Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models

Alexei Onatski*

July 7, 2013

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).

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

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

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 an estimator of the loss which is consistent under both strong and weak factors asymptotics, and propose to use it for model comparison and for selection of the number of factors. We show that the estimator of the number of factors that minimizes the loss estimate is 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

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

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.

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. Under the strong factors asymptotics, all the proposed estimators are consistent for L_p after a shift by a constant that does not depend on p. Moreover, the minimizers of these estimators are consistent for the true number of factors.

Under the weak factors asymptotics, two of the proposed 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 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.

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 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 $d_{1n} \geq d_{2n} \geq ... \geq d_{rn} > 0$ 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 $\Pr(\operatorname{tr}[ee']/(nT) > \varepsilon) \to 1$; (ii) for any $j, k \leq r$, $\Lambda'_{\cdot j} e F_{\cdot k}/\sqrt{nT} = O_p(1)$; (iii) $\mu_1(ee'/T) = O_p(1)$.

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 $j, k \leq r$ 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). In the Supplementary Appendix, we show 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 $L_{\cdot i}, ..., L_{\cdot j}$. 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, $E\left(e'e\right) = n\sigma^2 I_T$, 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* $n, T \to_c \infty$, $L_p = L_p^{(1)} + o_{\mathrm{P}}\left(1/T\right)$, *where*

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

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 $\sigma^2\left(1/T + 1/n\right)$ 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 $\mu_1\left(ee'/T\right) \to \sigma^2\left(1 + \sqrt{c}\right)^2$ as $n, T \to_c \infty$ (Yin et al, 1988). Replacing c in $\sigma^2\left(1 + \sqrt{c}\right)^2$ 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 section 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 $\delta_{1n} \geq \delta_{2n} \geq \ldots \geq \delta_{rn} > 0$ 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} can be thought of measuring 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 $\mathrm{tr}(A_n A_n') = n$ and $(A_n A_n') \Lambda = \Lambda$; (iii) B_T is such that $\mathrm{tr}(B_T' B_T) = T$ 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 (2012). 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 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 $\mathbf{1}\left[\cdot\right]$ 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 $n, T \rightarrow_c \infty$: (i) $G_A(x) \rightarrow \mathcal{G}_A(x)$ and $G_B(x) \rightarrow \mathcal{G}_B(x)$ for all $x \in \mathbb{R}$, (ii) $\mu_1\left(A_n A_n'\right) \rightarrow \bar{x}_A$ and $\mu_1\left(B_T' B_T\right) \rightarrow \bar{x}_B$, 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-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 $f(z)/z \rightarrow 1$ as $z \rightarrow \infty$; (iii) the elasticity $\mathrm{d}\ln f(z)/\mathrm{d}\ln z$ increases on $z > \bar{\delta}/\sigma^2$ and converges to one as $z \rightarrow \infty$.*

Proposition 2 *Suppose that assumptions A1w-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* $n, T \to_c \infty$, $L_p = L_p^{(1)} + o_{\mathrm{P}}(1/T)$, *where*

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

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 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, when $\mathcal{G}_A$ and $\mathcal{G}_B$ are degenerate probability distributions putting all mass at one (see Onatski (2006, Theorem 5)),[1]

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

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.* $N\left(0,\sigma^2\right)$. *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*

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

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. Instead, it is achieved at the maximum of $p \in \{0, 1, 2, ..., q\}$ such that $\delta_{pn}/n > \sigma^2\left(1/n+1/T\right) + 3\sigma^4/\left(T\delta_{pn}\right)$.

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

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

[1] Note that the constraints that A3w imposes on the densities of $\mathcal{G}_A$ and $\mathcal{G}_B$ in $\left(\underline{x}_A, \bar{x}_A\right)$ and $\left(\underline{x}_B, \bar{x}_B\right)$ are trivially satisfied even though $\mathcal{G}_A$ and $\mathcal{G}_B$ are not absolutely continuous with respect to Lebesgue measure because the intervals $\left(\underline{x}_A, \bar{x}_A\right)$ and $\left(\underline{x}_B, \bar{x}_B\right)$ are empty.

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

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

Hence, L_p equals $2KL\left(F, \Lambda; \hat{F}_{1:p}, \hat{\Lambda}_{1:p}\right)/(nT)$, which is the exact analog for the factor models of the loss used by Akaike (1973) to derive AIC.

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 2 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 $\mu_j\left(XX'/T\right)$. 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, $2\sum_{j=1}^{p}\mu_j \overset{a.s.}{\to} 2p\left(1+\sqrt{c}\right)^2$. 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 $p\left(n+T\right)\neq p(\sqrt{n}+\sqrt{T})^2$.

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 overspecified model are not identified (when the true loadings are identically zero, the corresponding factor may correspond to any point on a 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 $-2\ln L\left(X|\hat{\Lambda}_{1:p},\hat{F}_{1:p}\right)/\left(nT\right)$. Even though under the general assumptions A1-A3 or A1w-A3w, 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.

Let $\hat{r}$ be an estimator of the number of factors, such that $\operatorname{plim}\hat{r}=r$ under the strong factors asymptotics and $\operatorname{plim}\hat{r}=q$ under the weak factors asymptotics, where q is as defined in Proposition 2. One such $\hat{r}$ is the estimator of the number of factors developed in Onatski (2010). Below we study estimators of the loss L_p, shifted by a quantity that does not depend on p, of the following form

$$\hat{L}_p = \begin{cases} \sum_{j=1}^{p} \left(1 - 2\hat{\rho}_j\right) \mu_j / n & \text{for } p \leq \hat{r} \\ \hat{L}_{\hat{r}} + \sum_{j=\hat{r}+1}^{p} \mu_j / n & \text{for } p > \hat{r} \end{cases}, \tag{11}$$

where $\hat{\rho}_j$ are data dependent.

Consider the following two special cases of $\hat{L}_p$:

$$\begin{aligned} \underline{L}_p &= \hat{L}_p \text{ with } \hat{\rho}_j = 1, \text{ and} & (12) \\ \bar{L}_p &= \hat{L}_p \text{ with } \hat{\rho}_j = 1/(\mu_j^2 \max\{\hat{m}'\left(\mu_j\right), \widetilde{m}'\left(\mu_j\right)\}), & (13) \end{aligned}$$

where

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

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, as $n, T \to_c \infty$,*

$$\begin{aligned} \max_{0 \leq p < r} \left| L_p - \tilde{L}_p - (L_r - \tilde{L}_r) \right| &= o_{\mathrm{P}}(1) \text{ and} & (14) \\ \max_{r \leq p \leq r_{\max}} \left| L_p - \tilde{L}_p - (L_r - \tilde{L}_r) \right| &= o_{\mathrm{P}}(1/T). & (15) \end{aligned}$$

(ii) Under assumptions A1w-A3w, for any $\epsilon > 0$, as $n, T \to_c \infty$,

$$\begin{aligned} \Pr[\min_{0 \leq p \leq r_{\max}} \left((L_p - L_0) - \underline{L}_p\right) &\geq -\epsilon/n] \to 1 \text{ and} & (16) \\ \Pr[\max_{0 \leq p \leq r_{\max}} \left((L_p - L_0) - \bar{L}_p\right) &\leq \epsilon/n] \to 1. & (17) \end{aligned}$$

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 (14) 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 (15) 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 (13) converge in probability to one, which coincides with the value of $\hat{\rho}_j$ in (12).

