Asymptotics of the principal components estimator of large factor models with weakly influential factors

Alexei Onatski*
Faculty of Economics, University of Cambridge

First draft: November, 2005
This draft: May, 2011

Abstract

This paper introduces a drifting-parameter asymptotic framework to derive accurate approximations to the finite sample distribution of the principal components (PC) estimator in situations when factors' explanatory power does not strongly dominate the explanatory power of the cross-sectionally and temporally correlated idiosyncratic terms. Under our asymptotics, the PC estimator is inconsistent. We find explicit formulae for the amount of the inconsistency, and propose an estimator of the number of factors for which the PC estimator works reasonably well. For the special case when the idiosyncratic terms are cross-sectionally but not temporally correlated (or vice versa), we show that the coefficients in the OLS regressions of the PC estimates of factors (loadings) on the true factors (true loadings) are asymptotically normal, and find explicit formulae for the corresponding asymptotic covariance matrix. We explain how to estimate parameters of the derived asymptotic distributions. Our Monte Carlo analysis suggests that our asymptotic formulae and estimators work well even for relatively small n and T. We apply our theoretical results to test a hypothesis about the factor content of the US stock return data.

*Faculty of Economics, University of Cambridge, Sidgwick Avenue,Cambridge,CB3 9DD. E-mail: ao319@cam.ac.uk.
I am grateful to the co-editor, Takeshi Amemiya, the anonymous associate editor, and three anonymous referees for excellent, helpful comments.

JEL code: C13, C33. Key words: approximate factor models, principal components, weakly influential factors, weak factors, inconsistency, bias, asymptotic distribution, Marchenko-Pastur law.

1 Introduction

Approximate factor models have recently attracted an increasing amount of attention from researchers in macroeconomics and finance (see Reichlin (2003), Stock and Watson (2006) and Bai and Ng (2008) for a survey of numerous applications). The most popular technique for estimating factors in such models is the principal components (PC) analysis. Its consistency and asymptotic normality have been shown by Bai (2003). Unfortunately, as Monte Carlo experiments show (see, for example, Boivin and Ng (2006), Uhlig (2008) and Bai and Ng (2008)), the finite sample performance of the PC estimator is poor when the explanatory power of factors does not strongly dominate the explanatory power of the idiosyncratic terms. Such a situation is often encountered in practice. Its hallmark is the absence of clearly visible separation of the eigenvalues of the sample covariance matrix of the data into a group of large eigenvalues representing factor-related variation and a group of small eigenvalues representing idiosyncratic variation.

This paper shows how and why the principal component estimates for large factor models might not be appropriate. We develop asymptotic approximation to the finite sample biases of the PC estimator due to the relatively weak explanatory power of factors. We explicitly link these biases to the covariance structure of the idiosyncratic terms and show that they can be extremely large. We explain how to detect situations in which PC estimator breaks down, and how to estimate the parameters of our asymptotic approximations in cases when the PC estimator is only moderately biased. Our Monte Carlo experiments confirm good approximation quality of our asymptotics in finite samples with weakly influential factors.

Let us describe our main results in more detail. We consider static forms of the

approximate factor models[1] with k factors:

$$X_{it} = \sum_{j=1}^{k} L_{ij} F_{tj} + e_{it} \text{ with } i \in \mathbb{N} \text{ and } t \in \mathbb{N}, \tag{1}$$

where F_{tj} and L_{ij} are the values of the j-th factor at time t and of the loading of this factor on the i-th cross-sectional unit, respectively, and where e_{it} are possibly cross-sectionally and temporally correlated idiosyncratic components of X_{it}. Let X be an observed $n \times T$ matrix with elements X_{it}, and let F, L and e be unobserved $T \times k$, $n \times k$ and $n \times T$ matrices with elements F_{tj}, L_{ij} and e_{it}, respectively. Then we can write: $X = LF' + e$.

In this paper, we will treat both L and F as parameters of the distribution of X. In cases when factors are random, such an approach is equivalent to conditioning on a particular realization of F. Further, we will assume that the matrix of the idiosyncratic terms can be represented as $e = A\varepsilon B$, where A and B are relatively unrestricted $n \times n$ and $T \times T$ matrices and ε is an $n \times T$ matrix with i.i.d. $N(0, \sigma^2)$ entries. Similar assumptions have been previously made in Onatski (2009), Bai and Ng (2005) and Harding (2006). The assumption allows the idiosyncratic terms to be non-trivially correlated both cross-sectionally and over time. We discuss its relation to economic models in Section 2.

The asymptotic identification of the unobserved components of X is achieved by the following standard requirements. First, the factors and loadings are normalized so that $F'F/T = I_k$ and $L'L$ is a diagonal matrix with non-increasing elements along the diagonal. Such a normalization separately identifies F and L from the product LF' up to the simultaneous multiplication of the corresponding columns of F and L by -1. Second, the idiosyncratic terms are only weakly correlated so that:

$$\lim \sup_{n,T \to \infty} \max \text{eval}\, E\left(\frac{1}{T} ee'\right) < \infty, \tag{2}$$

where $\max \text{eval}\,(M)$ denotes the maximal eigenvalue of matrix M. Finally, the factors are pervasive in the sense that their cumulative loadings on n cross-sectional units

[1] The most general approximate factor models have factor loadings represented by possibly infinite lag polynomials (see Forni et al., 2000). When the order of such lag polynomials is bounded, the model can be rewritten in the static form, where the factor loadings are constants and factors are augmented by a set of their own lags. For a discussion of the terminology used in the factor model literature see, for example, Stock and Watson (2006).

rise proportionally to n :

$$\frac{1}{n}L'L \to S > 0. \tag{3}$$

The PC estimator of F, $\hat{F}$, is defined as $\sqrt{T}$ times the matrix of the k principal eigenvectors of a sample-covariance-type matrix $X'X/T$, and the PC estimator of L, $\hat{L}$, is defined as $X\hat{F}/T$. We would like to study the properties of the PC estimators in the situation when the factors' *finite sample* explanatory power, as measured by the diagonal elements of $L'L$, is weak, that is, only moderately larger than $\max \text{eval}\, E\left(\frac{1}{T}ee'\right)$.

Note that if we fix model (1) and let n and T go to infinity, assumptions (2) and (3) would imply that, asymptotically, any of the diagonal elements of $L'L$ is infinitely larger than $\max \text{eval}\, E\left(\frac{1}{T}ee'\right)$. Hence, such an asymptotics would not provide a useful approximation to the finite samples with relatively weak factors. We will therefore consider a different asymptotics, where models (1) are drifting as n and T tend to infinity so that the *finite sample* explanatory power of factors remains bounded. Formally, we will consider a sequence of models (1) indexed by the cross-sectional dimension n, so that

$$L^{(n)\prime}L^{(n)} - D \to 0 \tag{4}$$

as n and $T^{(n)}$ go to infinity proportionately, where D is a fixed diagonal matrix.

The above asymptotic device is similar to the well-known Pitman drift (see, for example, Davidson and MacKinnon (2004,.p.434)), which is usually used in the asymptotic power comparisons of consistent tests.[2] In this paper, however, a parameter drift is introduced to obtain an accurate asymptotic approximation to the finite sample distribution of a particular estimator. Such a strategy is not new. Bekker (1994) and Staiger and Stock (1997) use the drifting parameter device to obtain an accurate asymptotic approximation to the finite sample distribution of 2SLS and LIML estimators when there is a substantial amount of over-identification and when the instruments are weak, respectively.[3]

[2] As pointed out by McManus (1991), although the introduction of the local alternatives analysis is often attributed to Pitman, it seems to first appear in Neyman (1937). Considering shrinking neighborhoods in statistical experiments is the basis of the classical local asymptotic approach to statistical experiments associated with the name of Le Cam, among many others. This paper does not take the approach of the "asymptotic theory of statistical experiments". For such an approach in a context linked to factor models see Onatski et al. (2011).

[3] Another important goal of their analysis is to compare the performance of 2SLS and LIML. In our paper, however, obtaining an accurate asymptotic approximation to a particular estimator is

Bekker (1994, p.661) provides an illuminating discussion of the strategy of using drifting parameters to obtain an accurate asymptotic approximation to the exact distribution of a specific estimator. He stresses the point that the parameter sequence should be "designed to fit the finite sample distribution better and should not be considered as an assumption about the behavior of observations in case of further sampling". For the choice of the parameter sequence, he suggests the following principle: the parameter sequence should be designed "so that it generates acceptable approximations of known distributional properties of related statistics".

In this paper, as statistics related to the PC estimator, we consider the eigenvalues of the sample covariance matrix of the data. In applications, the eigenvalues typically do not separate into visually distinct groups of large and small eigenvalues. Such a behavior is consistent with the drifting-parameter asymptotics described above, but it does not accord with the standard fixed-parameter asymptotics, under which the distance between the two groups of eigenvalues must diverge to infinity.

The main focus of our analysis will be on the behavior of $\hat{\beta} \equiv (F'F)^{-1} F'\hat{F} = \frac{F'\hat{F}}{T}$, the matrix of the coefficients of the OLS regressions of the PC estimates of factors on the true factors, under the weakly influential factor asymptotic regime (4). As pointed out by Bai (2003, p.151), the i-th diagonal element of $\hat{\beta}$ can be considered as a measure of consistency of the PC estimator of the i-th factor. Under the standard asymptotics (3), $\hat{\beta}$ converges in probability to I_k, so that $\hat{F}$ is consistent for F.[4] Furthermore, under assumptions of Bai (2003), $\hat{\beta} - I_k = o_p\left(T^{-1/2}\right)$, which is a sufficient condition for the negligibility of the estimation error due to the replacement of the true factors by their PC estimates in factor-augmented regressions (see Bai, 2003, p.146).

In Theorem 1, we prove that $\operatorname{plim} \hat{\beta}$ under the weakly influential factor asymptotics is not I_k, but equals a diagonal matrix whose diagonal elements are strictly less than one.[5] Hence, the PC estimator is inconsistent under our asymptotics. We describe the diagonal elements of $\operatorname{plim} \hat{\beta}$ as specific functions of matrix D from (4), which measures the finite sample strength of factors; of σ^2, which scales the variance of the idiosyncratic terms; and of the limits of the empirical eigenvalue distributions of the matrices A and B, which encode the degree of the cross-sectional and temporal

the only goal.

[4] Here we assume that the sign indeterminacy of $\hat{F}$ is resolved by adding the normalization requirement that the diagonal elements of $\hat{F}'F$ are non-negative.

[5] All theorems in the paper, and Theorem 1 in particular, describe the asymptotics of $\hat{L}'\hat{L}$ and $\hat{\alpha} = (L'L)^{-1/2} L'\hat{L}\left(\hat{L}'\hat{L}\right)^{-1/2}$ in addition to that of $\hat{\beta}$.

correlation of the idiosyncratic terms. We show that when the strength D_{ii} of the i-th factor F_i is below a certain threshold, $\operatorname{plim} \hat{\beta}_{ii} = 0$ so that $\hat{F}_i$ is not just inconsistent, but orthogonal to F_i.

In Theorem 2, we study the asymptotic distribution of $\hat{\beta}$ around its probability limit. We show that, in the special case when the idiosyncratic terms are only cross-sectionally correlated, $\hat{\beta}$ is asymptotically normal and we find explicit formulae for the elements of the corresponding asymptotic covariance matrix.

Although very restrictive, the assumption of no temporal correlation in the idiosyncratic terms is acceptable in some applications. For example, Chamberlain and Rothschild (1983), who introduce the approximate factor model to the literature, model the idiosyncratic components of the excess stock returns as cross-sectionally but not temporally correlated random variables. Since the excess stock returns are poorly predictable, such an assumption can be viewed as a good first-order approximation. In Section 5, we use results of Theorem 2 to test a hypothesis that the celebrated Fama-French factors (see Fama and French, 1993) span the factor space of the US excess stock return data.

Our last theorem, Theorem 3, describes consistent estimators of the parameters of the asymptotic distributions derived in Theorem 2. As we explain in Section 4, under our weakly influential factor asymptotics, the true number of factors k is, in general, unidentified. What can be identified is the number of factors q that are not orthogonal to their PC estimates asymptotically. Therefore, Theorem 3 describes parameters of the asymptotic distributions of only those $\hat{\beta}_{ij}$ for which i and j are no larger than q.

We show that the estimator $\hat{q}$ of the number of factors proposed in Onatski (2009) is consistent for q under our asymptotics. Finding $\hat{q}$ for a particular dataset gives us a simple procedure for selecting the number of factors for which the PC estimator does not break down. In Section 5, we find that for the widely used Stock-Watson macroeconomic data, described in Watson (2003), $\hat{q} = 2$. We link this finding to the fact that, although Stock and Watson (2005) estimate no less than seven factors in their data, using more than two of the estimated factors in the forecasting exercises reported in Stock and Watson (2002) does not significantly improve the quality of the forecasts.

