, and since they diverge to infinity as $x \to \infty$, it is enough to prove that $\mathrm{d}\ln u_{2x}/\mathrm{d}\ln x$ and $\mathrm{d}\ln v_{2x}/\mathrm{d}\ln x$ are decreasing functions on $x > \bar{x}$.

Straightforward algebra shows that

$$\mathrm{d}\ln u_{2x}/\mathrm{d}\ln x = (1+V(x))/(1-U(x)V(x)), \text{ and} \tag{49}$$

$$\mathrm{d}\ln v_{2x}/\mathrm{d}\ln x = (1+U(x))/(1-U(x)V(x)), \tag{50}$$

where

$$V(x) = \int \frac{\lambda^2}{(v_{2x}-\lambda)^2}\mathrm{d}\mathcal{G}_B(\lambda) \Big/ \int \frac{\lambda}{v_{2x}-\lambda}\mathrm{d}\mathcal{G}_B(\lambda), \text{ and} \tag{51}$$

$$U(x) = \int \frac{\lambda^2}{(u_{2x}-\lambda)^2}\mathrm{d}\mathcal{G}_A(\lambda) \Big/ \int \frac{\lambda}{u_{2x}-\lambda}\mathrm{d}\mathcal{G}_A(\lambda). \tag{52}$$

Hence, it is enough to prove that $V(x)$ and $U(x)$ are decreasing functions on $x > \bar{x}$. Furthermore, since v_{2x} and u_{2x} are increasing functions of x on $x > \bar{x}$ it is enough to prove that

$$\frac{\mathrm{d}}{\mathrm{d}v}\ln\int \frac{\lambda^2}{(v-\lambda)^2}\mathrm{d}\mathcal{G}_B(\lambda) - \frac{\mathrm{d}}{\mathrm{d}v}\ln\int \frac{\lambda}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) < 0 \text{ for } v > \bar{v}, \text{ and} \tag{53}$$

$$\frac{\mathrm{d}}{\mathrm{d}u}\ln\int \frac{\lambda^2}{(u-\lambda)^2}\mathrm{d}\mathcal{G}_A(\lambda) - \frac{\mathrm{d}}{\mathrm{d}u}\ln\int \frac{\lambda}{u-\lambda}\mathrm{d}\mathcal{G}_A(\lambda) < 0 \text{ for } u > \bar{u}. \tag{54}$$

Below, we will establish (53). The proof of (54) is the same after v is replaced by u and $\mathcal{G}_B(\lambda)$ is replaced by $\mathcal{G}_A(\lambda)$.

For any $v > \bar{v}$, consider a function

$$h(\lambda) = \frac{1}{c_1}\frac{\lambda}{v-\lambda} - \frac{1}{c_2}\frac{\lambda^2}{(v-\lambda)^2},$$

where $\lambda \in (-\infty, v)$, $c_1 = \int \frac{\lambda}{v-\lambda} \mathrm{d}\mathcal{G}_B(\lambda)$, and $c_2 = \int \frac{\lambda^2}{(v-\lambda)^2} \mathrm{d}\mathcal{G}_B(\lambda)$. We have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} h(\lambda) = \frac{v}{(v-\lambda)^2}\left(\frac{1}{c_1} - \frac{1}{c_2}\frac{2\lambda}{v-\lambda}\right)$$

so that $\frac{\mathrm{d}}{\mathrm{d}\lambda} h(\lambda)$ is positive for $\lambda \in [0, vc_2/(2c_1 + c_2))$ and negative for $\lambda \in (vc_2/(2c_1+c_2), v)$. Since $h(0) = 0$ and $\int h(\lambda)\, \mathrm{d}\mathcal{G}_B(\lambda) = 0$, there must therefore exist $\tilde{\lambda} \in [\underline{x}_B, \bar{x}_B]$ such that $h(\lambda) \geq 0$ for $\lambda \in [\underline{x}_B, \tilde{\lambda}]$ and $h(\lambda) \leq 0$ for $\lambda \in [\tilde{\lambda}, \bar{x}_B]$. Hence, since $v > \bar{v} > \bar{x}_B$, we have

$$\begin{aligned}
\int h(\lambda)\frac{1}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) &= \int_{\underline{x}_B}^{\tilde{\lambda}} h(\lambda)\frac{1}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) + \int_{\tilde{\lambda}}^{\bar{x}_B} h(\lambda)\frac{1}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) \\
&\leq \int_{\underline{x}_B}^{\tilde{\lambda}} h(\lambda)\frac{1}{v-\tilde{\lambda}}\mathrm{d}\mathcal{G}_B(\lambda) + \int_{\tilde{\lambda}}^{\bar{x}_B} h(\lambda)\frac{1}{v-\tilde{\lambda}}\mathrm{d}\mathcal{G}_B(\lambda) \\
&= \frac{1}{v-\tilde{\lambda}}\int h(\lambda)\,\mathrm{d}\mathcal{G}_B(\lambda) = 0.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\int h(\lambda)\frac{1}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}v}\ln\int\frac{\lambda^2}{(v-\lambda)^2}\mathrm{d}\mathcal{G}_B(\lambda) - \frac{\mathrm{d}}{\mathrm{d}v}\ln\int\frac{\lambda}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda) \\
&> \frac{\mathrm{d}}{\mathrm{d}v}\ln\int\frac{\lambda^2}{(v-\lambda)^2}\mathrm{d}\mathcal{G}_B(\lambda) - \frac{\mathrm{d}}{\mathrm{d}v}\ln\int\frac{\lambda}{v-\lambda}\mathrm{d}\mathcal{G}_B(\lambda).
\end{aligned}$$