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 $\operatorname{plim} \hat{\rho}_j = \mathrm{d} \ln f \left(\delta_j/\sigma^2\right) / \mathrm{d}\ln(\delta_j/\sigma^2)$. 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

$$\begin{aligned} \hat{\rho}_j^{(A=I)} &= -(1 + \hat{m}(\mu_j)\hat{\sigma}^2 n/T)\hat{m}(\mu_j)/(\mu_j \hat{m}'(\mu_j)), \text{ and} \qquad (18) \\ \hat{\rho}_j^{(B=I)} &= -(1 + \widetilde{m}(\mu_j)\widetilde{\sigma}^2 T/n)\widetilde{m}(\mu_j)/(\mu_j \widetilde{m}'(\mu_j)). \qquad (19) \end{aligned}$$

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, as $n, T \to_c \infty$,*

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

(ii) Suppose that assumptions A1w-A3w hold. If $A = I_n$ or $B = I_T$, then, as $n, T \to_c \infty$, respectively,

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

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 $\hat{L}_p^{(B=I)}$ approximate 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. This estimator is called *asymptotically loss efficient* if

$$\frac{L_{\hat{p}}}{\min_{0 \le p \le r_{\max}} L_p} \xrightarrow{p} 1. \tag{20}$$

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

$$\Pr(\hat{p} = p^*) \to 1, \tag{21}$$

where $p^* = \arg\min_{0 \le p \le r_{\max}} L_p$. 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 (21) *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, estimators $\hat{r}, \underline{p}$, $\bar{p}$, $\hat{p}^{(A=I)}$, and $\hat{p}^{(B=I)}$ are consistent for r, and thus, optimal loss consistent.
(ii) Suppose that assumptions A1w-A3w hold. Then, estimator $\hat{r}$ 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, $L_{\bar{p}} - L_{p^*} \leq \bar{L}_{\bar{p}} - \underline{L}_{\underline{p}} + o_{\mathrm{P}}\left(1/T\right)$ and $L_{\underline{p}} - L_{p^*} \leq \bar{L}_{\underline{p}} - \bar{L}_{\bar{p}} + o_{\mathrm{P}}\left(1/T\right)$.
(iii) In the special cases where assumptions A1w-A3w hold and $A = I_n$ or $B = I_T$, estimators $\hat{p}^{(A=I)}$ or $\hat{p}^{(B=I)}$, respectively, are optimal loss consistent.

By definition, the minimum of $\hat{L}_p$ cannot be achieved for $p > \hat{r}$. Under the weak factors asymptotics, $\hat{r} \xrightarrow{p} q \leq r$. Therefore, 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 $\bar{p}$. Moreover, under the strong factors asymptotics, both $\underline{p}$ and $\bar{p}$ are *optimal loss consistent* and consistent for r.

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 (2012). The data are generated from

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

where the common component C_{it} and the idiosyncratic component e_{it} are independent and 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 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. $N(0,\sigma_v^2)$ with $\sigma_v^2 = (1-\rho^2)/(1+2J\beta^2)$, and $J = \min(n/20, 10)$. Following Ahn and Horenstein (2012), 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 $n = T = 50$, 100, or 200, and $r_{\max}$ 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 $n = T = 100$ and $(\rho,\beta) = (0.7,0)$. 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 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,

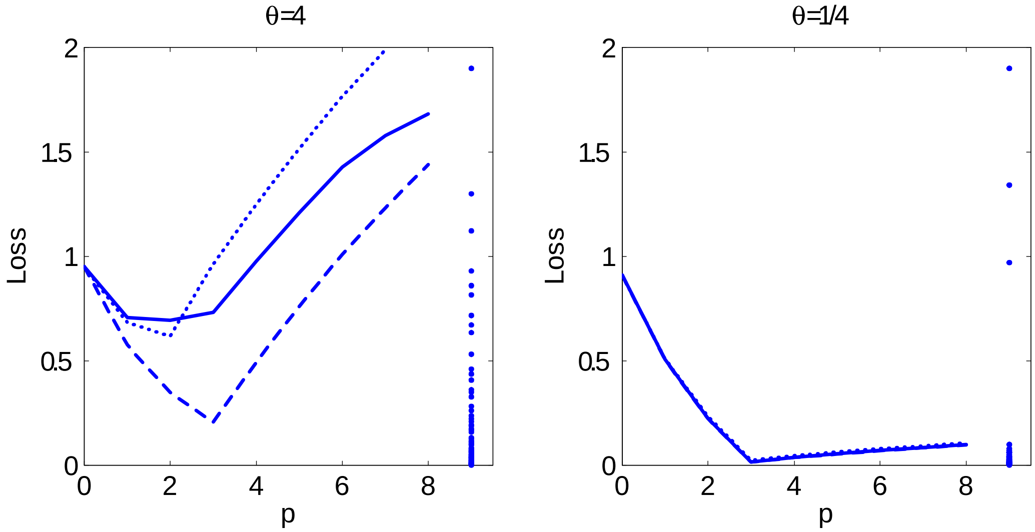

Figure 1: Particular realizations of the loss L_p (solid lines) and the corresponding asymptotic approximations $L_p^{(1)}$ (dotted lines – weak factors approximation, dashed lines – strong factors approximation). $n = T = 100$, $(\rho, \beta) = (0.7, 0)$.