Let us now describe some related literature. An alternative approach to modeling weakly influential factors has been recently proposed in DeMol et al. (2008). The au-

thors of that paper replace (2) by a weaker assumption: $\lim \sup_{n,T\to\infty} \max \text{eval}\, E\left(\frac{1}{n^{1-\alpha}T}ee'\right) < \infty$, where $0 < \alpha \leq 1$. Dividing the data by $n^{(1-\alpha)/2}$ and redefining L and e as $n^{(\alpha-1)/2}L$ and $n^{(\alpha-1)/2}e$, respectively, we see that such a modeling strategy is equivalent to maintaining (2) but assuming that $n^{-\alpha}L'L \to S > 0$ with $0 < \alpha \leq 1$. When $0 < \alpha < 1$, the PC estimator remains consistent but its rate of convergence decreases relative to the strong factor case: $\alpha = 1$.

Our approach differs from that of DeMol et al. (2008) in several respects. On the one hand, as explained above, we view our asymptotics as a Pitman drift device rather than as a description of the process of further sampling. Therefore, our assumptions about the asymptotic parameter drift is more difficult to justify than the asymptotic assumptions of DeMol et al. (2008). On the other hand, we analyze the case when $L'L \to S > 0$, which formally corresponds to the case $\alpha = 0$ not considered by DeMol et al. (2008). Such an extension leads to the inconsistency of the PC estimator and allows us to obtain sharper asymptotic description of the finite sample biases of the PC estimator documented in Boivin and Ng (2006), Uhlig (2008) and Bai and Ng (2008).

In the statistical literature, the weakly influential factors asymptotics of the PC estimators have been recently studied by Johnstone and Lu (2007) and by Paul (2007). For a 1-factor model with i.i.d. Gaussian factor and i.i.d Gaussian idiosyncratic terms, Johnstone and Lu (2007) show that the one-dimensional analog of our $\hat{\beta}$ remains separated from one as n and T go to infinity proportionately. Paul (2007) quantifies the amount of the inconsistency pointed out by Johnstone and Lu (2007) for the case of i.i.d. Gaussian data such that all but k distinct eigenvalues of the population covariance matrix are the same. For the same model, Paul (2007) finds the asymptotic distribution of the eigenvectors corresponding to the k largest eigenvalues.

In contrast to Johnstone and Lu (2007) and Paul (2007), our asymptotic analysis does not require the idiosyncratic terms be independent. Allowing the idiosyncratic terms to be correlated is crucial for macroeconomic and financial applications. Further, our proofs use a substantially different machinery than the proofs of Paul (2007), which spares us from the necessity of making Paul's (2007) assumption that factors are i.i.d. Gaussian. Finally, Paul (2007) does not show how to estimate the parameters of the asymptotic distributions that he obtains. In contrast, this paper describes such estimators.

The rest of the paper is organized as follows. In Section 2 we state our assumptions

and obtain our theoretical asymptotic results. In Section 3 we discuss the identification under our weakly influential factors asymptotics, and obtain consistent estimators of the parameters of the asymptotic distributions derived in Section 2. Section 4 contains a Monte Carlo analysis. Section 5 describes empirical applications of our theoretical results. Section 6 concludes. All proofs are relegated to the Technical Appendix available from the author's web site at http://www.columbia.edu/~ao2027.

2 Asymptotic distributions

In this section, we derive the asymptotic distributions of the coefficients from the OLS regressions of the PC estimates of factors on the true factors and of the PC estimates of the normalized factor loadings on the true normalized factor loadings. We assume that the estimates are based on finite samples of increasing dimensions from a sequence of approximate factor models (1). Finite samples of the cross-sectional size n and temporal size $T^{(n)}$ are summarized in $n \times T^{(n)}$ matrices $X^{(n)}$, which can be represented as:

$$X^{(n)} = L^{(n)} F^{(n)\prime} + e^{(n)}, \tag{5}$$

where parameters of the representation satisfy Assumptions 1, 2, and 3 described below. As explained in the Introduction, we treat both $L^{(n)}$ and $F^{(n)}$ as parameters of the distribution of $X^{(n)}$. In what follows, we will omit the superscript (n) from all notations to make them easier to read. The reader should, however, keep in mind that parameters of finite sample distributions may change as the sample size grows.

Assumption 1: *There exist positive constant c and a $k \times k$ diagonal matrix $D \equiv \text{diag}\left(d_1, ..., d_k\right)$, $d_1 > ... > d_k > 0$, such that, as $n \to \infty$:*

i) $n/T \to c$,

ii) $\frac{1}{T} F'F = I_k$ and $L'L$ is diagonal with $L_1'L_1 \geq ... \geq L_k'L_k$,

iii) $L'L \to D$.

In many applications of factor models, the cross-sectional size of the data is comparable to their temporal size. Part i) of the assumption requires n and T be comparable even asymptotically. Part ii) of the assumption is standard. It describes

the normalization of factors and loadings. Part iii) of the assumption describes the weakly influential factors asymptotics, discussed in the Introduction.

Once again, we would like to stress that the reader should not confuse the standard asymptotics, in which growing n and T correspond to the "natural" process of further series becoming observed, and the "auxiliary", non nested sequence used in this paper. In particular, under the weakly influential factor asymptotics, the entries L_{ij} of L may change as $n \to \infty$, and may converge to zero to satisfy (4).

Our next assumption imposes a structure on the matrix e of the idiosyncratic terms. We require that

$$e = A\varepsilon B, \tag{6}$$

where the $n \times T$ matrix ε has i.i.d. elements, and matrices A and B introduce the cross-sectional and temporal correlation in e. Such a modeling of the idiosyncratic correlation structure is convenient but restrictive. Indeed, the covariance matrix of the $nT \times 1$ vector of the stacked columns of e must have form $B'B \otimes AA'$. How well such a Kronecker product can approximate more interesting covariance structures depends on the details of these structures. For a general discussion of the quality of approximations with Kronecker products see Van Loan and Pitsianis (1993).

In economic applications, the covariance matrix of the vector of the stacked columns of e exactly equals the Kronecker product of two matrices only in special cases. For example, in the spirit of Forni and Lippi (1999, 2001), consider an n-industry constant-returns economy, where the productions X_{it} in industries $i = 1, ..., n$ at time t are given by the equations:

$$\begin{pmatrix} 1 & w_{12} & \dots & w_{1n} \\ w_{21} & 1 & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & 1 \end{pmatrix} \begin{pmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{nt} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} F_t + \begin{pmatrix} b_1(L)\varepsilon_{1t} \\ b_2(L)\varepsilon_{2t} \\ \vdots \\ b_n(L)\varepsilon_{nt} \end{pmatrix},$$

where F_t is a demand common shock, $b_i(L)\varepsilon_{it}$ are auto-correlated idiosyncratic productivity shocks, and ε_{it} are i.i.d. innovations to these shocks. For such a model, w_{ji} is the quantity of the i-th product necessary as a means of production to produce one unit of the j-th output. Inverting the input-output matrix W, we obtain:

$$X_t = \Lambda F_t + W^{-1}\varepsilon_t b(L),$$

where $b(L) \equiv \text{diag}\,(b_1(L), b_2(L), ..., b_n(L))$. In the special case when all the productivity shocks have the same dynamics described by the filter $b_i(L) = b_0 + b_1 L + b_2 L^2 + ...$, we can write: $e = A\varepsilon B$, where $A = W^{-1}$ and B is such that $B_{ij} = 0$ for for $i > j$ and $B_{ij} = b_{j-i}$ for for $i \leq j$.[6]

Under standard asymptotics (3), changing A and B, while holding the size of $\max \text{eval}\left(\frac{1}{T} ee'\right)$ fixed, does not have a first order effect on the quality of the PC estimators $\hat{F}$ and $\hat{L}$. This is not so under Assumption 1 iii). A change of the coordinates of the columns of F in the basis formed by the eigenvectors of $B'B$ would have a first-order effect on the quality of $\hat{F}$. The larger the projection of F on the eigenspaces of $B'B$ corresponding to larger eigenvalues and the smaller its projection on the eigenspaces corresponding to smaller eigenvalues, the better the performance of $\hat{F}$. Similarly, changing the position of the columns of L relative to the eigenvectors of $A'A$ has a first order effect on the quality of $\hat{L}$.

To keep our analysis as simple as possible, we will assume that the columns of F are eigenvectors of $B'B$. To avoid putting the PC estimator at an arbitrary advantage or disadvantage by such an assumption, we will require that the eigenvalues corresponding to the columns of F equal the average of the eigenvalues of $B'B$. Similarly, we will assume that the columns of L are eigenvectors of AA' with the corresponding eigenvalues equal to the average of the eigenvalues of $A'A$. We will check the robustness of our conclusions to such a simplifying assumption in the Monte Carlo section below. Finally, we will normalize A and B so that $\text{tr}\,(AA') = n$ and $\text{tr}\,(B'B) = T$. Such a normalization can always be achieved by scaling the variance of the entries of ε.

Assumption 2: *The matrices ε, A and B in the decomposition (6) are as follows.*

i) *ε is an $n \times T$ matrix with i.i.d. $N(0, \sigma^2)$ entries,*

ii) *A is such that $\text{tr}\,(AA') = n$ and $(AA')\,L = L$,*

iii) *B is such that $\text{tr}\,(B'B) = T$ and $(B'B)\,F = F$.*

Let us denote the eigenvalues of AA' and of $B'B$ as $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_T$. Further, let $G_A(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{a_i \leq x\}$ and $G_B(x) = \frac{1}{T}\sum_{i=1}^{T} 1\{b_i \leq x\}$ be the empirical distribution functions of the eigenvalues of AA' and of $B'B$.[7]

[6] Here we assume that the innovations ε_{it} for $t = 0, -1, -2, ...$ equal zero.

[7] Here $1\{\cdot\}$ denotes the indicator function of the set in the brackets.

Assumption 3: *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 (cdf) $\mathcal{G}_A(x)$ 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 $n \to \infty$:*

i) $G_A(x) \to \mathcal{G}_A(x)$ *and* $G_B(x) \to \mathcal{G}_B(x)$ *for all* $x \in \mathbb{R}$,

ii) $\max_{i \leq n} a_i \to \bar{x}_A$ *and* $\max_{i \leq T} b_i \to \bar{x}_B$,

iii) $\inf_{x \in (\underline{x}_A, \bar{x}_A)} \frac{d}{dx}\mathcal{G}_A(x) > 0$ *and* $\inf_{x \in (\underline{x}_B, \bar{x}_B)} \frac{d}{dx}\mathcal{G}_B(x) > 0$.

Parts i) and ii) of the assumption would be satisfied if $\{a_i,\ i = 1, ..., n\}$ and $\{b_i,\ i = 1, ..., T\}$ are random samples from $\mathcal{G}_A$ and $\mathcal{G}_B$. Part iii) of the assumption is made to make sure that the eigenvalues of ee'/T cluster together in the sense that, as $n \to \infty$, the distance between any two consequent eigenvalues of ee'/T converges to zero in probability. In particular, any finite number of the largest eigenvalues of ee'/T will cluster together as $n \to \infty$. As we explain in the Identification and Estimation section of this paper, such a condition is crucial for the identification under our weakly influential factor asymptotic regime.

Zhang (2006) shows that, under Assumptions 1 i), 2 i) and 3 i), the empirical distribution of the eigenvalues of $\frac{1}{\sigma^2 T}ee'$ converges to a cdf $\mathcal{G}(x)$, which can be uniquely determined from $\mathcal{G}_A(x)$ and $\mathcal{G}_B(x)$. Onatski (2009) shows that, if, in addition, Assumptions 3 ii) and iii) are satisfied,[8] any of the finite number of the largest eigenvalues of $\frac{1}{\sigma^2 T}ee'$ converges to the upper boundary of the support of $\mathcal{G}$, $\bar{x} \equiv \min\{x : \mathcal{G}(x) = 1\}$. Zhang's (2006) and Onatski's (2009) results play central role in our technical analysis below. Therefore, we will discuss them here in some detail.

A closed form expression for $\mathcal{G}(x)$ exists only in very special cases. For example, when both A and B are identity matrices,[9] $\mathcal{G}(x)$ is the so-called Marchenko-Pastur distribution, whose density is a known algebraic function of x (see Marchenko and Pastur, 1967). For general A and B, $\mathcal{G}(x)$ is described in terms of its Stieltjes transform, which is defined as $m(z) \equiv \int (\lambda - z)^{-1} d\mathcal{G}(\lambda)$, where $z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$. The Stieltjes transform is also well-defined for real z outside the support of the distribution

[8] Both Zhang (2006) and Onatski (2009) use weaker assumptions than Assumptions 2 i) and 3 i), ii) and iii). Here we sacrifice some generality to make our assumptions easier to interpret.

[9] In such a case, $\underline{x}_A = \bar{x}_A = 1$ and $\underline{x}_B = \bar{x}_B = 1$. Note that the case of the empty interior of $[\underline{x}_A, \bar{x}_A]$ and/or $[\underline{x}_B, \bar{x}_B]$ is not excluded by Assumption 3.

$\mathcal{G}$. The Frobenius-Perron inversion formula $\mathcal{G}\{[a,b]\} = \frac{1}{\pi}\lim_{\eta\to 0^+}\int_a^b \operatorname{Im} m(\xi+i\eta)\,d\xi$, where a and b are points of continuity of $\mathcal{G}(x)$, insures that we can reconstruct $\mathcal{G}(x)$ from $m(z)$.