Therefore, (53) holds.□

7.6 Proof of Proposition 2.

Similarly to the proof of Proposition 1, we start from the identity

$$\begin{aligned} L_p &= \sum_{j=1}^{p} \mu_j \left(X'X/ \left(nT \right) \right) + \sum_{j=1}^{r} d_{jn} - 2 \sum_{k=1}^{r} \sum_{j=1}^{p} (\Lambda'_{\cdot k} \hat{\Lambda}_{\cdot j})(F'_{\cdot k} \hat{F}_{\cdot j})/ \left(nT \right) \\ &= \sum_{j=1}^{p} \mu_j \left(X'X/ \left(nT \right) \right) + \sum_{j=1}^{r} \delta_{jn}/n - 2 \sum_{k=1}^{r} \sum_{j=1}^{p} \sqrt{\delta_{kn} \hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j}} \hat{\alpha}_{kj} \hat{\beta}_{kj}/n, \end{aligned}$$

where $\hat{\alpha}_{kj} = \Lambda'_{\cdot k} \hat{\Lambda}_{\cdot j}/(\|\Lambda_{\cdot k}\| \, ||\hat{\Lambda}_{\cdot j}||)$ and $\hat{\beta}_{kj} = F'_{\cdot k} \hat{F}_{\cdot j}/(\|F_{\cdot k}\| \, ||\hat{F}_{\cdot j}||)$. By Theorem 1 of Onatski (2012) (which, as shown in the above proof of Lemma 1, is valid not only under A1w-A3w, but also under A1w, A2w (nG), and A3w), for $k \neq j$, $\text{plim}\, \hat{\alpha}_{kj} = 0$, $\text{plim}\, \hat{\beta}_{kj} = 0$, and $\hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j} = O_{\mathrm{P}}(1)$. Therefore, we have

$$L_p = \sum_{j=1}^{p} \mu_j \left(X'X/ \left(nT \right) \right) + \sum_{j=1}^{r} \delta_{jn}/n - 2 \sum_{j=1}^{p} \sqrt{\delta_{jn} \hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j}} \hat{\alpha}_{jj} \hat{\beta}_{jj}/n + o_{\mathrm{P}}(1/T) . \quad (55)$$

Let q be the largest $p \in \{0, 1, ..., r\}$ such that $\delta_p > \bar{\delta}$. For $j \leq q$, by Theorem 1 of Onatski (2012),

$$\text{plim}(\sqrt{\delta_{jn} \hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j}} \hat{\alpha}_{jj} \hat{\beta}_{jj}) = \sqrt{\frac{\delta_j \sigma^2 f \left(\delta_j / \sigma^2 \right)}{\left(1 + \psi_j \theta_j \right) \left(1 + \psi_j \omega_j \right)}}, \quad (56)$$

where

$$\begin{aligned} \psi_j &= u_{2x}^{-1} v_{2x}^{-1} \left(\mathrm{d} \ln \left(u_{2x} v_{2x} \right) / \mathrm{d} \ln x - 1 \right), \\ \theta_j &= u_{2x} - \frac{\int \lambda / \left(u_{2x} - \lambda \right) \mathrm{d} \mathcal{G}_A \left(\lambda \right)}{\int \lambda / \left(u_{2x} - \lambda \right)^2 \mathrm{d} \mathcal{G}_A \left(\lambda \right)} \frac{u_{2x} - v_{2x}}{u_{2x} - 1}, \end{aligned}$$

$$\omega_j = v_{2x} - \frac{\int \lambda / (v_{2x} - \lambda) \, \mathrm{d}\mathcal{G}_B(\lambda)}{\int \lambda / (v_{2x} - \lambda)^2 \, \mathrm{d}\mathcal{G}_B(\lambda)} \frac{v_{2x} - u_{2x}}{v_{2x} - 1},$$

$x = f(\delta_j/\sigma^2)$, and u_{2x} and v_{2x} are as defined in the proof of Lemma 1. The above definitions of ψ_j, θ_j and ω_j are equivalent to those given in Onatski (2012), which can be shown using system of equations (7) in that paper.