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 $n = T = 100$. The other cases provide qualitatively similar information and we do not report them here. 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, espe-

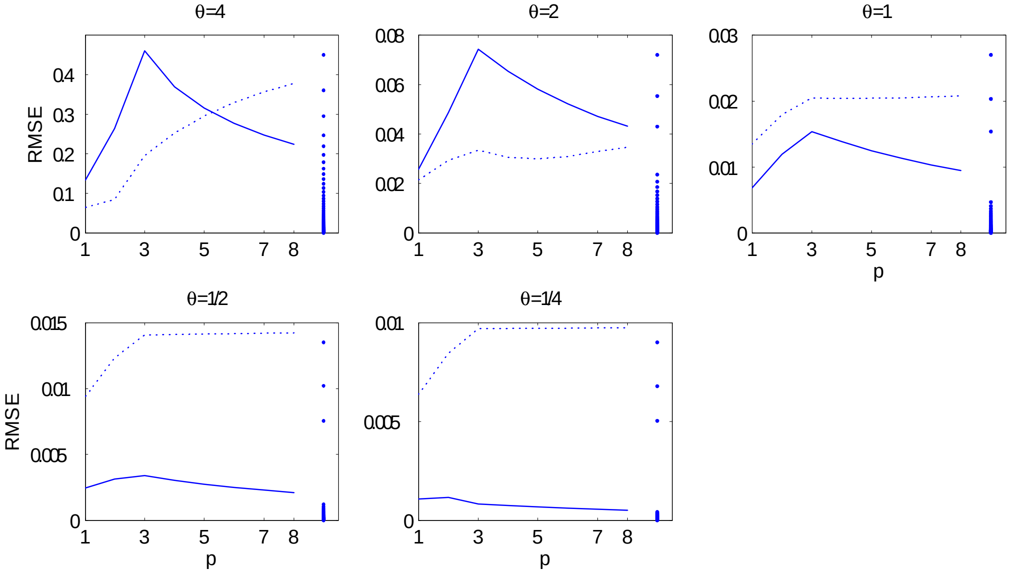

Figure 2: The root mean squared errors (over 1000 MC replications) of the strong (dashed lines) and weak (dotted lines) factors asymptotic approximations of L_p. The dots with abscissa p=9 show the MC average values of the sorted eigenvalues of the sample covariance matrix. $(\rho, \beta) = (0.7, 0)$, $n = T = 100$.

cially for $p \leq r$.

We now turn to the analysis of the proposed loss estimators. Figure 3 shows RMSE of $\bar{L}_p$ (solid lines), $\underline{L}_p$ (dotted lines), and $\hat{L}_p^{(A=I)}$ (dashed lines) for $n = T = 100$ and $(\rho, \beta) = (0.7, 0)$. Since $n = T$ in our experiments, $\hat{L}_p^{(B=I)}$ coincides with $\hat{L}_p^{(A=I)}$. In all our experiments, estimator $\hat{L}_p^{(A=I)}$ performs very similar to the simpler estimator $\bar{L}_p$. Therefore, in what follows, we focus on the analysis of $\bar{L}_p$ and $\underline{L}_p$. Since, as explained above, $\bar{L}_p$ and $\underline{L}_p$ estimate L_p only up to a shift, we shift the estimation error to zero at $p = q$ when computing RMSE. For relatively strong factors ($\theta \leq 2$), q was equal to $r = 3$ in most of our experiments. On Figure 3, this is reflected in the fact that RMSE is zero at $p = r = 3$.

From Figure 3, we see that estimator $\bar{L}_p$ dominates $\underline{L}_p$ for all values of θ. This situation remains the same for the other combinations of ρ and β, and the other sample sizes (not shown here). Therefore, in the rest of this section, we will restrict

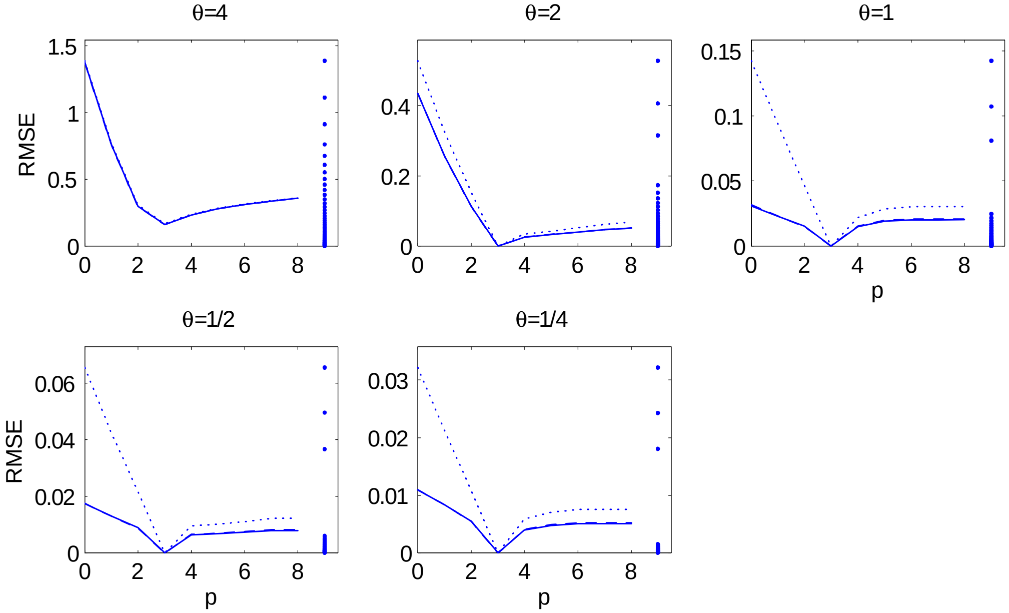

Figure 3: The root mean squared errors (over 1000 MC replications) of the shifted versions of $\bar{L}_p$.(solid lines), $\underline{L}_p$ (dotted lines), and $\hat{L}_p^{(A=I)}$ (dashed lines). $(\rho, \beta) = (0.7, 0)$, $n = T = 100$.

attention to $\bar{L}_p$. Comparing Figures 3 and 2, we see that the quality of $\bar{L}_p$ is comparable to the quality of the theoretical asymptotic approximations to the loss, except for relatively weak factors ($\theta \geq 2$) and $p < r$, where the quality of $\bar{L}_p$ is substantially worse.

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 4 illustrates this finding. It shows the same realization of L_p as on the left panel of Figure 1 (solid line) superimposed with a shifted version of $\bar{L}_p$. For the corresponding data replication, $q = 2$, so the graphs coincide at $p = q = 2$, by construction. However, our preliminary estimator $\hat{r} = 1 < q$. The dotted line shows what the value of $\bar{L}_p$ 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 for $p < r$.

Accurate estimation of q is difficult when factors are weak. The difficulty is well

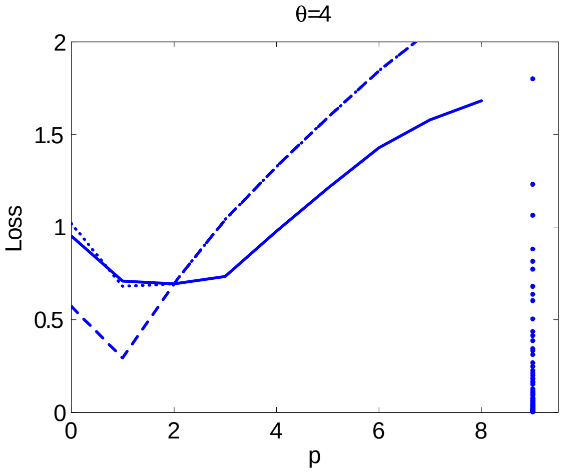

Figure 4: A realization of L_p (solid line) and $\bar{L}_p$ (dashed line). Dotted line corresponds to $\bar{L}_p$ based on the counterfactual $\hat{r} = q$. $(\rho, \beta) = (0.7, 0)$, $n = T = 100$.

illustrated by the relative position of the eigenvalues shown on Figure 4. 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 4 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) estimator has been formally shown to be consistent for q (see Section 3 of Onatski (2012)). 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 $\hat{r}$. Our benchmark is Onatski's (2010) ED estimator, used above. We also consider Bai and Ng (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 (2012), and Alessi et al's (2010) ABC estimator.