Let us denote the Stieltjes transforms of $\mathcal{G}_A(x)$ and $\mathcal{G}_B(x)$ as $m_A(z)$ and $m_B(z)$, respectively. Zhang (2006) proves that, for each $z \in \mathbb{C}^+$, $m(z)$, together with two other analytic on $\mathbb{C}^+$ functions $u(z)$ and $v(z)$ constitute a solution to the system:

$$\begin{cases} zm(z)+1 = u(z)\, m_A(u(z)) + 1 \\ zm(z)+1 = c^{-1}\left[v(z)\, m_B(v(z)) + 1\right] \\ zm(z)+1 = -c^{-1}\frac{z}{u(z)v(z)} \end{cases}, \tag{7}$$

which is unique in the set $\{(m(z),u(z),v(z)) : \operatorname{Im} m(z) > 0, \operatorname{Im}(u(z)) > 0, \operatorname{Im}(v(z)) > 0\}$.

For our analysis of the principal components estimator, the asymptotic behavior of the largest eigenvalues of $\frac{1}{\sigma^2 T}ee'$ is of particular interest. Onatski (2009) shows that, under Assumptions 1 i), 2 i) and 3, they converge to $\bar{x}$, the upper boundary of the support of $\mathcal{G}$. He explains how to find $\bar{x}$ directly from system (7). Details of Theorem 1 below are related to this procedure, and therefore, we describe it below.

After substituting the third equation of (7) into the first two equations, rearranging, and replacing the complex variables z, $u(z)$ and $v(z)$ by real variables x, u and v, we have:

$$\begin{cases} v = xc^{-1}u^{-1}\left(-um_A(u)-1\right)^{-1} \\ u = xv^{-1}\left(-vm_B(v)-1\right)^{-1} \end{cases}. \tag{8}$$

Note that $m_A(u)$ and $m_B(v)$ are well-defined if $u > \bar{x}_A$ and $v > \bar{x}_B$. Therefore, we will consider system (8) only in the domain $U = \{(u,v) : u > \bar{x}_A \text{ and } v > \bar{x}_B\}$. Onatski (2009) shows that, under Assumptions 1 i), 2 i) and 3, for any $x < \bar{x}$, system (8) has no solutions in U. For $x = \bar{x}$, there exists exactly one such solution, which we will denote as $\bar{u}, \bar{v}$. For $x > \bar{x}$, there exist two such solutions u_{1x}, v_{1x} and u_{2x}, v_{2x}, where $u_{2x} > u_{1x}$, $v_{2x} > v_{1x}$, and u_{2x} and v_{2x} equal the analytic continuations of $u(z)$ and $v(z)$ from (7) on the subset of real line $z \in (\bar{x}, +\infty)$, evaluated at $z = x$.

Since, as has been shown in Lemma A4 of Onatski (2009), $u^{-1}\left(-um_A(u)-1\right)^{-1}$ and $v^{-1}\left(-vm_B(v)-1\right)^{-1}$ are strictly concave functions of u and v, it is easy to solve (8) numerically for any real x, given distributions $\mathcal{G}_A$ and $\mathcal{G}_B$. The value of $\bar{x}$ can be numerically found as the smallest x such that the solution to (8) exists.

Our first theorem uses the above machinery to establish the probability limits

of the coefficients from the OLS regressions of $\hat{F}$ on F and of $\hat{\mathcal{L}} = \hat{L}\left(\hat{L}'\hat{L}\right)^{-1/2}$ on $\mathcal{L} = L\left(L'L\right)^{-1/2}$. To formulate the theorem, we will need the following new definitions. Let $q \in \{0, 1, ..., k\}$ be such that

$$\frac{d_i}{\sigma^2} > \bar{x}\left(1-\bar{u}^{-1}\right)\left(1-\bar{v}^{-1}\right) \text{ when } 1 \leq i \leq q \text{ and} \tag{9}$$

$$\frac{d_i}{\sigma^2} \leq \bar{x}\left(1-\bar{u}^{-1}\right)\left(1-\bar{v}^{-1}\right) \text{ when } q < i \leq k. \tag{10}$$

For any $1 \leq i \leq q$, let us define x_i, u_i and v_i as, respectively, such x, u_{2x} and v_{2x} that $\frac{d_i}{\sigma^2} = x\left(1-u_{2x}^{-1}\right)\left(1-v_{2x}^{-1}\right)$. Since the right hand side of the latter equality is a strictly increasing function of x, the values x_i, u_i and v_i are well-defined and can be found numerically, given $\mathcal{G}_A$ and $\mathcal{G}_B$.[10] Finally, let us define

$$\begin{aligned}
\psi_i &= c\left(m\left(x_i\right) + x_i m'(x_i)\right), \\
\theta_i &= u_i + \frac{1 + u_i m_A(u_i)}{m_A(u_i) + u_i m'_A(u_i)} \frac{u_i - v_i}{u_i - 1} \text{ and} \\
\omega_i &= v_i + \frac{1 + v_i m_B(v_i)}{m_B(v_i) + v_i m'_B(v_i)} \frac{v_i - u_i}{v_i - 1}.
\end{aligned}$$

Theorem 1: *Let Assumptions 1-3 hold and let* $\hat{\beta} = \left(F'F\right)^{-1}F'\hat{F}$ and $\hat{\alpha} = \left(\mathcal{L}'\mathcal{L}\right)^{-1}\mathcal{L}'\hat{\mathcal{L}}$. *Then, as* n *(and by Assumption 1 i) also* T*) goes to infinity,* $\hat{\beta}$, $\hat{\alpha}$ *and* $\hat{L}'\hat{L}$ *converge in probability to diagonal matrices such that:*

i) $\operatorname{plim}\hat{\beta}_{ii} = \left(1+\psi_i\theta_i\right)^{-1/2}$ *when* $i \leq q$, *and* $\operatorname{plim}\hat{\beta}_{ii} = 0$ *when* $i > q$,

ii) $\operatorname{plim}\hat{\alpha}_{ii} = \left(1+\psi_i\omega_i\right)^{-1/2}$ *when* $i \leq q$, *and* $\operatorname{plim}\hat{\alpha}_{ii} = 0$ *when* $i > q$, *and*

iii) $\operatorname{plim}\left(\hat{L}'\hat{L}\right)_{ii} = \sigma^2 x_i$ *when* $i \leq q$, *and* $\operatorname{plim}\left(\hat{L}'\hat{L}\right)_{ii} = \sigma^2\bar{x}$ *when* $i > q$.

Recall that, by definition of the PC estimator, the columns of $T^{-1/2}\hat{F}$ are unit-length eigenvectors of $X'X$ and the columns of $\hat{\mathcal{L}}$ are unit-length eigenvectors of XX'. Since unit-length eigenvectors are only defined up to a multiplication by -1, the signs of the entries of matrices $\hat{\beta}$ and $\hat{\alpha}$ are not well-defined without further normalization constraints. In Theorem 1, we assume that the direction of the columns of $\hat{F}$ and $\hat{\mathcal{L}}$ is chosen so that $\hat{\beta}_{ii}$ and $\hat{\alpha}_{ii}$ are non-negative for all $i = 1, ..., k$.

[10] A matlab code with a numerical procedure finding $\bar{x}, \bar{u}, \bar{v}, x_i, u_i$ and v_i for any cdf's $\mathcal{G}_A(x)$ and $\mathcal{G}_B(x)$ is available from the author upon request.

Theorem 1 shows that, for $i \leq q$, the probability limits of $\hat{\beta}_{ii}$ and $\hat{\alpha}_{ii}$ are strictly less than 1, which implies the inconsistency of the PC estimators of the i-th factor and of the corresponding factor loadings. For $i > q$, not only the PC estimators are inconsistent, but they are asymptotically orthogonal to the true factor and loadings. By definition of q, such a situation occurs only when the explanatory power of the i-th factor is so weak that d_i is smaller than the threshold $\sigma^2 \bar{x} \left(1 - \bar{u}^{-1}\right) \left(1 - \bar{v}^{-1}\right)$. When the explanatory power of the factor increases so that d_i is above the threshold, the PC estimates start to be non-trivially correlated to the true factor and loadings. When d_i diverges to $+\infty$, $\psi_i \theta_i$ and $\psi_i \omega_i$ converge to zero, and the inconsistency of the PC estimates vanishes.

Let us illustrate the results of Theorem 1 using a simple example. Suppose data are generated by the following 1-factor model:

$$\begin{aligned} X_{it} &= \sqrt{dT} \mathcal{L}_{1i} \mathcal{F}_{1t} + e_{it}, \text{ where} \\ e_{it} &= \rho_1 e_{i-1,t} + \left(1 - \rho_1^2\right)^{1/2} \xi_{it} \text{ and} \\ \xi_{it} &= \rho_2 \xi_{it-1} + \left(1 - \rho_2^2\right)^{1/2} \eta_{it}, \ \eta_{it} \sim \text{iid } N(0,1) \end{aligned} . \tag{11}$$

Here $\sqrt{d}\mathcal{L}_{1i}$ are the loadings, $\sqrt{T}\mathcal{F}_{1t}$ are the values of the factor at time t, and the idiosyncratic terms e_{it} follow auto-regressions both temporally and cross-sectionally. Note that vec (e) is an $nT \times 1$ Gaussian vector with covariance matrix $T_2 \otimes T_1$, where T_1 and T_2 are Toeplitz matrices with i,j-th entries equal to $\rho_1^{|i-j|}$ and $\rho_2^{|i-j|}$. Therefore, e can be represented in the form $A\varepsilon B$, where $A = T_1^{1/2}$, $B = T_2^{1/2}$, and ε is an $n \times T$ matrix with i.i.d. $N(0,1)$ entries.

As is required by Assumption 2, we will assume that $\mathcal{L}_1$ and $\mathcal{F}_1$ are the eigenvectors of $T_1 = AA'$ and $T_2 = B'B$ corresponding to the unit eigenvalue. Clearly, this is a restrictive assumption since the eigenvectors of the Toeplitz matrices have very special form. In the Monte Carlo section of the paper, we check the robustness of our results to the situations when the loadings and the factors are not related to the eigenspaces of the matrices AA' and $B'B$.

As is well-known (see, for example, Grenander and Szego, 1958), the empirical distribution of the eigenvalues of symmetric Toeplitz matrices converges as the dimensionality of the matrix tends to infinity. For the special case of T_1, the inverse of the limiting cdf at $p \in (0,1)$ equals $\frac{1-\rho_1^2}{1+2\rho_1 \cos(p\pi)+\rho_1^2}$ so that the upper boundary of the support of the limiting distribution is finite and equals $\frac{1+\rho_1}{1-\rho_1}$, and the density of the

limiting distribution in the interior of its support is no smaller than $\frac{(1-\rho_1)^3}{2\pi\rho_1(1+\rho_1)} > 0$. Furthermore, the largest eigenvalue of T_1 converges to $\frac{1+\rho_1}{1-\rho_1}$ as the size of T_1 grows. For T_2, the above facts hold with ρ_2 replacing ρ_1. Hence, Assumption 3 holds for the data generating process described in (11).

Let us set $n/T = 2$. Further, let $\rho_1 = 0.5$ and $\rho_2 = 0.9$ so that there is a mild degree of the cross-sectional correlation and a high degree of the temporal correlation in the idiosyncratic terms. Setting ρ_2 so high will make our numerical example below sharper. A high degree of the auto-correlation of the residuals is sometimes observed in applications of the factor analysis. For example, this is the case for the data in Boivin et al. (2008) who use European quarterly macroeconomic time series to analyze the effect of Euro on monetary transmission mechanism. Even after extracting seven factors from that data, the residuals remain highly auto-correlated.

For the above setting of the parameters of (11), we have computed the probability limits of $\hat{\beta}$ and $\hat{L}'\hat{L}$ described in Theorem 1.[11] Figure 1 shows these probability limits as functions of $L'L$, which equals d in our example. From the left panel of the figure, we see that the PC estimator remains asymptotically orthogonal to the true factor until d becomes as large as 42. To interpret this finding note that d equals $nR^2/(1-R^2)$, where R^2 is the population R^2 of the factor, defined as $\frac{L'L}{\operatorname{tr} E(XX'/T)}$. Such an equality follows from the facts that $L'L = d$ and $\operatorname{tr} E(XX'/T) = L'L + n$, by Assumption 2. Hence, $d = 42$ corresponds to the factor's population R^2 equal to $\frac{42}{42+n}$, which is, approximately, 0.30 for $n = 100$.