From definitions (51) and (52) of $V(x)$ and $U(x)$, we have

$$\frac{\int \lambda / (u_{2x} - \lambda) \, \mathrm{d}\mathcal{G}_A(\lambda)}{\int \lambda / (u_{2x} - \lambda)^2 \, \mathrm{d}\mathcal{G}_A(\lambda)} = \frac{u_{2x}}{1 + U(x)} \text{ and } \frac{\int \lambda / (v_{2x} - \lambda) \, \mathrm{d}\mathcal{G}_B(\lambda)}{\int \lambda / (v_{2x} - \lambda)^2 \, \mathrm{d}\mathcal{G}_B(\lambda)} = \frac{v_{2x}}{1 + V(x)}.$$

Using these equalities, the above definitions of ψ_j, θ_j, ω_j, x, and equalities (48), (49) and (50), we obtain

$$\left(1 + \psi_j \theta_j\right)\left(1 + \psi_j \omega_j\right) = x^{-1} g(x) \left(\mathrm{d} \ln g(x) / \mathrm{d} \ln x\right)^2,$$

where $g(x)$ is the inverse function of $f(z)$. Together with (56) and the fact that $g(x) = \delta_j/\sigma^2$, this implies that

$$\begin{aligned} \mathrm{plim}(\sqrt{\delta_{jn} \hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j}} \hat{\alpha}_{jj} \hat{\beta}_{jj}) &= \sqrt{\frac{\delta_j \sigma^2 x^2}{g(x) \left(\mathrm{d} \ln g(x) / \mathrm{d} \ln x\right)^2}} \\ &= \sigma^2 f\left(\delta_j/\sigma^2\right) \mathrm{d} \ln f(\delta_j/\sigma^2) / \mathrm{d} \ln\left(\delta_j/\sigma^2\right) \\ &= \delta_j f'\left(\delta_j/\sigma^2\right). \end{aligned}$$

Therefore, since $z f'(z)$ is an analytic function of z on $z > \bar{z} = \bar{\delta}/\sigma^2$,

$$\mathrm{plim}(\sqrt{\delta_{jn} \hat{\Lambda}'_{\cdot j} \hat{\Lambda}_{\cdot j}} \hat{\alpha}_{jj} \hat{\beta}_{jj}) = \delta_{jn} f'\left(\delta_{jn}/\sigma^2\right) + o_{\mathrm{P}}(1).$$

For $p \leq q$, the statement of Proposition 2 follows from the latter equality and (55).

For $j > q$, by Theorem 1 of Onatski (2012), $\text{plim}(\sqrt{\delta_{jn}\hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j}}\hat{\alpha}_{jj}\hat{\beta}_{jj}) = 0$. Hence, for $p > q$, the statement of Proposition 2 follows from (55) as well. □

7.7 Proof of Corollary 2.

As in the proof of Corollary 1, for any fixed $j > r$, $\mu_j(X'X)/T = \sigma^2\left(1 + \sqrt{n/T}\right)^2 + o_{\mathrm{P}}(1)$. This fact, equation (8), and Proposition 2 imply Corollary 2.□

7.8 Proof of Proposition 3.

First, we prove (i). By Proposition 1, for $p > r$,

$$L_p - L_r = \sum_{j=r+1}^{p} \mu_j/n + o_{\mathrm{P}}(1/T). \tag{57}$$

On the other hand, by (12), $\tilde{L}_p - \tilde{L}_{\hat{r}} = \sum_{j=\hat{r}+1}^{p} \mu_j/n$. Since by assumption, $\Pr(r = \hat{r}) \to 1$, we must have

$$\tilde{L}_p - \tilde{L}_r = \sum_{j=r+1}^{p} \mu_j/n + o_{\mathrm{P}}(1/T). \tag{58}$$

Equation (16) follows from equations (57) and (58), and from the trivial observation that $L_p - \tilde{L}_p - \left(L_r - \tilde{L}_r\right) = 0$ when $p = r$.

For $p \leq r$, since $\Pr(r = \hat{r}) \to 1$, $\underline{L}_p + \sum_{j=1}^{p} \mu_j/n = o_{\mathrm{P}}(1)$. By Lemma 2, $\sum_{j=1}^{p} \mu_j/n = \sum_{j=1}^{p} d_j + o_{\mathrm{P}}(1)$, and hence, $\underline{L}_p - \underline{L}_r = \sum_{j=p+1}^{r} d_j + o_{\mathrm{P}}(1)$. On the other hand, by Proposition 1, $L_p - L_r = \sum_{j=p+1}^{r} d_j + o_{\mathrm{P}}(1)$. Therefore, (15) holds for $\tilde{L}_p = \underline{L}_p$. For $\tilde{L}_p = \bar{L}_p$, equality (15) can be proven similarly, using the fact that, under the strong factors asymptotics, $\hat{\rho}_j \xrightarrow{p} 1$.

Now, let us prove (ii). By Proposition 2 and by (5), for $p \leq q$, $L_p - L_0 =$

$\sum_{j=1}^{p}\left(1-2\rho_j\right)\mu_j/n + o_{\mathrm{P}}\left(1/n\right),$ where $\rho_j = \mathrm{d}\ln f\left(z\right)/\mathrm{dln}(z)$ evaluated at $z = \delta_{jn}/\sigma^2$. On the other hand, by Lemma 1 (iii), $\mathrm{d}\ln f\left(z\right)/\mathrm{dln}(z)$ at $z = \delta_{jn}/\sigma^2$ is smaller than one, and hence, $\sum_{j=1}^{p}\left(1-2\rho_j\right)\mu_j/n \geq \underline{L}_p$ for $p \leq \min\left\{\hat{r}, q\right\}$. Since by assumption, $\Pr\left(\hat{r}=q\right) \to 1$, we have, for any $\epsilon > 0$, $\Pr[\min_{0\leq p\leq q}\left(\left(L_p - L_0\right) - \underline{L}_p\right) \geq -\epsilon/n] \to 1$. To establish (17), it remains to note that, by Proposition 2, for $p > q$, $L_p - L_0 = L_q - L_0 + \sum_{j=q+1}^{p}\mu_j/n + o_{\mathrm{P}}\left(1/n\right)$ and, by (12), $\underline{L}_p = \underline{L}_q + \sum_{j=q+1}^{p}\mu_j/n + o_{\mathrm{P}}\left(1/n\right)$.