Comparing quality of different loss estimates requires some care. Ideally, the measure of quality should depend on the use to which the estimate will be put.

n, T	θ	ED	ER	GR	ABC	BIC	PC1	PC2	IC1	IC2
50	4	32.7	32.5	32.8	32.7	29.3	50.6	46.6	51.8	34.4
	2	32.0	21.6	20.2	30.0	11.9	23.2	21.0	23.7	17.9
	1	16.5	10.2	8.4	14.7	5.2	11.8	10.7	12.0	9.6
	1/2	2.5	3.6	2.6	3.3	2.8	6.1	5.6	6.2	5.1
	1/4	1.0	1.1	0.9	1.3	1.5	3.1	2.9	3.2	2.7
100	4	32.1	37.0	36.7	25.9	24.2	22.7	17.3	22.3	12.3
	2	9.6	9.5	6.8	4.9	4.6	11.8	9.0	11.9	4.4
	1	1.2	1.1	1.1	1.7	1.1	6.1	4.7	6.1	2.4
	1/2	0.7	0.6	0.6	1.0	0.6	3.1	2.4	3.1	1.3
	1/4	0.4	0.4	0.4	0.5	0.4	1.6	1.2	1.6	0.7
200	4	2.3	35.9	31.6	2.4	18.1	7.6	3.7	3.7	1.6
	2	0.8	0.7	0.7	1.1	0.7	3.9	1.9	1.9	0.7
	1	0.5	0.4	0.4	0.6	0.4	2.0	0.9	1.0	0.4
	1/2	0.3	0.3	0.3	0.3	0.3	1.0	0.5	0.5	0.3
	1/4	0.2	0.2	0.2	0.2	0.2	0.5	0.3	0.3	0.2

Table 1: Values of $\frac{100}{r_{\max}}\sum_{j=1}^{r_{\max}} E_{MC}\left(\bar{L}_j - \bar{L}_{j-1} - (L_j - L_{j-1})\right)^2$ corresponding to different preliminary estimators of q.

One might, for example, use the estimate to compare the losses incurred by models with $j-1$ and j number of factors. In such a case, it is natural to measure the quality of $\bar{L}_p$ by the square root of $E_{MC}\left(\bar{L}_j - \bar{L}_{j-1} - (L_j - L_{j-1})\right)^2$, where E_{MC} denotes the operator of taking mean over MC replications. Table 1 reports values of $\frac{100}{r_{\max}}\sum_{j=1}^{r_{\max}} E_{MC}\left(\bar{L}_j - \bar{L}_{j-1} - (L_j - L_{j-1})\right)^2$ for $\bar{L}_p$ based on different versions of $\hat{r}$, when $(\rho, \beta) = (0.7, 0)$. Results corresponding to the other combinations of ρ and β are qualitatively similar and we do not report them to save space.

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 variation 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 n and T. Overall, ED, ER, GR, and ABC show more robust performance. However, none of these estimators clearly dominates.

Another basis for comparison of different estimates of the loss is related to the

n, T	θ	ED	ER	GR	ABC	BIC	PC1	PC2	IC1	IC2
50	4	1.11	1.11	1.16	1.13	1.42	2.51	2.28	2.57	1.56
	2	1.60	1.34	1.38	1.54	1.18	2.17	1.95	2.21	1.65
	1	1.78	1.37	1.29	1.70	1.28	2.90	2.56	2.96	2.28
	1/2	1.13	1.16	1.08	1.32	1.44	3.66	3.24	3.73	2.91
	1/4	1.09	1.03	1.01	1.30	1.54	4.08	3.58	4.17	3.27
100	4	1.49	1.65	1.64	1.33	1.27	2.05	1.57	2.02	1.11
	2	1.26	1.28	1.14	1.14	1.07	2.96	2.10	2.99	1.22
	1	1.02	1.00	1.00	1.13	1.00	3.68	2.52	3.72	1.35
	1/2	1.02	1.00	1.00	1.19	1.00	4.08	2.76	4.13	1.42
	1/4	1.03	1.00	1.00	1.20	1.00	4.32	2.89	4.42	1.47
200	4	1.01	4.34	3.62	1.04	2.05	1.56	1.08	1.11	1.00
	2	1.01	1.00	1.00	1.05	1.00	1.76	1.11	1.16	1.00
	1	1.01	1.00	1.00	1.04	1.00	1.86	1.10	1.16	1.00
	1/2	1.01	1.00	1.00	1.06	1.00	1.92	1.14	1.19	1.00
	1/4	1.01	1.00	1.00	1.07	1.00	1.99	1.15	1.22	1.00

Table 2: Value of $E_{MC}L_{\hat{p}}/E_{MC}L_{p^*}$ corresponding to different preliminary estimators of q

quality of the corresponding *loss efficient* estimates of the number of factors. Let $\hat{p} = \arg\min_{0\leq p\leq r_{\max}} \bar{L}_p$ and $p^* = \arg\min_{0\leq p\leq r_{\max}} L_p$. Then, a natural measure of quality of $\bar{L}_p$ is $E_{MC}L_{\hat{p}}/E_{MC}L_{p^*}$. Table 2 reports this quality measure for $\bar{L}_p$ based on different versions of $\hat{r}$, when $(\rho, \beta) = (0.7, 0)$. Results corresponding to the other combinations of ρ and β are similar, and are omitted to save space. Loss estimates $\bar{L}_p$ based on ED, ER, GR, and ABC work reasonably well, and better than those based on the other preliminary estimators. However, the choice between ED, ER, GR, and ABC is not obvious.