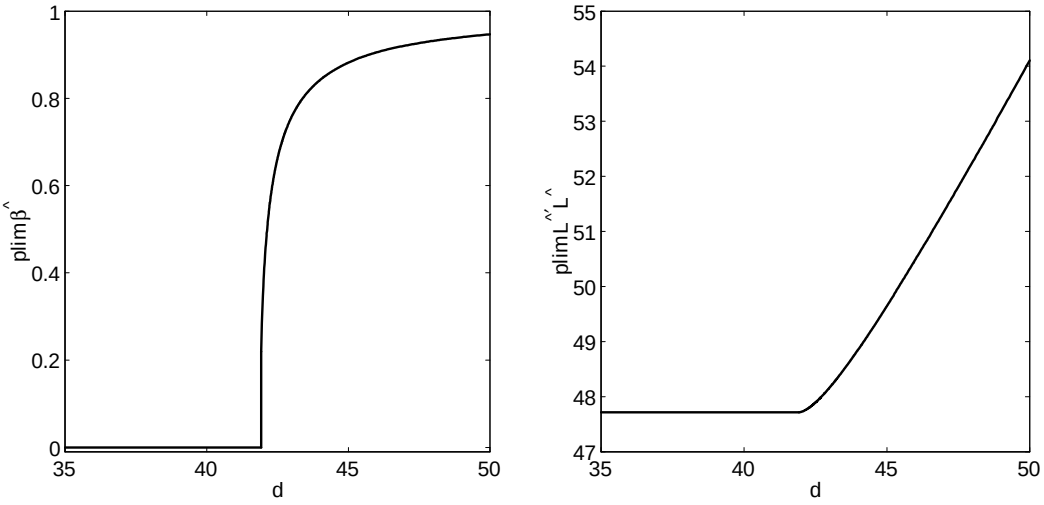
Furthermore, the right panel of Figure 1 reveals that the sample R^2 from fitting a single factor to the data may be very large even in cases when the factor is, in fact, very weak. Indeed, the sample R^2 can be approximated by the ratio $\frac{\hat{L}'\hat{L}}{\hat{L}'\hat{L}+n}$. Hence, in our example, even if there are no factors in the data at all so that $d = 0$, the sample R^2 from fitting one "factor" to the data would be around 0.32 for $n = 100$.

For smaller values of ρ_1 and ρ_2, the probability limits $\operatorname{plim}\hat{\beta}$ and $\operatorname{plim}\hat{L}'\hat{L}$ start increasing for smaller values of d. Hence, the problems of the PC estimator will be less extreme. In the special case when $\rho_1 = \rho_2 = 0$ so that there is no correlation in the idiosyncratic terms, $\operatorname{plim}\hat{\beta}$ remains zero only up to the threshold $d = 1.41$. For $d < 1.41$, $\operatorname{plim}\hat{L}'\hat{L}$ is only 5.83.

We now turn to our next result, which describes the asymptotic distribution of

[11] We do not illustrate our results on $\hat{\alpha}$ because they can be obtained from those on $\hat{\beta}$ by interchanging the cross-sectional and temporal parameters of the model.

Figure 1: The probability limits of $\hat{\beta}$ and $\hat{L}'\hat{L}$ as functions of d.



$\hat{\beta}$, $\hat{\alpha}$ and $\hat{L}'\hat{L}$ around their probability limits. Let us make the following additional assumption.

Assumption 4: *As* $n \to \infty$, $n/T - c = o\left(n^{-1/2}\right)$, $L'L - D = o\left(n^{-1/2}\right)$, $\sup_{x \in R} |G_A(x) - \mathcal{G}_A(x)| = o\left(n^{-1/2}\right)$ *and* $\sup_{x \in R} |G_B(x) - \mathcal{G}_B(x)| = o\left(n^{-1/2}\right)$.

By requiring that the convergence of n/T, $L'L$, $G_A(x)$ and $G_B(x)$ to the limits introduced in Assumptions 1 and 3 is fast, Assumption 4 eliminates any possible effects of this convergence on our asymptotic results. As we mentioned above, we view our asymptotics as a device for obtaining accurate approximations to finite sample distributions rather than as a description of the process of further sampling. From this perspective, consequential assumptions about the rates of convergence of n/T, $L'L$, $G_A(x)$ and $G_B(x)$ would be desirable only to the extent that they generate some known distributional properties of statistics related to the PC estimator. We do not see how this is possible, and therefore, make the simplest possible assumption that the rates of the convergence of n/T, $L'L$, $G_A(x)$ and $G_B(x)$ are fast enough not to interfere with our asymptotic analysis.

To formulate Theorem 2, we need to introduce new notation. First, denote the matrices of the first q columns of $\hat{\beta}$, $\hat{\alpha}$ and $\hat{L}$ as $\hat{\beta}_{1:q}$, $\hat{\alpha}_{1:q}$ and $\hat{L}_{1:q}$, respectively, where q is as defined in (9) and (10). Further, let $m_i(r) = \int (x_i - \lambda)^{-r} d\mathcal{G}(\lambda)$, and $m_{ij}(1,1) = \int (x_i - \lambda)^{-1} (x_j - \lambda)^{-1} d\mathcal{G}(\lambda)$, where x_i with $i = 1, ..., q$ are as in Theorem 1 and $\mathcal{G}$ is the limiting distribution of the eigenvalues of $XX'/(\sigma^2 T)$. Let

$\tilde{\mathcal{G}}$ be the limiting distribution of the eigenvalues of $X'X/(\sigma^2 T)$ (it differs from $\mathcal{G}$ only by the probability it assigns to zero). Finally, let $\tilde{m}_i(r) = \int (x_i - \lambda)^{-r} d\tilde{\mathcal{G}}(\lambda)$, and $\tilde{m}_{ij}(1,1) = \int (x_i - \lambda)^{-1} (x_j - \lambda)^{-1} d\tilde{\mathcal{G}}(\lambda)$.

Theorem 2: *Let Assumptions 1, 2, 3 and 4 hold. Then, as $n \to \infty$:*

i) *If A is unconstrained other than by Assumptions 1-4, but $B = I_T$:*

$$\operatorname{diag}\sqrt{T}\left(\hat{L}'_{1:q}\hat{L}_{1:q} - \operatorname{plim}\hat{L}'_{1:q}\hat{L}_{1:q}\right) \xrightarrow{d} N(0,\Omega) \text{ and}$$
$$\operatorname{vec}\sqrt{T}\left(\hat{\beta}_{1:q} - \operatorname{plim}\hat{\beta}_{1:q}\right) \xrightarrow{d} N(0,\Sigma_\beta),$$

where Ω is diagonal with $\Omega_{jj} = 2(d_j + \sigma^2)^2 \frac{\tilde{m}_j^2(1)}{\tilde{m}_j(2)} - 2d_j^2\left(\frac{\tilde{m}_j^2(1)}{\tilde{m}_j(2)}\right)^2$; $\operatorname{plim}\hat{L}'_{1:q}\hat{L}_{1:q}$ is diagonal with $\operatorname{plim}\hat{L}'_j\hat{L}_j = \sigma^2 x_j$, where $x = x_j$ is the largest solution of the equation $\left(\frac{d_j}{\sigma^2} + 1\right)\int (x-\lambda)^{-1} d\tilde{\mathcal{G}}(\lambda) = 1$; $\operatorname{plim}\hat{\beta}_{1:q}$ has all elements zero except $\operatorname{plim}\hat{\beta}_{jj} = \sqrt{\frac{d_j}{d_j+\sigma^2}\frac{\tilde{m}_j^2(1)}{\tilde{m}_j(2)}}$, where $j = 1, ..., q$; finally, Σ_β is such that the asymptotic covariance between $\sqrt{T}\left(\hat{\beta}_{ij} - \operatorname{plim}\hat{\beta}_{ij}\right)$ and $\sqrt{T}\left(\hat{\beta}_{st} - \operatorname{plim}\hat{\beta}_{st}\right)$ equals:

- $\frac{d_j^2+d_i\sigma^2}{(d_i-d_j)^2} - \frac{d_j^3}{(d_i-d_j)^2(d_j+\sigma^2)}\frac{\tilde{m}_j^2(1)}{\tilde{m}_j(2)}$ *when* $i = s \neq j = t$;
- $\frac{(d_i d_j)^{3/2}}{(d_i-d_j)^2\sqrt{(d_i+\sigma^2)(d_j+\sigma^2)}}\frac{\tilde{m}_j(1)\tilde{m}_i(1)}{\sqrt{\tilde{m}_j(2)\tilde{m}_i(2)}} - \frac{\sqrt{d_i d_j(d_i+\sigma^2)(d_j+\sigma^2)}}{(d_i-d_j)^2}\frac{\tilde{m}_{ij}(1,1)}{\sqrt{\tilde{m}_j(2)\tilde{m}_i(2)}}$ *when* $i = t \neq j = s$;
- $\frac{\sigma^2}{d_i+\sigma^2} + \frac{d_i}{2(d_i+\sigma^2)}\frac{\tilde{m}_i^2(1)\tilde{m}_i(4)}{\tilde{m}_i^3(2)} - \frac{2d_i}{(d_i+\sigma^2)^3}\frac{\tilde{m}_i^2(1)}{\tilde{m}_i(2)}\left(d_i\frac{\tilde{m}_i(1)\tilde{m}_i(3)}{\tilde{m}_i^2(2)} - \frac{d_i}{2} + \sigma^2\right)^2$ *when* $i = j = s = t$;
- *and zero otherwise.*

ii) *If B is unconstrained other than by Assumptions 1-4, but $A = I_n$:*

$$\operatorname{diag}\sqrt{n}\left(\hat{L}'_{1:q}\hat{L}_{1:q} - \operatorname{plim}\hat{L}'_{1:q}\hat{L}_{1:q}\right) \xrightarrow{d} N(0,\Omega) \text{ and}$$
$$\operatorname{vec}\sqrt{n}\left(\hat{\alpha}_{1:q} - \operatorname{plim}\hat{\alpha}_{1:q}\right) \xrightarrow{d} N(0,\Sigma_\alpha),$$

where Ω is diagonal with $\Omega_{jj} = 2(d_j + \sigma^2 c)^2 \frac{m_j^2(1)}{m_j(2)} - 2d_j^2\left(\frac{m_j^2(1)}{m_j(2)}\right)^2$; $\operatorname{plim}\hat{L}'_{1:q}\hat{L}_{1:q}$ is diagonal with $\operatorname{plim}\hat{L}'_j\hat{L}_j = \sigma^2 x_j$, where $x = x_j$ is the largest solution of the equation $\left(\frac{d_i}{\sigma^2} + c\right)\int (x-\lambda)^{-1} d\mathcal{G}(\lambda) = 1$; $\operatorname{plim}\hat{\alpha}_{1:q}$ has all elements zero

except $\operatorname{plim}\hat{\alpha}_{jj} = \sqrt{\frac{d_i}{d_i+c\sigma^2}\frac{m_i^2(1)}{m_i(2)}}$, *where* $j = 1, ..., q$; *finally,* Σ_α *is such that the asymptotic covariance between* $\sqrt{n}\left(\hat{\alpha}_{ij} - \operatorname{plim}\hat{\alpha}_{ij}\right)$ *and* $\sqrt{n}\left(\hat{\alpha}_{st} - \operatorname{plim}\hat{\alpha}_{st}\right)$ *equals*:

- $\frac{d_j^2+c\sigma^2 d_i}{(d_i-d_j)^2} - \frac{d_j^3}{(d_i-d_j)^2(d_j+c\sigma^2)}\frac{m_j^2(1)}{m_j(2)}$ *when* $i = s \neq j = t$;
- $\frac{(d_i d_j)^{3/2}}{(d_i-d_j)^2\sqrt{(d_i+c\sigma^2)(d_j+c\sigma^2)}}\frac{m_j(1)m_i(1)}{\sqrt{m_j(2)m_i(2)}} - \frac{\sqrt{d_i d_j(d_i+c\sigma^2)(d_j+c\sigma^2)}}{(d_i-d_j)^2}\frac{m_{ij}(1,1)}{\sqrt{m_j(2)m_i(2)}}$ *when* $i = t \neq j = s$;
- $\frac{c\sigma^2}{d_i+c\sigma^2} + \frac{d_i}{2(d_i+c\sigma^2)}\frac{m_i^2(1)m_i(4)}{m_i^3(2)} - \frac{2d_i}{(d_i+c\sigma^2)^3}\frac{m_i^2(1)}{m_i(2)}\left(d_i\frac{m_i(1)m_i(3)}{m_i^2(2)} - \frac{d_i}{2} + c\sigma^2\right)^2$ *when* $i = j = s = t$;
- *and zero otherwise.*

The assumption $B = I_T$, made in part i) of Theorem 2, corresponds to the case when the idiosyncratic terms are not temporally correlated but may be correlated cross-sectionally. Although we were able to establish the asymptotic distributions of $\hat{\beta}_{1:q}$ and of $\hat{L}'_{1:q}\hat{L}_{1:q}$, we do not know how to obtain the asymptotic distribution of $\hat{\alpha}_{1:q}$ in such a case. Similarly, for the opposite case when $A = I_n$, considered in part ii), we were able to establish only the asymptotic distributions of $\hat{\alpha}_{1:q}$ and $\hat{L}'_{1:q}\hat{L}_{1:q}$, but not that of $\hat{\beta}_{1:q}$. In the very special case, when both A and B are identity matrices, so that neither cross-sectional nor temporal correlation is present in the idiosyncratic terms, the asymptotic distributions of $\hat{\beta}_{1:q}$ and $\hat{\alpha}_{1:q}$ are described simultaneously by formulae in parts i) and ii),[12] and the asymptotic distribution for $\hat{L}'_{1:q}\hat{L}_{1:q}$ provided by part i) equals that provided by part ii) after scaling by $\lim\sqrt{T/n} = c^{-1/2}$.

Note that part ii) of Theorem 2 can be obtained from part i) by interchanging the cross-sectional and temporal parameters of the model. For formulae describing Σ_α, this amounts to replacing d_i and $\tilde{m}_r$ in the corresponding formulae for Σ_β by d_i/c and m_r.