The convergence (18) can be proven similarly using the following fact. As shown in Lemma 4 below, $\operatorname{plim}\rho_j \geq \operatorname{plim}\hat{\rho}_j$ for $j \leq q$, where $\hat{\rho}_j = 1/(\mu_j^2 \max\{\hat{m}'\left(\mu_j\right), \widetilde{m}'\left(\mu_j\right)\})$ as in the definition (14) of $\bar{L}_p$.□

Lemma 4 *Suppose that assumptions A1w, A2w or A2w (nG), and A3w hold. Let* $\rho_j = \mathrm{d}\ln f\left(z\right)/\mathrm{dln}z$ *evaluated at* $z = \delta_{jn}/\sigma^2$, *and* $\hat{\rho}_j = 1/(\mu_j^2 \max\{\hat{m}'\left(\mu_j\right), \widetilde{m}'\left(\mu_j\right)\})$. *Then, for any* $j \leq q$,

$$\operatorname{plim}\rho_j \geq \operatorname{plim}\hat{\rho}_j. \tag{59}$$

Furthermore, let $\hat{\rho}_j^{(A=I)}$ *and* $\hat{\rho}_j^{(B=I)}$ *be as defined in (19-20). Then, if* $A = I_n$ *or* $B = I_T$, *for any* $j \leq q$, *we have, respectively,*

$$\operatorname{plim}\rho_j = \operatorname{plim}\hat{\rho}_j^{(A=I)}, \text{ and} \tag{60}$$

$$\operatorname{plim}\rho_j = \operatorname{plim}\hat{\rho}_j^{(B=I)}. \tag{61}$$

Proof: Let us denote $\mathrm{d}\ln f\left(z\right)/\mathrm{dln}z$ evaluated at $z = z_j = \delta_j/\sigma^2$ as $\mathrm{d}\ln f\left(z_j\right)/\mathrm{dln}z$. By Lemma 1, $\operatorname{plim}\rho_j = \mathrm{d}\ln f\left(z_j\right)/\mathrm{dln}z$. Further, using notation of the proof of Lemma 1, $\mathrm{d}\ln f\left(z_j\right)/\mathrm{dln}z = 1/\left(\mathrm{d}\ln g\left(x_j\right)/\mathrm{d}\ln x\right)$, where $x_j = f\left(z_j\right)$, and, for $x > \bar{x}$,

$$\begin{aligned}\frac{\mathrm{d}\ln g(x)}{\mathrm{d}\ln x} &= 1+\frac{1}{u_{2x}-1}\frac{\mathrm{d}\ln u_{2x}}{\mathrm{d}\ln x}+\frac{1}{v_{2x}-1}\frac{\mathrm{d}\ln v_{2x}}{\mathrm{d}\ln x} \\ &\leq 1+\max\left\{\frac{1}{u_{2x}-1},\frac{1}{v_{2x}-1}\right\}\left(\frac{\mathrm{d}\ln u_{2x}}{\mathrm{d}\ln x}+\frac{\mathrm{d}\ln v_{2x}}{\mathrm{d}\ln x}\right). \end{aligned} \quad (62)$$

The system of equations (7) in Onatski (2012) implies that

$$\begin{cases} -xm(x)-1=-u_{2x}\int(\lambda-u_{2x})^{-1}\mathrm{d}\mathcal{G}_A(\lambda)-1 \\ -xm(x)-1=c^{-1}\left(-v_{2x}\int(\lambda-v_{2x})^{-1}\mathrm{d}\mathcal{G}_B(\lambda)-1\right) \\ -xm(x)-1=x\left(cu_{2x}v_{2x}\right)^{-1}\end{cases}, \quad (63)$$

where $m(x)=\int(\lambda-x)^{-1}\mathrm{d}\mathcal{G}(\lambda)$ and $\mathcal{G}(\lambda)$ is the cumulative distribution function of the limit of the empirical distribution of the eigenvalues of $ee'/(\sigma^2T)$ as $n,T\to_c\infty$. Therefore,

$$\frac{\mathrm{d}\ln u_{2x}}{\mathrm{d}\ln x}+\frac{\mathrm{d}\ln v_{2x}}{\mathrm{d}\ln x}=1-\frac{\mathrm{d}\ln(-xm(x)-1)}{\mathrm{d}\ln x}. \quad (64)$$