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 5 reports $\bar{L}_p$ (solid line) and $\hat{L}_p^{(B=I)}$ (dotted line) normalized 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

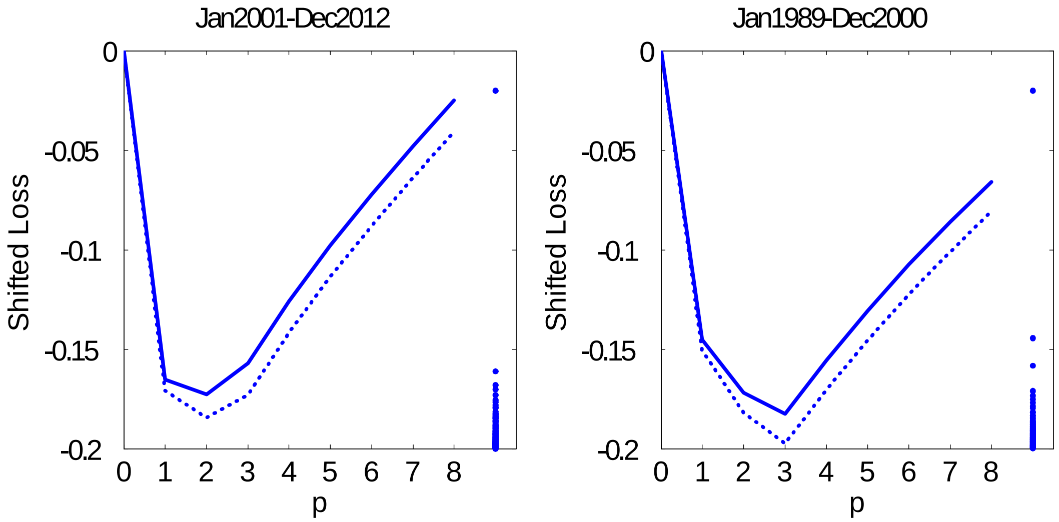

Figure 5: Estimated loss of the PC estimator of a factor model of excess returns. $\bar{L}_p$ – solid lines, $\hat{L}_p^{(B=I)}$ – dotted lines. $\hat{r}$ is set to 3.

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.

Interestingly, the entire exercise repeated for the Jan1989-Dec2000 data yields different results. As the right panel of Figure 5 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 $O_p(1/T) \sim O_p(1/n)$ 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.

The quadratic loss from the estimation of the common component is not the only interesting loss that can be considered in the factor model context. A large number of 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 would be the squared forecast error.

Finally, 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.

7 Appendix

For any matrix A, let $\|A\|$ denote the spectral norm of A, that is $\|A\|$ equals the maximum singular value of A. Let $e^{(j,k)} = \Lambda_{\cdot j}' eF_{\cdot k}/\sqrt{d_{jn}nT}$. Supplementary Appendix contains proofs of the following two auxiliary lemmas.

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

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

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 projection on the space spanned by $F_{\cdot j}$, and let P_0 be the projection on the subspace of $\mathbb{R}^T$ orthogonal to all columns of F. Similarly, 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 $n, T \to_c \infty$,*

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

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

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

$$\begin{aligned} L_p &= \operatorname{tr}[\hat{\Lambda}_{1:p}'\hat{\Lambda}_{1:p}]/n + \operatorname{tr}[\Lambda'\Lambda]/n - 2\operatorname{tr}[\hat{\Lambda}_{1:p}\hat{F}_{1:p}'F\Lambda']/(nT) \\ &= \sum_{j=1}^{p}\mu_j\left(X'X/(nT)\right) + \sum_{j=1}^{r}d_{jn} - 2\sum_{k=1}^{r}\sum_{j=1}^{p}(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot j})(F_{\cdot k}'\hat{F}_{\cdot j})/(nT). \qquad (22) \end{aligned}$$

To simplify (22), consider the identity

$$\begin{aligned}\left|(\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot j})(F'_{\cdot k}\hat{F}_{\cdot j})\right|^2 &= \|\Lambda_{\cdot k}\|^2\,||\hat{\Lambda}_{\cdot j}||^2\,\mathrm{tr}[Q_k\hat{Q}_j]\,\|F_{\cdot k}\|^2\,||\hat{F}_{\cdot j}||^2\,\mathrm{tr}[P_k\hat{P}_j] \\ &= d_{kn}nT^2\mu_j\left(X'X/T\right)\mathrm{tr}[Q_k\hat{Q}_j]\,\mathrm{tr}[P_k\hat{P}_j]. \qquad (23)\end{aligned}$$

For $j \leq r$ and $j \neq k$, by Lemmas 2 and 3, $\mu_j\left(X'X/T\right)\mathrm{tr}[Q_k\hat{Q}_j]\,\mathrm{tr}[P_k\hat{P}_j] = o_{\mathrm{P}}\left(1/T\right)$. Therefore

$$(\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot j})(F'_{\cdot k}\hat{F}_{\cdot j})/\left(nT\right) = o_{\mathrm{P}}\left(1/T\right). \qquad (24)$$

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,

$$\mu_{i+s-1}^{1/2}\left((A+B)\left(A+B\right)'\right) \leq \mu_i^{1/2}\left(AA'\right) + \mu_s^{1/2}\left(BB'\right), \qquad (25)$$

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 $\mu_{r+1}\left(F\Lambda'\Lambda F'/(nT)\right) = 0$, we get

$$\mu_j(X'X)/\left(nT\right) \leq \mu_{j-r}(e'e)/\left(nT\right). \qquad (26)$$

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

$$\mu_j(X'X)/\left(nT\right) \geq \mu_{j+r}(e'e)/\left(nT\right). \qquad (27)$$

Further, for $j > r$, we have $0 \leq \mathrm{tr}[P_k\hat{P}_j] \leq \mathrm{tr}[P_k\hat{P}_j] + \mathrm{tr}[P_k(I_T - \hat{P}_j - \hat{P}_k)] = 1 - \mathrm{tr}[P_k\hat{P}_k]$. Hence, by Lemma 3, $\mathrm{tr}[P_k\hat{P}_j] = O_{\mathrm{P}}\left(1/T\right)$. Similarly, $\mathrm{tr}[Q_k\hat{Q}_j] = O_{\mathrm{P}}\left(1/T\right)$. The latter two equalities together with (23) and the fact that, by (26), $\mu_j\left(X'X/T\right) \leq \mu_1\left(e'e/T\right) = O_{\mathrm{P}}\left(1\right)$ imply (24).