Assuming that either A or B is an identity matrix is restrictive. However, as we have explained in the Introduction, such an assumption is acceptable in some applications. In Section 5, we make the assumption that $B = I_T$ and use results of part i) of Theorem 2 to test a hypothesis that the celebrated Fama-French factors (see Fama and French, 1993) span the factor space of the US excess stock return data.

[12] In such a special case, the formulas of Theorem 2 simplify. A previous version of the paper, which is available from the author's webpage, reports such a simplified formulas.

Generalizing Theorem 2 to the case when both A and B are not identity requires conceptual changes in our proofs. Therefore, we leave this important topic for future research.

Results of Theorem 2 can be used to obtain the asymptotic distributions of the principal components estimator of factors at particular time periods and of factor loadings corresponding to specific cross-sectional units. We will do such an extension elsewhere, and will keep our focus on the behavior of $\hat{\beta}$ and $\hat{\alpha}$ in what follows.

3 Identification and estimation

The asymptotic distributions obtained in the previous section quantify the potential problems with the PC estimator when factors are weakly influential. However, if the distributions are to be used for inference, their parameters have to be estimated. This section explains how to obtain such an estimates when the true number of factors is bounded above by a known fixed number $k_{\max}$.

The crucial parameter in Theorems 1 and 2 is q, which is the number of factors that are not orthogonal to their PC estimators asymptotically. Under our weakly influential factor asymptotics, q can be identified as the number of eigenvalues out of the $k_{\max}$ largest eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{k_{\max}}$ of XX'/T that remain isolated asymptotically as n and T go to infinity. Indeed, under Assumptions 1, 2 and 3, the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_q$ converge to distinct limits $\sigma^2 x_1, ..., \sigma^2 x_q$, respectively, whereas the eigenvalues $\lambda_{q+1}, ..., \lambda_{k_{\max}}$ all converge to the same limit $\sigma^2 \bar{x}$.

The above identification scheme implies that q can be consistently estimated by Onatski's (2009) estimator of the number of factors:

$$\hat{q}(\delta) = \max\{0 \leq i \leq k_{\max} : \lambda_i - \lambda_{i+1} \geq \delta\},$$

where δ is a calibrated scaling parameter, and λ_0 is defined as $+\infty$. The estimator $\hat{q}(\delta)$ is consistent for q as long as $\sigma^2 (x_q - \bar{x}) > \delta$. Note that, under Assumptions 1,2 and 3, the calibration algorithm[13] that Onatski (2009) recommends using in practice results in δ that converges to zero as n and T go to infinity. Hence, if such a calibration

[13] The algorithm is based on an insight from the large random matrix theory, and is as follows. First, set $j = k_{\max} + 1$. Then, iterate the following steps until convergence: 1) set $\delta = 2|\hat{\gamma}|$, where $\hat{\gamma}$ is the OLS estimate of the slope in the regression of $\lambda_j, \lambda_{j+1}, ..., \lambda_{j+4}$ on the constant and $(j-1)^{2/3}, ..., (j+3)^{2/3}$, 2) set $j = \hat{q}(\delta) + 1$.

is used, the inequality $\sigma^2(x_q - \bar{x}) > \delta$ is satisfied asymptotically.

It is instructive to compare the above identification of q with the identification of the number of factors k under standard asymptotics (3). In the standard case, the identification of k is based on the fact that the eigenvalues $\lambda_1, ..., \lambda_k$ diverge to infinity while λ_{k+1} remains bounded. Under our alternative asymptotics, none of the eigenvalues diverge to infinity. However, the first q eigenvalues remain isolated and separated from the other eigenvalues, which cluster together. The proposed estimator of q counts the number of the isolated eigenvalues which are at least at distance δ from the cluster.

The true number of factors k is not identified under our asymptotics in the sense that there is no procedure allowing us to consistently estimate k from the sequence of the finite samples $X^{(n)}$ satisfying Assumptions 1,2,3 and 4. In a separate research project (see Onatski et al., 2011) we show that in the case when $k > q$, although there exist statistical tests that have non-trivial power of rejecting the nulls that the true number of factors is no larger than q, the power of such tests does never approach one. Had a consistent procedure for determining k existed, a test that would reject the above nulls whenever the estimate of the true number of factors is larger than q would have an asymptotic power of 1, contradicting our results.

From practical perspective, identifying q may be more important than identifying k. Indeed, in the case when $k > q$, the eigenvalues $\lambda_{q+1}, ..., \lambda_k$ do not separate from the cluster because the corresponding factors do not have sufficient explanatory power. Moreover, as Theorem 1 shows, such a weak factors will be orthogonal to their PC estimators asymptotically. Hence, as long as the principal components estimator is used, q may be treated as an "effective" number of factors in the data.

In principle, the standard identification scheme based on "exploding eigenvalues" can coexist with the identification based on the "asymptotically isolated eigenvalues".[14] In particular, it is possible to consider a generalization of (1):

$$X_{it} = \sum_{j=1}^{k_1} L_{ij} F_{tj} + \sum_{j=1}^{k_2} \tilde{L}_{ij} \tilde{F}_{tj} + e_{it}, \tag{12}$$

where $k_1 + k_2 = q$, the first k_1 of the largest eigenvalues of XX'/T explode as $n \to \infty$, and the next k_2 of the largest eigenvalues remain isolated from the cluster of smaller

[14] I am grateful to an anonymous referee for pointing out this possibility.

eigenvalues. In such a model, F_{tj} and $\tilde{F}_{tj}$ would be interpreted as strongly and weakly influential factors, respectively. Further, the convergence $\tilde{L}'\tilde{L} \to \tilde{D}$, where $\tilde{L}$ is the matrix of the loadings of the weakly influential factors, would be interpreted directly as defining a non-standard concept of "weakly influential factors" rather than as describing a non-standard Pitman-drift-like asymptotic device. The identification of the "weakly influential factors" themselves as opposed to merely the identification of their number k_2 would require an extra assumption such as our Assumption 2. Finding a form of such an identification assumption that would be convenient without being overly restrictive and reinterpreting the results of this paper in the context of model (12) are interesting exercises left for future research.

Once a consistent estimator $\hat{q}$ of q is obtained, it is relatively easy to estimate the parameters of the distributions obtained in the previous section. Theorem 3 below describes consistent estimates of the probability limits and of the asymptotic variances from Theorem 2. We will use the following new notation and definitions. Let λ_i be the i-th largest eigenvalue of XX'/T. In cases when $T > n$, let us define $\lambda_i = 0$ for $n < i \leq T$. Further, let $\hat{c} = n/T$, $\hat{\sigma}^2 = \sum_{j=\hat{q}+1}^{T} \lambda_j / (n - \hat{q})$,and, for any non-negative integer r and any non-negative integer $i \leq q$, let $\widehat{\tilde{m}}_i(r) = \frac{\hat{\sigma}^{2r}}{T-\hat{q}} \sum_{j=\hat{q}+1}^{T} (\lambda_i - \lambda_j)^{-r}$, $\widehat{\tilde{m}}_{is}(1,1) = \frac{\hat{\sigma}^4}{T-\hat{q}} \sum_{j=\hat{q}+1}^{T} (\lambda_i - \lambda_j)^{-1} (\lambda_s - \lambda_j)^{-1}$, $\hat{m}_i(r) = \frac{\hat{\sigma}^{2r}}{n-\hat{q}} \sum_{j=\hat{q}+1}^{n} (\lambda_i - \lambda_j)^{-r}$ and $\hat{m}_{is}(1,1) = \frac{\hat{\sigma}^4}{n-\hat{q}} \sum_{j=\hat{q}+1}^{n} (\lambda_i - \lambda_j)^{-1} (\lambda_s - \lambda_j)^{-1}$.

Theorem 3. *Suppose that Assumptions 1, 2, 3 and 4 hold, and let $\hat{q}$ be a consistent estimator of q. Then, for any $i \leq q$, $j \leq q$, $s \leq q$ and $t \leq q$:*

i) *If $B = I_T$, we have: d_j is consistently estimated by $\hat{d}_j = \hat{\sigma}^2 \left(1/\widehat{\tilde{m}}_j(1) - 1\right)$, $\operatorname{plim} \hat{\beta}_{jj}$ is consistently estimated by $\sqrt{\frac{\hat{d}_j}{\hat{d}_j + \hat{\sigma}^2} \frac{\widehat{\tilde{m}}_j^2(1)}{\widehat{\tilde{m}}_j(2)}}$, and consistent estimators of Ω_{jj} and of the asymptotic covariances between $\sqrt{T}\left(\hat{\beta}_{ij} - \operatorname{plim} \hat{\beta}_{ij}\right)$ and $\sqrt{T}\left(\hat{\beta}_{st} - \operatorname{plim} \hat{\beta}_{st}\right)$ are obtained by replacing parameters σ^2, d_i and $\tilde{m}_i(r)$ in the formulae of Theorem 2 by $\hat{\sigma}^2$, $\widehat{\tilde{d}}_i$ and $\widehat{\tilde{m}}_i(r)$.*

ii) *If $A = I_n$, we have: d_i is consistently estimated by $\hat{d}_i = \hat{\sigma}^2 (1/\hat{m}_i(1) - \hat{c})$, $\operatorname{plim} \hat{\alpha}_{ii}$ is consistently estimated by $\sqrt{\frac{\hat{d}_i}{\hat{d}_i + \hat{c}\hat{\sigma}^2} \frac{\hat{m}_i^2(1)}{\hat{m}_i(2)}}$, and consistent estimators of Ω_{ii} and of the asymptotic covariances between $\sqrt{n}(\hat{\alpha}_{ij} - \operatorname{plim} \hat{\alpha}_{ij})$ and $\sqrt{n}(\hat{\alpha}_{st} - \operatorname{plim} \hat{\alpha}_{st})$ are obtained by replacing parameters c, σ^2, d_i and $m_i(r)$ in the formulae of Theorem 2 by their estimates $\hat{c}$, $\hat{\sigma}^2$, $\hat{d}_i$ and $\hat{m}_i(r)$.*

The full proof of Theorem 3 as well as of all other results of this paper is given in the Technical Appendix. The key consistency result, which, together with the continuous mapping theorem, implies the validity of Theorem 3, is the consistency of $\widehat{\widetilde{m}}_i(r)$ and $\hat{m}_i(r)$ for $\tilde{m}_i(r)$ and $m_i(r)$, respectively. Such a consistency follows from Theorem 1 iii) and from the weak convergence of the empirical distribution of $\lambda_{\hat{q}+1}/\sigma^2, ..., \lambda_n/\sigma^2$ to $\mathcal{G}$, which, in turn, easily follows from the results of Zhang (2006) that were discussed above.

The quantity $\widehat{\widetilde{m}}_i^2(1)/\widehat{\widetilde{m}}_i(2)$, which appears in our estimate of $\operatorname{plim}\hat{\beta}_{ii}$, has an interesting interpretation[15] of a particular measure of dispersion of the "idiosyncratic" eigenvalues $\lambda_{\hat{q}+1}, \lambda_{\hat{q}}, ..., \lambda_T$ of $X'X/T$. Jensen's inequality implies that $\widehat{\widetilde{m}}_i^2(1)/\widehat{\widetilde{m}}_i(2)$ is smaller than 1 as long as not all of $\lambda_{\hat{q}+1}, \lambda_{\hat{q}}, ..., \lambda_T$ are equal to each other. Note that the smaller the $\widehat{\widetilde{m}}_i^2(1)/\widehat{\widetilde{m}}_i(2)$, the larger the estimated asymptotic bias of the PC estimator of the i-th factor. The link between $\widehat{\widetilde{m}}_i^2(1)/\widehat{\widetilde{m}}_i(2)$ and the empirical distribution of $\lambda_{\hat{q}+1}, \lambda_{\hat{q}}, ..., \lambda_T$ can potentially be further exploited to obtain a simple procedure for assessing the quality of the PC estimates based on the visual inspection of the scree plot introduced by Cattell (1966). We leave such a development for future research.

4 A Monte Carlo study

In this section, we use Monte Carlo (MC) experiments to study the finite sample approximation quality of the asymptotic results derived in Theorems 1, 2 and 3. In all our experiments, the idiosyncratic terms follow auto-regressions both temporally and cross-sectionally as described in (11). Hence, $e = A\varepsilon B$, where AA' and $B'B$ equal the Toeplitz matrices with i,j-th elements $\rho_1^{|i-j|}$ and $\rho_2^{|i-j|}$, respectively, and $\sigma^2 \equiv \operatorname{Var}(\varepsilon_{ij}) = 1$.

We consider models with three factors, and always normalize factors F and loadings L so that $\frac{1}{T}F'F = I_3$ and $L'L$ equals a diagonal matrix with diagonal elements $d_1 = 30$, $d_2 = 20$ and d_3 varying on a grid in the interval $[0, 15]$. Such a normalization implies that, for the cross-section size n, the proportion of the variance of the data explained by the first, second and third factors equals $\frac{30}{30+20+d_3+n}$, $\frac{20}{30+20+d_3+n}$ and $\frac{d_3}{30+20+d_3+n}$, respectively.