Moreover, using Jensen's inequality and the normalizations $\int\mathrm{d}\mathcal{G}_A(\lambda)=\int\mathrm{d}\mathcal{G}_B(\lambda)=1$ in the first two equations of system (63), we obtain

$$\frac{1}{u_{2x}-1}\leq -xm(x)-1 \text{ and } \frac{1}{v_{2x}-1}\leq c\left(-xm(x)-1\right). \quad (65)$$

From (62), (64), and (65), we get

$$\mathrm{d}\ln g(x)/\mathrm{d}\ln x\leq 1+\max\{c,1\}\left(-1+x^2m'(x)\right). \quad (66)$$

Let $\underline{m}(x)=cm(x)-(1-c)/x$. That is, $\underline{m}(x)=\int(\lambda-x)^{-1}\mathrm{d}\underline{\mathcal{G}}(\lambda)$, where $\underline{\mathcal{G}}(\lambda)$ is the cdf of the limit of the empirical distribution of the eigenvalues of $e'e/(\sigma^2T)$ (as opposed to $ee'/(\sigma^2T)$) as $n,T\to_c\infty$. We have $\underline{m}'(x)=cm'(x)+(1-c)/x^2$ so that

$x^2 m'(x) = c^{-1} x^2 \underline{m}'(x) - c^{-1} + 1$. Using this equality in (66) when $c \geq 1$, we obtain $\mathrm{d}\ln g(x)/\mathrm{d}\ln x \leq x^2 \underline{m}'(x)$. When $c < 1$, (66) simplifies to $\mathrm{d}\ln g(x)/\mathrm{d}\ln x \leq x^2 m'(x)$. Therefore, we have $\mathrm{d}\ln g(x)/\mathrm{d}\ln x \leq x^2 \max\{m'(x), \underline{m}'(x)\}$ for $x > \bar{x}$, and thus

$$\mathrm{d}\ln f(z_j)/\mathrm{d}\ln z \geq 1/\left(x_j^2 \max\{m'(x_j), \underline{m}'(x_j)\}\right). \tag{67}$$

From (40) and (41) and the definitions of q and $\hat{r}$, we see that $\mathcal{G}(\lambda)$ and $\underline{\mathcal{G}}(\lambda)$ are the cdf's of the limits of the empirical distributions of $\mu_{\hat{r}+1}/\sigma^2, ..., \mu_n/\sigma^2$ and of $\mu_{\hat{r}+1}/\sigma^2, ..., \mu_T/\sigma^2$, respectively. Hence,

$$x_j^2 m'(x_j) - \frac{1}{n-\hat{r}} \sum_{i=\hat{r}+1}^{n} \left(\sigma^2 x_j\right)^2 / \left(\sigma^2 x_j - \mu_i\right)^2 \xrightarrow{p} 0, \text{ and}$$
$$x_j^2 \underline{m}'(x_j) - \frac{1}{T-\hat{r}} \sum_{i=\hat{r}+1}^{T} \left(\sigma^2 x_j\right)^2 / \left(\sigma^2 x_j - \mu_i\right)^2 \xrightarrow{p} 0.$$

Since $\sigma^2 x_j = \operatorname{plim} \sigma^2 f(z_{jn}) = \operatorname{plim} \mu_j$, the latter two convergences imply that

$$x_j^2 m'(x_j) - \mu_j^2 \hat{m}'\left(\mu_j\right) \xrightarrow{p} 0 \text{ and } x_j^2 \underline{m}'(x_j) - \mu_j^2 \widetilde{m}'\left(\mu_j\right) \xrightarrow{p} 0. \tag{68}$$

Finally, (67) and (68) imply (59).

Now, assume that $A = I_n$. Then the first equation of (63) can be written as $-xm(x) - 1 = 1/(u_{2x} - 1)$. Therefore,

$$\frac{1}{u_{2x}-1} \mathrm{d}\ln u_{2x}/\mathrm{d}\ln x = -1 - xm'(x)/m(x). \tag{69}$$

Furthermore, from the third equation of (63), we have $1/(u_{2x} - 1) = x\left(cu_{2x}v_{2x}\right)^{-1}$ so that $v_{2x} = x(u_{2x} - 1)/(cu_{2x})$. Therefore, after some algebra, we get

$$\frac{1}{v_{2x}-1}\mathrm{d}\ln v_{2x}/\mathrm{d}\ln x = cxm'(x)/(1+cm(x)). \tag{70}$$

Combining (69) and (70), we obtain $\mathrm{d}\ln g(x)/\mathrm{d}\ln x = -xm'(x)/(m(x)(1+cm(x)))$, and therefore,

$$\operatorname{plim} \mathrm{d}\ln f(z_{jn})/\mathrm{d}\ln(z) = -\frac{m(x_j)(1+cm(x_j))}{xm'(x_j)}. \tag{71}$$

On the other hand, similarly to (68),

$$\frac{m(x_j)}{xm'(x_j)} - \frac{\hat{m}(\mu_j)}{\mu_j \hat{m}'(\mu_j)} \xrightarrow{p} 0 \text{ and } cm(x_j) - (n/T)\hat{\sigma}^2\hat{m}(\mu_j) \xrightarrow{p} 0,$$

where $\hat{\sigma}^2 = (n-\hat{r})^{-1}\sum_{i=\hat{r}+1}^{n}\mu_i$. Thus,

$$\operatorname{plim} \hat{\rho}_j^{(A=I)} = -\frac{m(x_j)(1+cm(x_j))}{xm'(x_j)}, \tag{72}$$

and (71) and (72) imply (60).