Using (24) together with (22), we obtain

$$L_p = \sum_{j=1}^{p}\mu_j\left(X'X/\left(nT\right)\right) + \sum_{j=1}^{r}d_{jn} - 2\sum_{k=1}^{\min\{p,r\}}(\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot k})(F'_{\cdot k}\hat{F}_{\cdot k})/\left(nT\right) + o_{\mathrm{P}}\left(1/T\right). \qquad (28)$$

Now, by (23), $(\Lambda'_{\cdot k}\hat{\Lambda}_{\cdot k})(F'_{\cdot k}\hat{F}_{\cdot k}) = \pm(d_{kn}nT^2\mu_k\left(X'X/T\right)\mathrm{tr}[Q_k\hat{Q}_k]\,\mathrm{tr}[P_k\hat{P}_k])^{1/2}$. On the other hand,

$$\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot k} = \Lambda_{\cdot k}'X\hat{F}_{\cdot k}/T = d_{kn}nF_{\cdot k}'\hat{F}_{\cdot k}/T + \Lambda_{\cdot k}'e\hat{F}_{\cdot k}/T = d_{kn}nF_{\cdot k}'\hat{F}_{\cdot k}/T + O_{\mathrm{P}}(1). \quad (29)$$

To see that the latter equality holds, note that

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

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 (29) 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,

$$(\Lambda_{\cdot k}'\hat{\Lambda}_{\cdot k})(F_{\cdot k}'\hat{F}_{\cdot k}) = (d_{kn}nT^2\mu_k(X'X/T)\,\mathrm{tr}[Q_k\hat{Q}_k]\,\mathrm{tr}[P_k\hat{P}_k])^{1/2}. \quad (30)$$

Using (28), (30), and Lemmas 2 and 3, we obtain

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

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

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

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

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

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 (26) and (27) 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.□

Proof of Lemma 1.

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

$$\begin{cases} v = x g_1 (u) \\ u = x g_2 (v) \end{cases}, \tag{31}$$

where $g_1 (u) = (c \int \lambda u / (u-\lambda) \, \mathrm{d}\mathcal{G}_A (\lambda))^{-1}$ and $g_2 (v) = (\int \lambda v / (v-\lambda) \, \mathrm{d}\mathcal{G}_B (\lambda))^{-1}$. 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 (31) 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 (31) 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

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

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

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

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

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

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

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 $\mathrm{d}\ln g(x)/\mathrm{dln}x \to 1$ as $x \to \infty$. From the definition of $g(x)$, we get

$$\mathrm{d}\ln g(x)/\mathrm{d}\ln x = 1 + \frac{1}{u_{2x} - 1}\mathrm{d}\ln u_{2x}/\mathrm{d}\ln x + \frac{1}{v_{2x} - 1}\mathrm{d}\ln v_{2x}/\mathrm{d}\ln x. \qquad (32)$$

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 $\mathrm{d}\ln u_{2x}/\mathrm{dln}x$ and $\mathrm{d}\ln v_{2x}/\mathrm{dln}x$ are decreasing functions on $x > \bar{x}$.

Straightforward algebra shows that

$$\begin{aligned} \mathrm{d}\ln u_{2x}/\mathrm{d}\ln x &= \left(1+V\left(x\right)\right)/\left(1-U\left(x\right)V\left(x\right)\right), \text{ and} \qquad (33) \\ \mathrm{d}\ln v_{2x}/\mathrm{d}\ln x &= \left(1+U\left(x\right)\right)/\left(1-U\left(x\right)V\left(x\right)\right), \qquad (34) \end{aligned}$$

where

$$\begin{aligned} V(x) &= \int \frac{\lambda^2}{\left(v_{2x}-\lambda\right)^2}\mathrm{d}\mathcal{G}_B\left(\lambda\right)/\int \frac{\lambda}{v_{2x}-\lambda}\mathrm{d}\mathcal{G}_B\left(\lambda\right), \text{ and} \qquad (35) \\ U\left(x\right) &= \int \frac{\lambda^2}{\left(u_{2x}-\lambda\right)^2}\mathrm{d}\mathcal{G}_A\left(\lambda\right)/\int \frac{\lambda}{u_{2x}-\lambda}\mathrm{d}\mathcal{G}_A\left(\lambda\right). \qquad (36) \end{aligned}$$

Hence, it is enough to prove that $V\left(x\right)$ and $U\left(x\right)$ 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

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}v}\ln\int \frac{\lambda^2}{\left(v-\lambda\right)^2}\mathrm{d}\mathcal{G}_B\left(\lambda\right) - \frac{\mathrm{d}}{\mathrm{d}v}\ln\int \frac{\lambda}{v-\lambda}\mathrm{d}\mathcal{G}_B\left(\lambda\right) &< 0 \text{ for } v > \bar{v}, \text{ and} \qquad (37) \\ \frac{\mathrm{d}}{\mathrm{d}u}\ln\int \frac{\lambda^2}{\left(u-\lambda\right)^2}\mathrm{d}\mathcal{G}_A\left(\lambda\right) - \frac{\mathrm{d}}{\mathrm{d}u}\ln\int \frac{\lambda}{u-\lambda}\mathrm{d}\mathcal{G}_A\left(\lambda\right) &< 0 \text{ for } u > \bar{u}. \qquad (38) \end{aligned}$$

Below, we will establish (37). The proof of (38) is the same after v is replaced by u and $\mathcal{G}_B\left(\lambda\right)$ is replaced by $\mathcal{G}_A\left(\lambda\right)$.

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

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

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

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

so that $\frac{\mathrm{d}}{\mathrm{d}\lambda}h\left(\lambda\right)$ is positive for $\lambda \in [0, vc_2/\left(2c_1+c_2\right))$ and negative for $\lambda \in \left(vc_2/\left(2c_1+c_2\right), v\right)$.

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, (37) holds.□

Proof of Proposition 2.

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

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

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), for $k \neq j$, $\operatorname{plim}\hat{\alpha}_{kj} = 0$, $\operatorname{plim}\hat{\beta}_{kj} = 0$, and $\hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j} = O_{\mathrm{P}}(1)$. Therefore, we have

$$L_p = \sum_{j=1}^{p}\mu_j\left(X'X/(nT)\right) + \sum_{j=1}^{r}\delta_{jn}/n - 2\sum_{j=1}^{p}\sqrt{\delta_{jn}\hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j}}\hat{\alpha}_{jj}\hat{\beta}_{jj}/n + o_{\mathrm{P}}(1/T). \quad (39)$$

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),

$$\operatorname{plim}(\sqrt{\delta_{jn}\hat{\Lambda}'_{\cdot j}\hat{\Lambda}_{\cdot j}}\hat{\alpha}_{jj}\hat{\beta}_{jj}) = \sqrt{\frac{\delta_j \sigma^2 f\left(\delta_j/\sigma^2\right)}{\left(1+\psi_j\theta_j\right)\left(1+\psi_j\omega_j\right)}}, \tag{40}$$

where

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

$x = f\left(\delta_j/\sigma^2\right)$, 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 (35) and (36) of $V(x)$ and $U(x)$, we have

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

Using these equalities, the above definitions of ψ_j, θ_j, ω_j, x, and equalities (32), (33) and (34), 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 (40) and the fact that $g\left(x\right) = \delta_j/\sigma^2$, this implies that

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

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

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

For $p \leq q$, the statement of Proposition 2 follows from the latter equality and (39).