In our first experiment, we set $\rho_1 = \rho_2 = 0.5$ so that there is a moderate amount of

[15] The quantity $\hat{m}_i^2(1)/\hat{m}_i(2)$ has a similar interpretation.

the cross-sectional and temporal correlation in the idiosyncratic terms. Similar to the setting of our example (11), we make the factors and loadings equal to eigenvectors of $B'B$ and of AA' corresponding to the three eigenvalues that are the closest to unity. We relax this setting in our further MC experiments.

The left panel of Figure 2 compares MC means, based on 5,000 MC replications, of the diagonal elements $\hat{\beta}_{ii}$ of the 3×3 matrix $\hat{\beta} = (F'F)^{-1} F'\hat{F}$ with the corresponding probability limits derived in Theorem 1.[16] The MC means and the probability limits are plotted against d_3, which is a measure of the explanatory power of the third factor. The dashed lines correspond to the probability limits, the solid lines correspond to the MC means, and the two dotted lines correspond to the 5-th and the 95-th quantiles of the MC distribution of $\hat{\beta}_{33}$. The upper, middle and lower panels correspond to sample sizes $(n, T) = (50, 25)$, $(100, 50)$ and $(200, 100)$, respectively.

As d_3 decreases from 15 to zero, $\operatorname{plim} \hat{\beta}_{33}$ decreases from a value close to one to zero. Such a decrease is very abrupt around $d_3 = 7.5$. In contrast, $\operatorname{plim} \hat{\beta}_{ii}$ with $i = 1, 2$ remain constant at values close to one. It is because the first and the second factors remain strongly influential, d_1 and d_2 being fixed at a relatively high level.

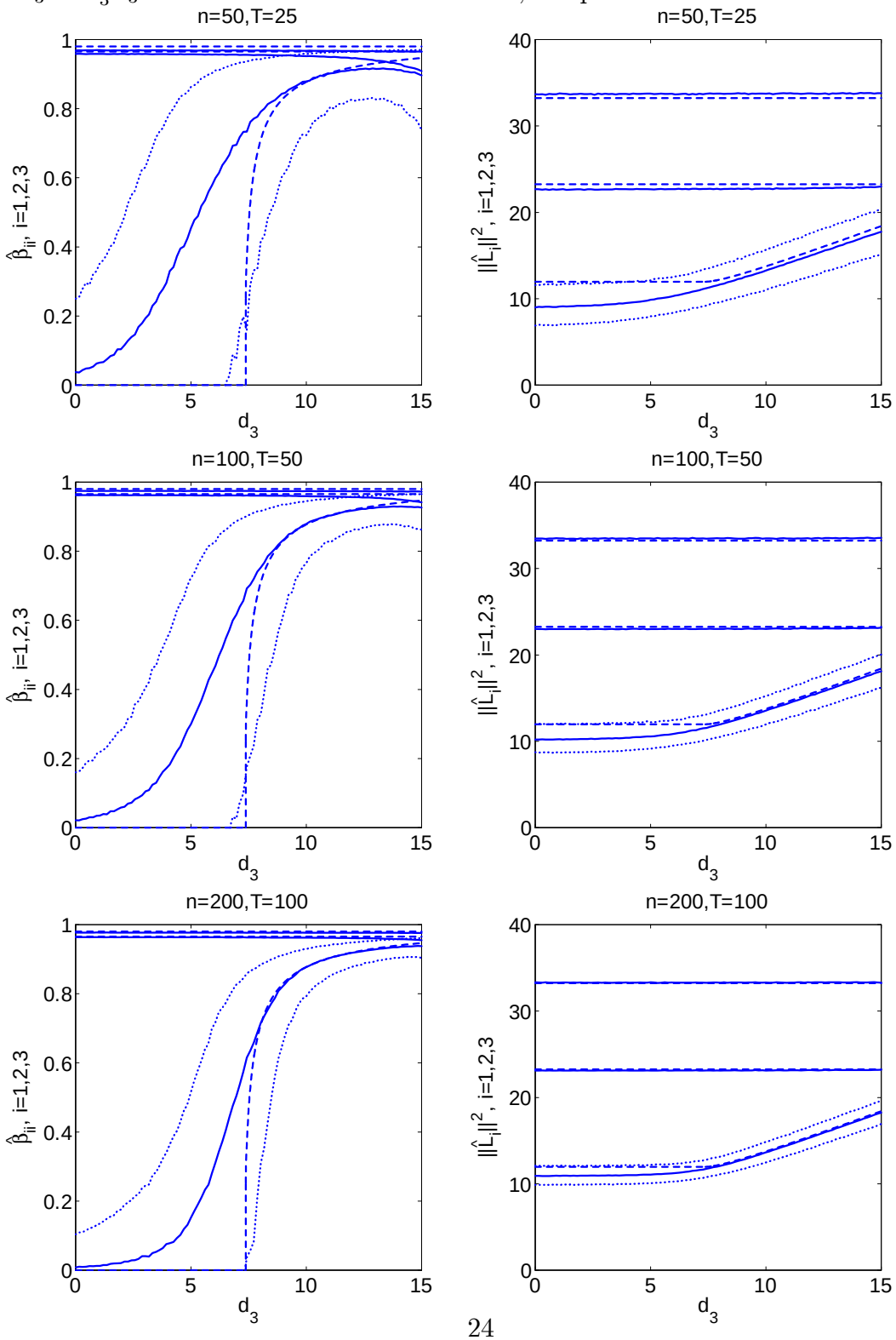
The MC means of $\hat{\beta}_{ii}$ approach the corresponding probability limits as n and T increase. The MC means of $\hat{\beta}_{ii}$ with $i = 1, 2$ are very well approximated by $\operatorname{plim} \hat{\beta}_{ii}$ for all values of d_3 on our grid. The MC mean of $\hat{\beta}_{33}$ is well approximated by $\operatorname{plim} \hat{\beta}_{33}$ for $d_3 > 7.5$. The quality of the approximation is especially good for $(n, T) = (200, 100)$. However, as d_3 decreases from values above 7.5 to values below 7.5, the MC mean of $\hat{\beta}_{33}$ declines much less abruptly than $\operatorname{plim} \hat{\beta}_{33}$.

The right panel of Figure 2 shows the probability limits and the MC means of $\hat{L}_i'\hat{L}_i$ with $i = 1, 2, 3$ as functions of d_3. The quality of the asymptotic approximations is good for all values of d_3. It is especially good for $d_3 > 7.5$.

We repeated the experiment for different values of ρ_1 and ρ_2. Qualitatively, the results remain the same. However, the MC means are increasingly better approximated by the probability limits from Theorem 1 as ρ_1 and ρ_2 decline. Vice versa, as ρ_1 and ρ_2 rise, the quality of the approximation deteriorates. Further, we again set $\rho_1 = \rho_2 = 0.5$, but consider a relatively fat-tailed and a skewed distribution for ε_{it}. Precisely, we consider Student's t distribution with five degrees of freedom, normalized to have unit variance, and the centered chi-squared distribution with one degree

[16] Although not shown on the graphs, the MC means of all the non-diagonal elements $\hat{\beta}_{ij}, i \neq j$ of matrix $\hat{\beta}$ are very close to zero, which is the theoretical probability limit.

Figure 2: The correlation coefficients $\hat{\beta}_{ii}$ between true and estimated factors (left panel) and the estimated explanatory power of factors $\hat{L}_i'\hat{L}_i$ (right panel) as functions of $d_3 \equiv L_3'L_3$. Theoretical values: dashed line, sample means: solid line.



of freedom, normalized to have unit variance. For such a non-normal distributions, the results are very similar to those shown in Figure 2, and we do not report them to save space.

Our second MC experiment assesses the impact of violations of parts ii) and iii) of Assumption 2 on the quality of the asymptotic approximations derived in Theorem 1. Recall that assumptions 2 ii) and 2 iii) require that the columns of L and F are eigenvectors of AA' and $B'B$ corresponding to unit eigenvalues, which is very restrictive. We now make L and F unrelated to the eigenvectors of AA' and $B'B$ by generating L and F as $n \times 3$ and $T \times 3$ matrices with i.i.d. standard Gaussian entries.[17] Figure 3 summarizes MC results for such a new setting.

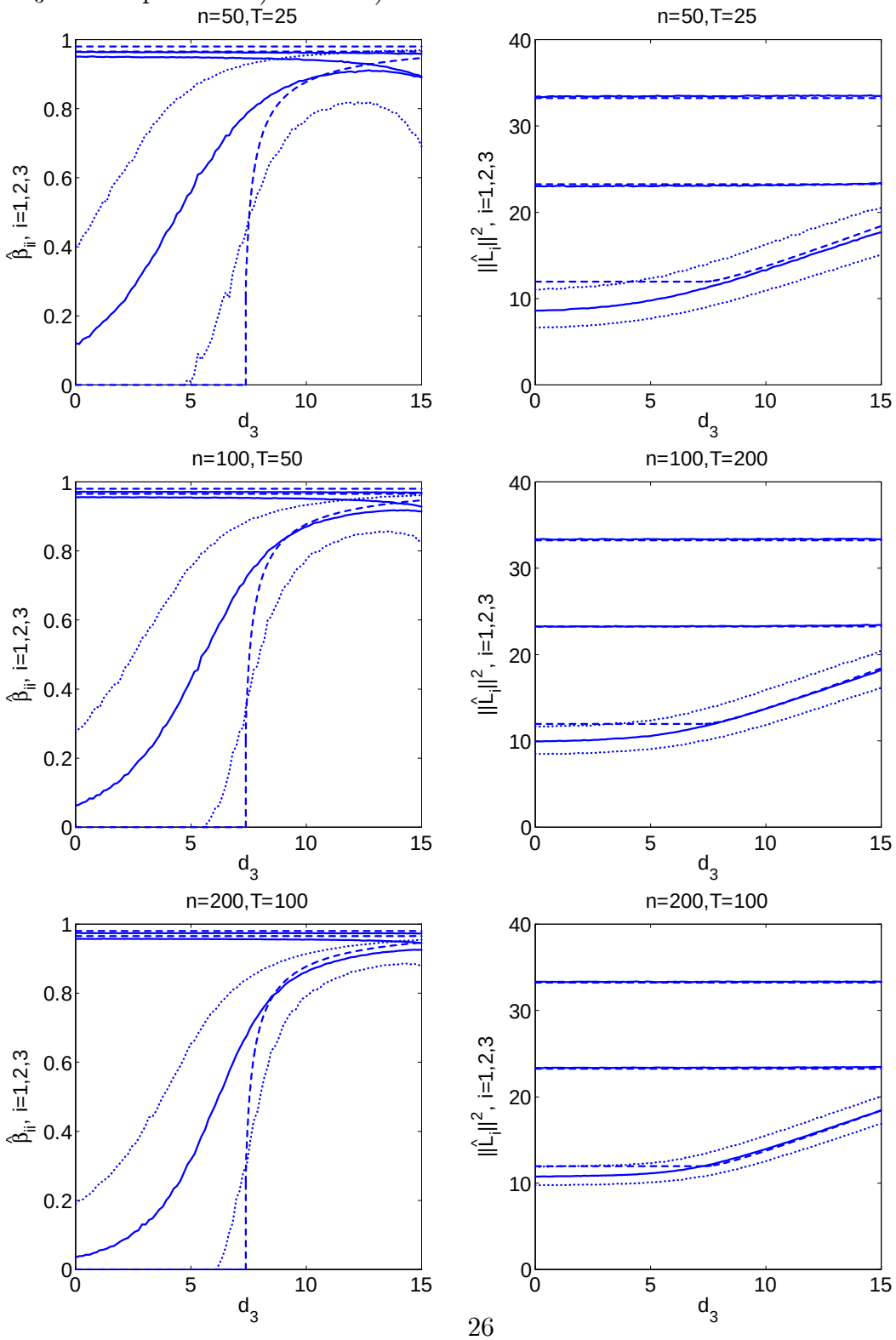
Comparing Figures 2 and 3, we see that the violation of parts ii) and iii) of Assumption 2 leads to a deterioration in the quality of the asymptotic approximations. The deterioration is more noticeable for $\hat{\beta}_{ii}$ than for $\hat{L}_i'\hat{L}_i$. As n and T rise, the solid and dashed lines on the left panel of Figure 3 become closer, but such a convergence appears to be slower than the one that can be seen on Figure 2.

Intuitively, when factors' projection on the eigenspaces of $B'B$ corresponding to relatively large eigenvalues increase, the factors and the principal eigenvectors of the covariance matrix of the idiosyncratic terms become more collinear. As a result, the principal component estimators of the factors become less "contaminated" by the noisy directions corresponding to the idiosyncratic influences. Vice versa, when factors' become more collinear with those eigenvectors of $B'B$ that correspond to relatively small eigenvalues, the "contamination" effect becomes more pronounced. These effects translate into relatively higher or relatively lower values of $\hat{\beta}_{ii}$, which contributes to the differences between Figures 2 and 3.