The equality (61) can be proven similarly to (60) after $-xm(x)-1$ in (63) is replaced by $(-x\underline{m}(x)-1)/c$. We omit details of such a similar proof to save space.□

7.9 Proof of Proposition 4

The proof of part (i) is similar to the proof of Proposition 3 (i), and we therefore omit it. For part (ii), since $\Pr(\hat{r}=q) \to 1$, Proposition 2 and equalities (60) and (61) of Lemma 4 imply that

$$
\begin{aligned}
\max_{0\le p\le r_{\max}} \left| L_p^{(1)} - \hat{L}_p^{(A=I)} - (L_q^{(1)} - \hat{L}_q^{(A=I)}) \right| &= o_{\mathrm{P}}(1/T), \text{ or} \\
\max_{0\le p\le r_{\max}} \left| L_p^{(1)} - \hat{L}_p^{(B=I)} - (L_q^{(1)} - \hat{L}_q^{(B=I)}) \right| &= o_{\mathrm{P}}(1/T),
\end{aligned}
$$

if $A = I_n$, or $B = I_T$, respectively. Part (ii) now follows from Proposition 2.□

7.10 Proof of Proposition 5.

First, let us prove (i). Let $p_0^* = \arg\min_{0\le p\le r_{\max}} L_p^{(1)}$, where $L_p^{(1)}$ is as defined in Proposition 1. From (3), using assumptions A1-A3, we have $\Pr(p_0^* = r) \to 1$ as $n, T \to_c \infty$. Since $L_p - L_p^{(1)} = o_{\mathrm{P}}(1/T)$ and $\min_{p<r} L_p^{(1)}$ is positive and bounded away from zero with probability approaching one (w.p.a.1), we have $\Pr(p^* \ge r) \to 1$, where $p^* = \arg\min_{0\le p\le r_{\max}} L_p$. To establish the *optimal loss consistency* of any estimator that is consistent for r, it remains to show that $\Pr(p^* > r) \to 0$.

In view of (3) and equality $L_p - L_p^{(1)} = o_{\mathrm{P}}(1/T)$, it is sufficient to prove that μ_{r+1} is positive and bounded away from zero w.p.a.1. By (41), $\mu_{r+1} > \mu_{2r+1}(e'e/T)$. On the other hand, there must exist $\epsilon > 0$ such that $\Pr\left(\mu_{2r+1}(e'e/T) \ge \epsilon\right) \to 1$. This follows from Assumption A3 (i) and the fact that $\mathrm{tr}\,(e'e)/(nT) \le 2r\mu_1(e'e)/(nT) + n\mu_{2r+1}(e'e)/(nT) = \mu_{2r+1}(e'e/T) + o_{\mathrm{P}}(1)$, where the last equality holds by A3 (iii). Therefore, under the strong factors asymptotics, any estimator that is consistent for r is *optimal loss consistent.*

Now, let $\hat{p}$ be one of the following estimators: $\underline{p}$, $\bar{p}$, $\hat{p}^{(A=I)}$, or $\hat{p}^{(B=I)}$. From parts (i) of Propositions 3 and 4, we have $\Pr(\hat{p} \ge r) \to 1$ because $\min_{0\le p<r} L_p$ is positive and stays away from zero w.p.a.1, whereas $\max_{r\le p\le r_{\max}} L_p \xrightarrow{p} 0$. On the other hand, $\Pr(\hat{p} > r) \to 0$, which is established similarly to $\Pr(p^* > r) \to 0$.

Hence, $\hat{p}$ is consistent for r, and thus, *optimal loss consistent* under the strong factors asymptotics.

Turning to the proof of part (ii), let $p_0^* = \arg\min_{0 \le p \le r_{\max}} L_p^{(1)}$, where now $L_p^{(1)}$ is as defined in Proposition 2. Let us show that $\Pr(p_0^* = p^*) \to 1$. By Proposition 2, $L_p^{(1)} - L_p = o_{\mathrm{P}}(1/T) = o_{\mathrm{P}}(1/n)$. Therefore, it is sufficient to show that there exists $\epsilon > 0$ such that

$$\Pr\left(\min_{0 \le p \le r_{\max}, p \ne p_0^*} L_p^{(1)} - L_{p_0^*}^{(1)} > \epsilon/n\right) \to 1. \tag{73}$$

If $p_0^* < q$, this follows from (6) and Lemma 1. If $p_0^* = q$, this follows from (6), Lemma 1, and the fact that μ_{q+1} is bounded away from zero w.p.a.1, which we will now establish. Note that p_0^* cannot be larger than q by (6).