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 (39) as well. □

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.□

Proof of Proposition 3.

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

$$L_p - L_r = \sum_{j=r+1}^{p}\mu_j/n + o_{\mathrm{P}}(1/T). \tag{41}$$

On the other hand, by (11), $\tilde{L}_p - \tilde{L}_{\hat{r}} = \sum_{j=\hat{r}+1}^{p}\mu_j/n$. Since, by definition of $\hat{r}$, under the strong factors asymptotics, $\Pr(r = \hat{r}) \to 1$, we must have

$$\tilde{L}_p - \tilde{L}_r = \sum_{j=r+1}^{p}\mu_j/n + o_{\mathrm{P}}(1/T). \tag{42}$$

Equation (15) follows from equations (41) and (42), 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, (14) holds for $\tilde{L}_p = \underline{L}_p$. For $\tilde{L}_p = \bar{L}_p$, equality (14) 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 the definition of $\hat{r}$, under the weak factors asymptotics, $\Pr\left(\hat{r}=q\right)\to 1$, 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 (16), 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 (11), $\underline{L}_p = \underline{L}_q + \sum_{j=q+1}^{p}\mu_j/n + o_{\mathrm{P}}\left(1/n\right)$.

The convergence (17) can be proven similarly using the following fact. As shown in Lemma 4 below, $\mathrm{plim}\,\rho_j \geq \mathrm{plim}\,\hat{\rho}_j$ 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 (13) of $\bar{L}_p$.□

Lemma 4 *Suppose that assumptions A1w-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* $j \leq q$,

$$\mathrm{plim}\,\rho_j \geq \mathrm{plim}\,\hat{\rho}_j. \tag{43}$$

Furthermore, let $\hat{\rho}_j^{(A=I)}$ *and* $\hat{\rho}_j^{(B=I)}$ *be as defined in (18-19). Then, if* $A = I_n$ *or* $B = I_T$, *for any* $j \leq q$, *we have, respectively,*

$$\mathrm{plim}\,\rho_j = \mathrm{plim}\,\hat{\rho}_j^{(A=I)}, \text{ and} \tag{44}$$

$$\mathrm{plim}\,\rho_j = \mathrm{plim}\,\hat{\rho}_j^{(B=I)}. \tag{45}$$

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, $\mathrm{plim}\,\rho_j = \mathrm{d}\ln f\left(z_j\right)/\mathrm{dln}z$. 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 $x > \bar{x}$,

$$\begin{aligned}\frac{\mathrm{d}\ln g\left(x\right)}{\mathrm{d}\ln x} &= 1 + \frac{1}{u_{2x}-1}\frac{\mathrm{d}\ln u_{2x}}{\mathrm{d}\ln x} + \frac{1}{v_{2x}-1}\frac{\mathrm{d}\ln v_{2x}}{\mathrm{d}\ln x} \\ &\leq 1 + \max\left\{\frac{1}{u_{2x}-1}, \frac{1}{v_{2x}-1}\right\}\left(\frac{\mathrm{d}\ln u_{2x}}{\mathrm{d}\ln x} + \frac{\mathrm{d}\ln v_{2x}}{\mathrm{d}\ln x}\right). \end{aligned} \tag{46}$$

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

$$\left\{ \begin{array}{l} -xm(x) - 1 = -u_{2x} \int (\lambda - u_{2x})^{-1} \mathrm{d}\mathcal{G}_A (\lambda) - 1 \\ -xm(x) - 1 = c^{-1} \left(-v_{2x} \int (\lambda - v_{2x})^{-1} \mathrm{d}\mathcal{G}_B (\lambda) - 1 \right) \\ -xm(x) - 1 = x \left(c u_{2x} v_{2x} \right)^{-1} \end{array} \right. , \tag{47}$$

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 $ee'/(\sigma^2 T)$ as $n, T \rightarrow_c \infty$. Therefore,

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

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 (47), we obtain

$$\frac{1}{u_{2x} - 1} \leq -xm(x) - 1 \text{ and } \frac{1}{v_{2x} - 1} \leq c \left(-xm(x) - 1 \right). \tag{49}$$

From (46), (48), and (49), we get

$$\mathrm{d} \ln g\left(x\right) / \mathrm{d} \ln x \leq 1 + \max \left\{ c, 1 \right\} \left(-1 + x^2 m'(x) \right). \tag{50}$$

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 $e'e/(\sigma^2 T)$ (as opposed to $ee'/(\sigma^2 T)$) as $n, T \rightarrow_c \infty$. 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 (50) when $c \geq 1$, we obtain $\mathrm{d} \ln g\left(x\right) / \mathrm{dln} x \leq x^2 \underline{m}'(x)$. When $c < 1$, (50) simplifies to $\mathrm{d} \ln g\left(x\right) / \mathrm{dln} x \leq x^2 m'(x)$. Therefore, we have $\mathrm{d} \ln g\left(x\right) / \mathrm{dln} x \leq x^2 \max \left\{ m'(x), \underline{m}'(x) \right\}$ for $x > \bar{x}$, and thus

$$\mathrm{d} \ln f\left(z_j\right) / \mathrm{d} \ln z \geq 1 / \left(x_j^2 \max \left\{ m'(x_j), \underline{m}'(x_j) \right\} \right). \tag{51}$$

From (26) and (27) 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,

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

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

$$x_j^2 m'(x_j) - \mu_j^2 \hat{m}'\left(\mu_j\right) \xrightarrow{p} 0 \text{ and } x_j^2 \underline{m}'(x_j) - \mu_j^2 \widetilde{m}'\left(\mu_j\right) \xrightarrow{p} 0. \tag{52}$$

Finally, (51) and (52) imply (43).

Now, assume that $A = I_n$. Then the first equation of (47) can be written as $-xm(x) - 1 = 1/\left(u_{2x} - 1\right)$. Therefore,

$$\frac{1}{u_{2x} - 1} \mathrm{d} \ln u_{2x} / \mathrm{d} \ln x = -1 - xm'\left(x\right) / m\left(x\right). \tag{53}$$

Furthermore, from the third equation of (47), we have $1/\left(u_{2x} - 1\right) = x\left(cu_{2x}v_{2x}\right)^{-1}$ so that $v_{2x} = x\left(u_{2x} - 1\right) / \left(cu_{2x}\right)$. Therefore, after some algebra, we get

$$\frac{1}{v_{2x} - 1} \mathrm{d} \ln v_{2x} / \mathrm{d} \ln x = cxm'\left(x\right) / \left(1 + cm(x)\right). \tag{54}$$

Combining (53) and (54), we obtain $\mathrm{d} \ln g\left(x\right) / \mathrm{dln} x = -xm'\left(x\right) / \left(m(x)\left(1 + cm(x)\right)\right)$, and therefore,

$$\operatorname{plim} \mathrm{d} \ln f\left(z_{jn}\right) / \mathrm{d} \ln\left(z\right) = -\frac{m(x_j)\left(1 + cm(x_j)\right)}{xm'\left(x_j\right)}. \tag{55}$$

On the other hand, similarly to (52),

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

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

$$\operatorname{plim} \hat{\rho}_j^{(A=I)} = -\frac{m(x_j)\left(1 + cm(x_j)\right)}{xm'\left(x_j\right)}, \tag{56}$$

and (55) and (56) imply (44).