The above intuition can be checked as follows. Note that a simple indicator that contains information about the relationship between F_3 and the eigenstructure of $B'B$ is $\frac{1}{T}F_3'B'BF_3$. Had assumptions 2 ii) and 2 iii) been respected, $\frac{1}{T}F_3'B'BF_3$ would have been equal to one. When $\frac{1}{T}F_3'B'BF_3$ is larger or smaller than one, the projection of F_3 on the eigenvectors of $B'B$ corresponding to relatively large eigenvalues increases or decreases, respectively. Therefore, if our intuition is correct, the value of $\frac{1}{T}F_3'B'BF_3$ should be positively correlated with that of $\hat{\beta}_{33}$. To check this, we run a regression of

[17] We then normalize the matrices of factors and loadings by redefining them as $F\left(\frac{1}{T}F'F\right)^{-1/2}$ and $L\left(L'L\right)^{-1/2}D^{1/2}$, where $D = \text{diag}\left(d_1, d_2, d_3\right)$. Hence, the normalized factors satisfy $\frac{1}{T}F'F = I_3$, and the normalized loadings satisfy $L_i'L_j = 0$ for $i \neq j$ and $L_i'L_i = d_i$.

Figure 3: The correlation coefficients $\hat{\beta}_{ii}$ between true and estimated factors (left panel) and the estimated explanatory power of factors $\hat{L}_i'\hat{L}_i$ (right panel) as functions of d_3. Assumptions 2 ii) and 2 iii) are violated.



5,000 MC replications of $\hat{\beta}_{33}$ on constant and the corresponding values of $\frac{1}{T}F_3'B'BF_3$. The estimated slope coefficient and the corresponding 95% confidence band are shown on the left panel of Figure 4 as functions of d_3. Confirming our intuition, the slope of the estimated regression is significant and positive for the majority of values of d_3 in the studied range. The value of the slope is relatively larger for those d_3 which correspond to relatively high variation in the dependent variable $\hat{\beta}_{33}$.[18]

The right panel of Figure 4 contains more detailed information about the effect of $\frac{1}{T}F_3'B'BF_3$ on $\hat{\beta}_{33}$ when $d_3 = 5$. The figure reports estimates of the densities of the conditional distribution of $\hat{\beta}_{33}$ given relatively small and relatively large values of $\frac{1}{T}F_3'B'BF_3$. Precisely, we sort the MC replications according to the value of $\frac{1}{T}F_3'B'BF_3$ starting from the lowest value and ending with the highest value. Then we use MATLAB's 'ksdensity' code to get the kernel density estimates for $\hat{\beta}_{33}$ using only the first third (dashed line) and only the last third (solid line) of the so sorted MC sample. In accordance with our intuition, the estimated conditional distribution of $\hat{\beta}_{33}$ given relatively small values of $\frac{1}{T}F_3'B'BF_3$ puts more mass on relatively smaller values of $\hat{\beta}_{33}$.

Note that when F_3 is generated randomly, as in the above experiment, and normalized so that $\frac{1}{T}F_3'F_3 = 1$, the value of $\frac{1}{T}F_3'B'BF_3$ is approximately symmetrically distributed around one. Therefore, the unconditional distribution of $\hat{\beta}_{33}$ is a "well-balanced" mixture of the conditional distributions corresponding to relatively small and relatively large values of $\frac{1}{T}F_3'B'BF_3$. As a result, the locations of the unconditional distribution of $\hat{\beta}_{33}$ on Figures 2 and 3 are only moderately different.

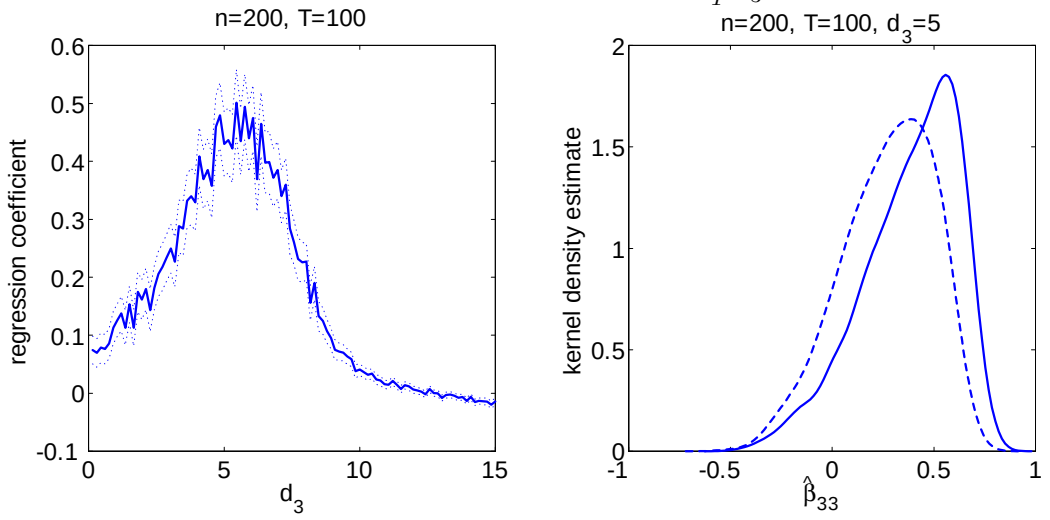
We can generate larger differences by using factors that are systematically related to the eigenstructure of $B'B$. For example, if we generate F_3 as T observations of an auto-regression with positive auto-regressive coefficient ρ_F, then, for our choice of B, the values of $\frac{1}{T}F_3'B'BF_3$ would tend to be larger than one, and we would expect $\hat{\beta}_{33}$ shifting towards larger values. Similarly, for negative values of ρ_F we would expect a shift towards smaller values.

Figure 5 reports the results of such an MC experiment. The left panel of the figure

[18] Large or small d_3 correspond to, respectively, strongly or very weakly influential factors, for which $\hat{\beta}_{33}$ is close to 1 or to 0 for all MC replications.

In addition to the above described regression, we have also run a regression of $\hat{\beta}_{33}$ on the constant, $\frac{1}{T}F_i'B'BF_i$, $i = 1, 2, 3$ and $\frac{1}{d_i}L_i'AA'L_i$, $i = 1, 2, 3$. All the slope coefficients in such a regression except the coefficients on $\frac{1}{T}F_3'B'BF_3$ and $\frac{1}{d_i}L_3'AA'L_3$ were insignificant. The coefficient on $\frac{1}{d_i}L_3'AA'L_3$ was significantly negative, but small in absolute value, for relatively large values of d_3.

Figure 4: The coefficient in the OLS regression of the MC replications of $\hat{\beta}_{33}$ on $\frac{1}{T}F_3'B'BF_3$ (left panel); and the estimated densities of the conditional distribution of $\hat{\beta}_{33}$ given relatively small and relatively large values of $\frac{1}{T}F_3'B'BF_3$ for fixed $d_3 = 5$.

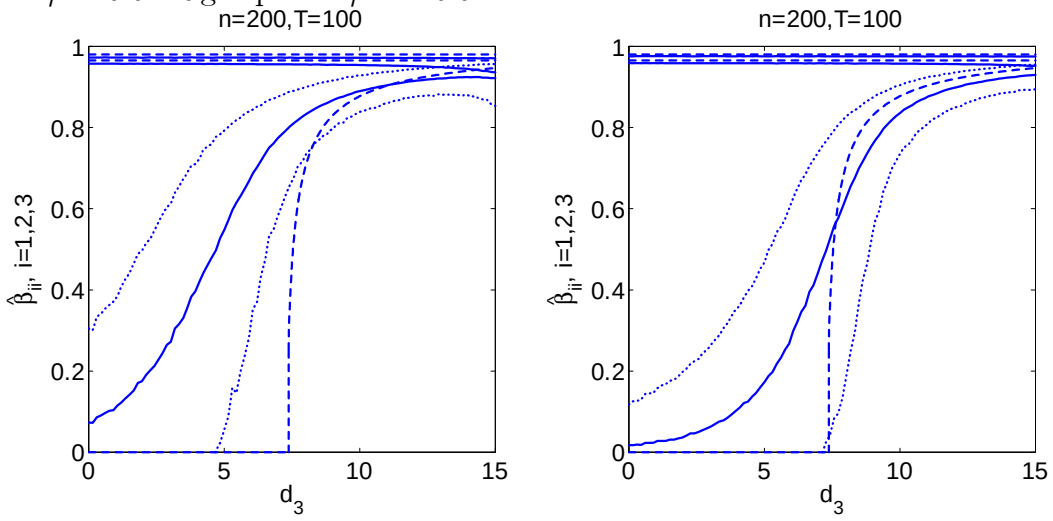


corresponds to $\rho_F = 0.5$ and the right panel corresponds to $\rho_F = -0.5$. As expected, the graphs of the MC mean and percentiles of $\hat{\beta}_{33}$ on the left picture are substantially higher than those on the right picture. Further, for $7.5 < d_3 < 10$, the MC mean of $\hat{\beta}_{33}$ is above plim $\hat{\beta}_{33}$ on the left picture, but below it on the right picture. Hence, the direction of the small sample bias of the plim $\hat{\beta}_{33}$ may change depending on particular details of the violation of assumptions 2 ii) and 2 iii). In future research, it would be interesting to replace Assumption 2 by a less restrictive assumption that would incorporate some a priori information about possible links between F and B.

In the remaining MC experiments, we will study the finite sample approximation quality of the formulae derived in Theorems 2 and 3. First, we compare the asymptotic distributions for $\hat{\beta}$ and $\hat{L}'\hat{L}$ obtained in Theorem 2 with the corresponding finite sample distributions. The general MC setting remains as above. However, we now set $\rho_1 = 0.5$ and $\rho_2 = 0$ so that there is only cross-sectional correlation in the idiosyncratic terms, which accords with the assumption of part i) of Theorem 2. We generate loadings L and factors F as $n \times 3$ and $T \times 3$ matrices with i.i.d. standard Gaussian entries so that Assumption 2ii) is violated.[19] We will keep this setting for the rest of

[19]Note that when $\rho_2 = 0$, Assumption 2iii) holds for any choice of F. Indeed, in such a case, $B'B = I_T$ so that any vector is an eigenvector of $B'B$ corresponding to the unit eigenvalue. In fact, since the distribution of ε is invariant with respect to multiplication of ε from the right by any orthogonal matrix, the results of our Monte Carlo experiment would not depend on the choice of F.

Figure 5: The correlation coefficients $\hat{\beta}_{ii}$ between true and estimated factors. Left panel: $\rho = 0.5$. Right panel: $\rho = -0.5$.



the Monte Carlo experiments in this section.

The upper panel of Figure 6 reports the means and the 5th and the 95th quantiles of the asymptotic and MC distributions of $\hat{\beta}_{33}$ and $\hat{L}_3'\hat{L}_3$ as functions of d_3. The parameters of the asymptotic distributions are shown by dashed lines, and those of the MC distributions are shown by solid lines. The range of d_3 is limited to the region where $\operatorname{plim} \hat{\beta}_{33} > 0$ so that the asymptotic distribution of $\hat{\beta}_{33}$ is available from Theorem 2 i).

The lower panel of Figure 6 provides finer details of the asymptotic approximation for $d_3 = 8$, which is in the middle of the shown range of d_3. Precisely, it superimposes the asymptotic Gaussian density with the histogram of the corresponding finite sample distribution, which is scaled so that the area of all the bars sums up to 1, and therefore, the direct comparison of the density and the histogram is possible.

Figure 6 shows that the quality of the asymptotic approximation to the finite sample distributions of $\hat{\beta}_{33}$ and $\hat{L}_3'\hat{L}_3$ is good for almost all d_3 in the shown range. The approximation is poor only for d_3 immediately above the threshold below which $\operatorname{plim} \hat{\beta}_{33} = 0$. In such a region, the variances of the asymptotic distributions of $\hat{\beta}_{33}$ and of $\hat{L}_3'\hat{L}_3$ are, respectively, much larger and much smaller than the variances of the corresponding MC distributions. As can be seen from the lower panel of Figure 6, the finite sample distribution of $\hat{\beta}_{33}$ is skewed to the left. The finite sample distribution of $\hat{L}_3'\hat{L}_3$ appears to be much more symmetric around its mean.

Figure 6: Asymptotic and MC distributions of $\hat{\beta}_{33}$ (left panel) and $\hat{L}_3'\hat{L}_3$ (right panel). Upper panel: means and 5th and 95th quantiles. Lower panel: distributions in the middle of the grid for d_3.