By (41), $\mu_{q+1} > \mu_{2q+1}(e'e/T)$. Onatski (2010) proves that under assumptions A1w, A2w or A2w (nG), A3w, the empirical distribution of the eigenvalues of $e'e/T$ converges to a fixed distribution, and that $\mu_1(ee'/T)$ converges to the finite upper boundary of the support of this limiting distribution. This implies that, for any fixed k, $\mu_k(e'e/T)$ converges to the same finite value. Taking $k = 2q+1$, we see that $\mu_{2q+1}(e'e/T)$ is bounded away from zero w.p.a.1.

To summarize, we have just established the fact that $\Pr(p_0^* = p^*) \to 1$ under the weak factors asymptotics. Now note that by an appropriate choice of A, B, and of $\delta_1, ..., \delta_r$, we can make p_0^* converge to any integer between 0 and q. Indeed, the minimum of $L_p^{(1)}$ is asymptotically achieved at the largest $p \le q$ such that inequality (7) holds. For the special case where $A = I_n$ and $B = I_T$, $f(z)$ is given by (8), and the value of such a largest p can be set to an arbitrary integer between zero and q by an appropriate choice of $\delta_1, ..., \delta_r$. Hence $\hat{r}$, which is consistent for q under the weak factors asymptotics, is not, in general, *optimal loss consistent.*

Next, by the definitions of $L_p^{(1)}$, $\bar{L}_p$, and $\underline{L}_p$, and by Lemma 4, we have for any p such that $1 \leq p \leq \min\{\hat{r}, q\}$,

$$\bar{L}_{p-1} - \bar{L}_p \leq L_{p-1}^{(1)} - L_p^{(1)} + o_{\mathrm{P}}(1/T) \text{ and} \tag{74}$$

$$\underline{L}_{p-1} - \underline{L}_p \geq L_{p-1}^{(1)} - L_p^{(1)} + o_{\mathrm{P}}(1/T). \tag{75}$$

Furthermore, for any $p \geq \max\{\hat{r}, q\}$,

$$\bar{L}_{p+1} - \bar{L}_p = L_{p+1}^{(1)} - L_p^{(1)} \text{ and} \tag{76}$$

$$\underline{L}_{p+1} - \underline{L}_p = L_{p+1}^{(1)} - L_p^{(1)}. \tag{77}$$

Since $\hat{r}$ is consistent for q, inequalities (74)-(75), equalities (76)-(77), and the convergence (73) imply that $\Pr\left(\underline{p} \leq p_0^* \leq \bar{p}\right) \to 1$, that $L_{\bar{p}}^{(1)} - L_{p_0^*}^{(1)} \leq \bar{L}_{\bar{p}} - \underline{L}_{\underline{p}} + o_{\mathrm{P}}(1/T)$, and that $L_{\underline{p}}^{(1)} - L_{p_0^*}^{(1)} \leq \bar{L}_{\underline{p}} - \bar{L}_{\bar{p}} + o_{\mathrm{P}}(1/T)$. Part (ii) of Proposition 5 now follows from the facts that $\Pr(p_0^* = p^*) \to 1$ and $L_p^{(1)} - L_p = o_{\mathrm{P}}(1/T)$. The latter two facts, together with (73) and Proposition 4 (ii) also imply part (iii).□

Acknowledgement *I would like to thank Serena Ng, Yuichi Kitamura, Giovanni Urga, and three anonymous referees for helpful suggestions. I am grateful to J.M. Keynes Fellowships Fund, University of Cambridge, for financial support.*

References

[1] Ahn, S.C., and A.R. Horenstein, 2013, Eigenvalue Ratio Test for the Number of Factors. Econometrica 81, 1203 – 1227.

[2] Akaike, H., 1973, Information Theory and an Extension of the Maximum Likelihood Principle, in: Petrov, B. and F. Csáki (Eds.), Second International Symposium on Information Theory, 267-281. Akadémiai Kiadó, Budapest.

[3] Alessi, L., M. Barigozzi, and M. Capasso, 2010, Improved Penalization for Determining the Number of Factors in Approximate Factor Models. Statistics and Probability Letters 80, 1806 – 1813.

[4] Bai, Z.D., 1999, Methodologies in Spectral Analysis of Large Dimensional Random Matrices, a review. Statistica Sinica 9, 611 – 677.

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

[6] Bai, J. and S. Ng, 2006, Determining the Number of Factors in Approximate Factor Models, Errata. Mimeo, Columbia University.

[7] Bai, J. and S. Ng, 2008, Large Dimensional Factor Analysis. Foundations and Trends in Econometrics 3, 89 – 163.

[8] Bai, J. and S. Ng, 2010, Selecting Instrumental Variables in a Data Rich Environment. Econometric Theory 26, 1577 – 1606.

[9] Belloni, A., Chen, D., Chernozhukov, V., and C. Hansen, 2012, Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. Econometrica 80, 2369 – 2429.

[10] Boivin, J. and S. Ng, 2006, Are more data always better for factor analysis. Journal of Econometrics 132, 169 – 194.

[11] Caner, M. and X. Han, 2014, Selecting the Correct Number of Factors in Approximate Factor Models: The Large Panel Case with Group Bridge Estimator. Journal of Business and Economics Statistics 32, 359 – 374.

[12] Chudik, A., M. H. Pesaran, and E. Tosetti, 2011, Weak and strong cross section dependence and estimation of large panels. Econometrics Journal 14, C45 – C90.