The equality (45) can be proven similarly to (44) after $-xm(x) - 1$ in (47) is replaced by $\left(-x\underline{m}(x) - 1\right)/c$. We omit details of such a similar proof to save space.□

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 (44) and (45) of Lemma 4 imply that

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

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

Proof of Proposition 5.

First, let us prove (i). Let $p_0^* = \arg\min_{0 \leq p \leq r_{\max}} L_p^{(1)}$, 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 $\Pr(p^* \geq r) \to 1$, where $p^* = \arg\min_{0 \leq p \leq r_{\max}} L_p$. 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 (27), $\mu_{r+1} > \mu_{2r+1}(e'e/T)$. On the other hand, there must exist $\epsilon > 0$ such that $\Pr\left(\mu_{2r+1}(e'e/T) \geq \epsilon\right) \to 1$. This follows from Assumption A3 (i) and the fact that $\operatorname{tr}(e'e)/(nT) \leq 2r\mu_1(e'e)/(nT) + n\mu_{2r+1}(e'e)/(nT) = \mu_{2r+1}(e'e/T) + o_{\mathrm{P}}(1)$, 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 $\Pr(\hat{p} \geq r) \to 1$ because $\min_{0 \leq p < r} L_p$ is positive and stays away from zero w.p.a.1, whereas $\max_{r \leq p \leq r_{\max}} L_p \xrightarrow{p} 0$. 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 $p_0^* = \arg\min_{0\leq p\leq r_{\max}} L_p^{(1)}$, where now $L_p^{(1)}$ is as defined in Proposition 2. Let us show that $\Pr\left(p_0^* = p^*\right) \to 1$. By Proposition 2, $L_p^{(1)} - L_p = o_{\mathrm{P}}\left(1/T\right) = o_{\mathrm{P}}\left(1/n\right)$. Therefore, it is sufficient to show that there exists $\epsilon > 0$ such that

$$\Pr\left(\min_{0\leq p\leq r_{\max}, p\neq p_0^*} L_p^{(1)} - L_{p_0^*}^{(1)} > \epsilon/n\right) \to 1. \tag{57}$$

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 (27), $\mu_{q+1} > \mu_{2q+1}(e'e/T)$. Onatski (2010) proves that under assumptions A1w-A3w, the empirical distribution of the eigenvalues of $e'e/T$ converges to a fixed distribution, and that $\mu_1\left(ee'/T\right)$ 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 $\Pr\left(p_0^* = p^*\right) \to 1$ 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 $p \leq q$ 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 $1 \leq p \leq \min\left\{\hat{r}, q\right\}$,

$$\bar{L}_{p-1} - \bar{L}_p \leq L_{p-1}^{(1)} - L_p^{(1)} + o_{\mathrm{P}}\left(1/T\right) \text{ and} \tag{58}$$

$$\underline{L}_{p-1} - \underline{L}_p \geq L_{p-1}^{(1)} - L_p^{(1)} + o_{\mathrm{P}}\left(1/T\right). \tag{59}$$

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

$$\bar{L}_{p+1} - \bar{L}_p = L_{p+1}^{(1)} - L_p^{(1)} \text{ and} \tag{60}$$

$$\underline{L}_{p+1} - \underline{L}_p = L_{p+1}^{(1)} - L_p^{(1)}. \tag{61}$$

Since $\hat{r}$ is consistent for q, inequalities (58)-(59), equalities (60)-(61), and the convergence (57) 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 (57) and Proposition 4 (ii) also imply part (iii).□

References

[1] Ahn, S.C., and Horenstein, A.R. (2012) "Eigenvalue Ratio Test for the Number of Factors", forthcoming in *Econometrica*.

[2] Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle", in Petrov, B. and Csáki, F. (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 Ng, S (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 Ng, S (2008). "Large Dimensional Factor Analysis", Foundations and Trends in Econometrics 3, 89-163.

[8] Boivin, J. and S. Ng (2006) "Are more data always better for factor analysis", *Journal of Econometrics* 132, 169-194.

[9] 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.

[10] Connor, G. (1995), "The three types of factor models: A comparison of their explanatory power," *Financial Analysts Journal* 51, 42-46.

[11] Connor, G. and Korajczyk, R. (1993) "A test for the number of factors in an approximate factor model", *Journal of Finance* 58, 1263-1291.

[12] 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

[13] 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.

[14] Forni, M., and L. Reichlin (2001) "Federal policies and local economies: Europe and the US," *European Economic Review* 45, 109-134.

[15] 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.

[16] Hansen, B.E. (1996) "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis", *Econometrica* 64, 413-430.

[17] Hansen, B.E. (2007) "Least squares model averaging", *Econometrica* 75, 1175-1189.

[18] 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.

[19] Horn, R.A., and C.R. Johnson (1991), *Topics in Matrix Analysis*, Cambridge University Press.

[20] Huang, R. and Jo, H. (1995). "Data frequency and the number of factors in stock returns", *Journal of Banking and Finance* 19, 987-1003.

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

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

[23] Kunitomo, N., and T. Yamamoto, (1985), "Properties of Predictors in Misspecified Autoregressive Time Series Models", *Journal of the American Statistical Association* 80, 941-950.

[24] Lancaster, T. (2000) "The Incidental Parameter Problem Since 1948", *Journal of Econometrics* 95, 391-413.

[25] 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.

[26] Mallows, C.L. (1973) "Some comments on C_p", *Technometrics* 15, 661-675.

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

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

[29] Onatski, A. (2010), "Determining the Number of Factors From Empirical Distribution of Eigenvalues", *Review of Economics and Statistics* 92 (4), 1004-1016.

[30] Onatski, A. (2012), "Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors", *Journal of Econometrics* 168, 244-258

[31] Onatski, A., Moreira, M. J., and M. Hallin (2013) "Asymptotic Power of Sphericity Tests for High-dimensional Data", *Annals of Statistics* 41 (3), 1204-1231.

[32] Phillips, P.C.B. (1979), "The sampling distribution of forecasts from a first-order autoregression", *Journal of Econometrics* 9, 241-261.

[33] Shao, J. (1997) "An asymptotic theory for linear model selection", *Statistica Sinica* 7, 221-264.

[34] 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.

[35] 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

[36] 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.

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

[38] 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.