[13] Connor, G., 1995, The three types of factor models: A comparison of their explanatory power. Financial Analysts Journal 51, 42 – 46.

[14] Connor, G. and R. Korajczyk, 1993, A test for the number of factors in an approximate factor model. Journal of Finance 58, 1263 – 1291.

[15] deLeeuw, J., 1992, Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle, in: Kotz, S., and N.L. Johnson, (Eds.), Breakthroughs in Statistics Volume 1. Foundations and Basic Theory. Springer Series in Statistics, Perspectives in Statistics. Springer-Verlag: New York, 599 – 609.

[16] DeMol, C., Giannone, D. and L. Reichlin, 2008, Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146, 318 – 328.

[17] Forni, M., M. Hallin, M. Lippi, and L. Reichlin, 2000, The Generalized Dynamic-Factor Model: Identification and Estimation. Review of Economics and Statistics 82, 540 – 554.

[18] Forni, M., and L. Reichlin, 2001, Federal policies and local economies: Europe and the US. European Economic Review 45, 109 – 134.

[19] Foerster, A., P-D Sarte, and M. Watson, 2011, Sectoral vs. Aggregate Shocks: A Structural Factor Analysis of Industrial Production. Journal of Political Economy 119, 1 – 38.

[20] Hansen, B.E., 1996, Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. Econometrica 64, 413 – 430.

[21] Hansen, B.E., 2007, Least squares model averaging. Econometrica 75, 1175 – 1189.

[22] Heaton, C., and V. Solo, 2006, Estimation of Approximate Factor Models: Is it Important to Have a Large Number of Variables? Research Paper 0605, Macquarie University.

[23] Horn, R.A., and C.R. Johnson, 1985, Matrix Analysis, Cambridge University Press.

[24] Horn, R.A., and C.R. Johnson, 1991, Topics in Matrix Analysis, Cambridge University Press.

[25] Huang, R. and H. Jo, 1995, Data frequency and the number of factors in stock returns. Journal of Banking and Finance 19, 987 – 1003.

[26] Kapetanios, G. and M. Marcellino, 2010, Factor-GMM estimation with large sets of possibly weak instruments. Computational Statistics and Data Analysis 54, 2655 – 2675.

[27] Kato, T., 1980, Perturbation Theory for Linear Operators. Second Edition. Springer-Verlag. Berlin, Heidelberg, New York.

[28] Krantz, S.G., 1992, Function Theory of Several Complex Variables, American mathematical Society, Providence, Rhode Island.

[29] Kunitomo, N., and T. Yamamoto, 1985, Properties of Predictors in Misspecified Autoregressive Time Series Models. Journal of the American Statistical Association 80, 941 – 950.

[30] Lancaster, T., 2000, The Incidental Parameter Problem Since 1948. Journal of Econometrics 95, 391 – 413.

[31] Li, Ker-Chau, 1987, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Annals of Statistics 15, 958 – 975.

[32] Mallows, C.L., 1973, Some comments on C_p. Technometrics 15, 661 – 675.

[33] Moon, H.R., and M. Weidner, 2010, Dynamic Linear Panel Regression Models with Interactive Fixed Effects, manuscript, UCL.

[34] Moon, H.R., and M. Weidner, 2010a, A Supplementary Appendix to "Dynamic Linear Panel Regression Models with Interactive Fixed Effects", manuscript, UCL.

[35] Onatski, A., 2006, The Principal Components Estimation of Large Factor Models when Factors are Weak, mimeo, University of Cambridge.

[36] Onatski, A., 2010, Determining the Number of Factors From Empirical Distribution of Eigenvalues. Review of Economics and Statistics 92, 1004 – 1016.

[37] Onatski, A., 2012, Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors. Journal of Econometrics 168, 244 – 258.

[38] Onatski, A., Moreira, M. J., and M. Hallin, 2013, Asymptotic Power of Sphericity Tests for High-dimensional Data. Annals of Statistics 41, 1204 – 1231.

[39] Phillips, P.C.B., 1979, The sampling distribution of forecasts from a first-order autoregression. Journal of Econometrics 9, 241 – 261.

[40] Shao, J., 1997, An asymptotic theory for linear model selection. Statistica Sinica 7, 221 – 264.

[41] Shibata, R., 1980, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Annals of Statistics 8, 147 – 164.

[42] Stock, J. and M. Watson, 2006, Forecasting with Many Predictors, in: G. Elliot, C.W.J. Granger, and A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol 1, 515 – 554.

[43] Stock, J. and M. Watson, 2011, Dynamic Factor Models, in: M. P. Clements and D. F. Hendry (Eds.), Oxford Handbook of Forecasting, Oxford: Oxford University Press.

[44] Van Loan, C. F. and N.P. Pitsianis, 1993, Approximation with Kronecker products, in: M. S. Moonen and G. H. Golub (Eds.), Linear Algebra for Large Scale and Real Time Applications, Kluwer Publications, 293 – 314.

[45] Yin, Y.Q., Z.D. Bai, and P.R. Krishnaiah, 1988, On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Theory of Probability and Related Fields 78, 509 – 521.