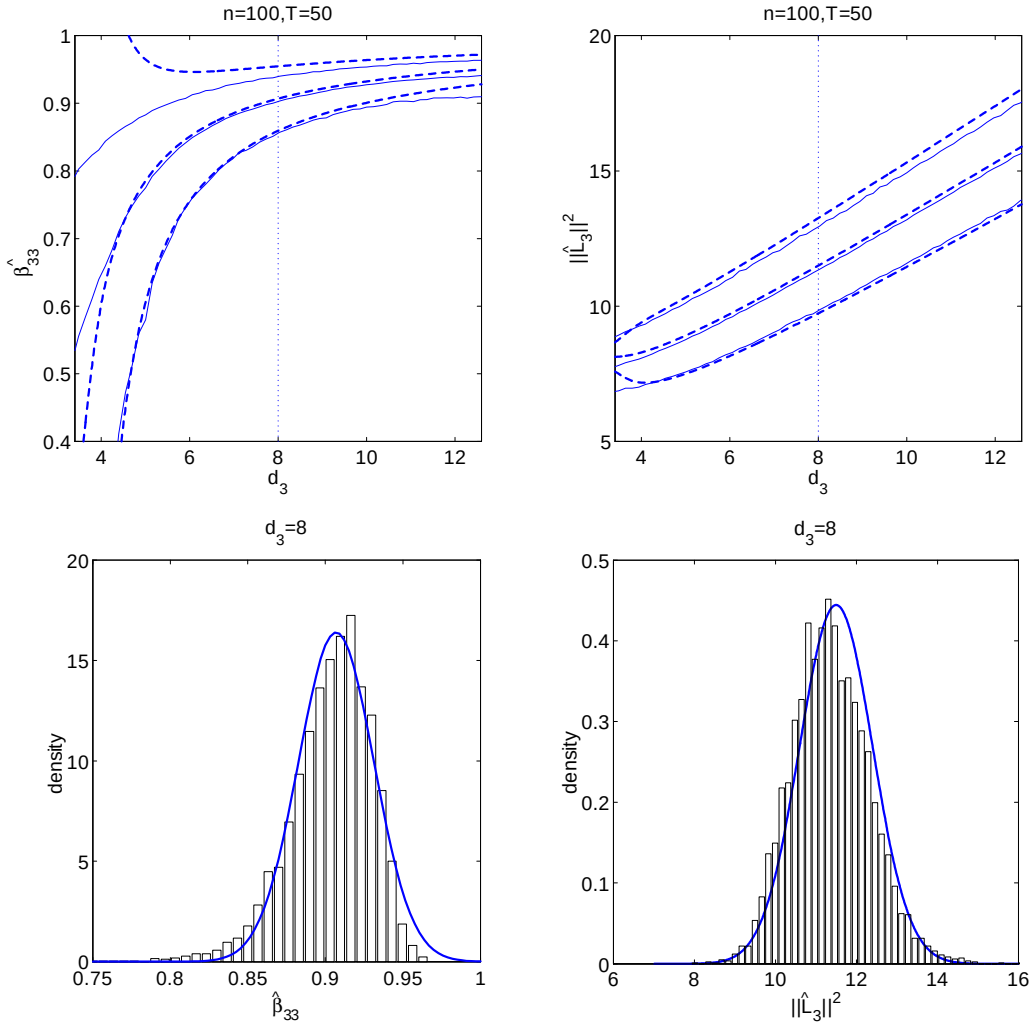
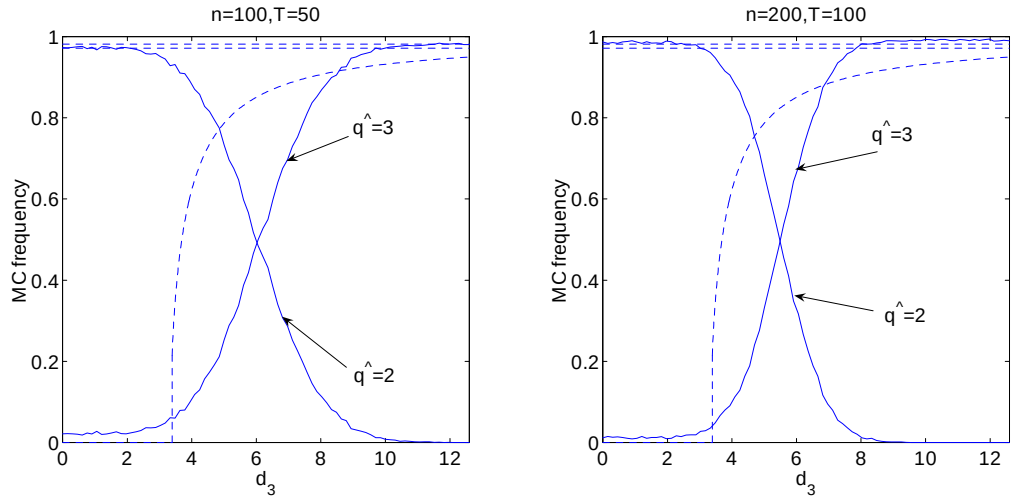


Figure 7: The MC frequencies of different values of $\hat{q}$, which is the proposed estimator of the number of factors that are positively correlated with their PC estimates. The plots of plim $\hat{\beta}_{ii}$ for $i = 1, 2$ and 3 are included as dashed lines for convenience.

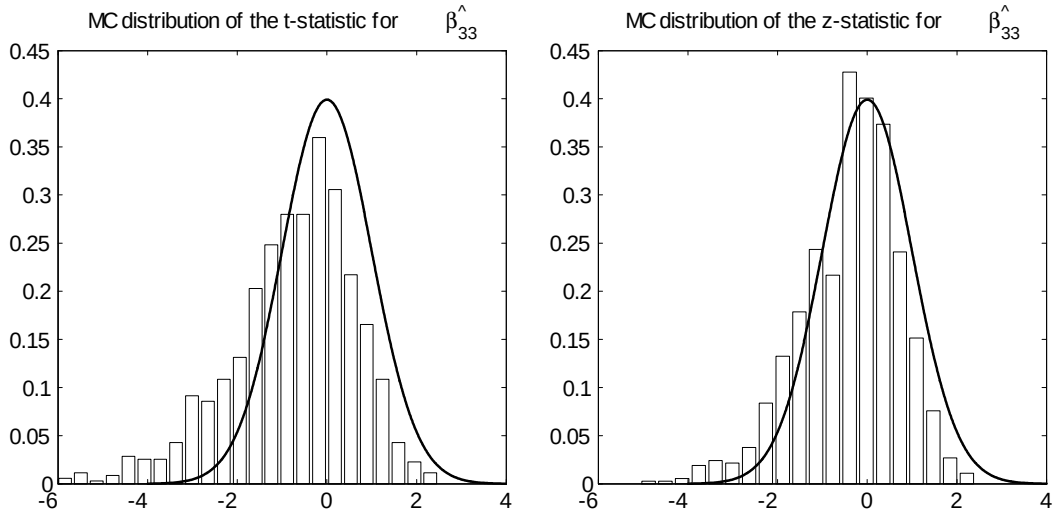


We now turn to the analysis of the finite sample approximation quality of the asymptotic distribution of $\hat{\beta}$ with parameters estimated as in Theorem 3. First, we would like to know how well the proposed estimator of q works in finite samples. Figure 7 shows the MC frequencies of different values of $\hat{q}$ as functions of the strength of the third factor d_3. The setting of the MC experiment is the same as that used to produce Figure 6. We see that when plim $\hat{\beta}_{33} = 0$ so that $q = 2$, almost all MC replications result in $\hat{q} = 2$. As plim $\hat{\beta}_{33}$ increases so that q becomes equal to 3, the MC frequency of getting $\hat{q} = 3$ increases to 1. Such an increase is faster the larger the sample size. For $n = 200$ and $T = 100$, the proportion of the MC replications in which $\hat{q} = 3$ becomes larger than 50% when plim $\hat{\beta}_{33}$ becomes larger than 0.81. It becomes larger than 90% when plim $\hat{\beta}_{33}$ becomes larger than 0.88.

In our last MC experiment, we study the finite sample distribution of the t-statistic and z-statistic for $\hat{\beta}_{33}$, which are defined as follows. Under the assumptions of Theorem 2, the asymptotic distribution of $\sqrt{T}\left(\hat{\beta}_{33} - \text{plim}\,\hat{\beta}_{33}\right)$ is normal with mean zero and variance Avar $\hat{\beta}_{33}$. Hence, we define the z-statistic for $\hat{\beta}_{33}$ as $\frac{\hat{\beta}_{33} - \text{plim}\,\hat{\beta}_{33}}{\sqrt{\text{Avar}\,\hat{\beta}_{33}/T}}$. We define the t-statistic for $\hat{\beta}_{33}$ as the z-statistic with values of plim $\hat{\beta}_{33}$ and Avar $\hat{\beta}_{33}$ replaced by their estimates from Theorem 3. As n and T go to infinity, both z and t statistics must converge in distribution to the standard normal random variables.

Figure 8 shows the MC distributions of the z and t statistics for $\hat{\beta}_{33}$ for the case

Figure 8: The MC distributions of $t = \frac{\hat{\beta}_{33} - \widehat{\text{plim}}\hat{\beta}_{33}}{\sqrt{\widehat{\text{Avar}}\hat{\beta}_{33}/T}}$ and $z = \frac{\hat{\beta}_{33} - \text{plim}\,\hat{\beta}_{33}}{\sqrt{\text{Avar}\,\hat{\beta}_{33}/T}}$. The case of $d_3 = 8$, $n = 200$ and $T = 100$. The solid line shows the density of $N(0,1)$.



of $d_3 = 8$, $n = 200$ and $T = 100$. For such a case, 989 out of 1000 MC replications result in $\hat{q} = 3$. For each of these replications, we compute z and t using formulae for $\text{plim}\,\hat{\beta}_{33}$ and $\text{Avar}\,\hat{\beta}_{33}$ from Theorem 2 and formulae for $\widehat{\text{plim}}\hat{\beta}_{33}$ and $\widehat{\text{Avar}}\hat{\beta}_{33}$ from Theorem 3. We discard the 11 MC replications with $\hat{q} < 3$ because for them, $\widehat{\text{plim}}\hat{\beta}_{33} = 0$ and $\widehat{\text{Avar}}\hat{\beta}_{33}$ is not defined.

From Figure 8, we see that our estimation procedure results in the t statistic whose finite sample distribution is similar to the finite sample distribution of the corresponding z statistic, which, of course, is infeasible. Both finite sample distributions have relatively more mass in the negative area, and both are skewed to the left. The skewness is more visible for the finite sample distribution of the t statistic. Overall, replacing $\text{plim}\,\hat{\beta}_{33}$ and $\text{Avar}\,\hat{\beta}_{33}$ by the estimates derived in Theorem 3 does not lead to large changes in the distribution of z statistic.

5 An empirical illustration

In this section, we use our theoretical results to test the adequacy of the celebrated Fama-French three-factor model of stock returns. In an influential paper, Fama and French (1993) suggest that the non-diversifiable risk in the excess stock returns can be explained by the so-called "market", "size" and "book-to-market" factors. They

propose a simple procedure to estimate these factors. First, they sort stocks according to the size and book-to-market ratio of the corresponding firms. Then, they form portfolios of stocks with similar size and book-to-market characteristics. Finally, they compute the "size" and the "book-to-market" factors as the differences between, respectively, the returns on large and small size portfolios, and the returns on large and small book-to-market portfolios. The "market" factor is computed as the capitalization-weighted index of the excess stock returns.

The Fama-French three-factor model has generated much discussion and a substantial amount of new work in the empirical finance literature. However, as pointed out in Connor and Linton (2007, p.695), "there is... no rigorous statistical theory to justify... [their factor estimation method] with regard even to consistency". The results of this paper allow us to test a hypothesis that the Fama-French factors do span the factor space of the excess stock returns.

The idea of the test is to compute the matrix $\hat{\beta}$ of the coefficients in the OLS regression of the PC estimates of the factors in a large panel of excess stock returns on the Fama-French factors, and to compare the value of $\operatorname{tr}\hat{\beta}'\hat{\beta}$ with a theoretical critical value that can be obtained from Theorems 2 and 3. If the Fama-French factors are poor proxies for the true factors, then the entries of $\hat{\beta}$ are likely to be relatively small, and the value of $\operatorname{tr}\hat{\beta}'\hat{\beta}$ should be below the critical value. If $\operatorname{tr}\hat{\beta}'\hat{\beta}$ is above the critical value, we will not reject the null that the Fama-French factors span the space of the true factors.

Our stock return data consist of monthly excess returns on 1148 stocks ($n = 1148$) traded on the NYSE, AMEX, and NASDAQ during the period from January 1983 to December 2003 ($T = 252$), obtained from CRSP data base. We included a stock in the data set if it was traded during the whole sample period. We assume that the idiosyncratic components of the excess stock returns, although cross-sectionally correlated, are not correlated over time. Since the predictability of the excess returns would imply arbitrage opportunities, such an assumption is plausible. We obtain the data on the three Fama-French factors from Kenneth French's website.

Our estimate of q, which is the number of factors whose asymptotic correlation with the corresponding PC estimate is not zero, equals two. This is consistent with Bai and Ng (2002) who also detect two pervasive factors in stock returns data. Note that the fact that $\hat{q} = 2$, which is less than the number of the Fama-French factors, does not by itself invalidates the Fama-French three-factor model. It is possible that

a particular linear combination of the Fama-French factors is simply not influential enough to be detected from our data.

Let us denote a $T \times 2$ matrix whose columns equal the PC estimates of the two factors detected from our data as $\hat{F}_{1:2}$. Further, let the $T \times 3$ matrix of the Fama-French factors be equal to $F \cdot H$, where H is an unknown non-singular matrix defined so that F satisfy the factor-normalization Assumption 1 ii). Finally, let $P_{1:2} = FH\left(H'F'FH\right)^{-1} H'F'\hat{F}_{1:2}$ be the liner projection of $\hat{F}_{1:2}$ on $F \cdot H$. Note that, since H is non-singular and since $\frac{1}{T}F'F = I_T$, $P'_{1:2}P_{1:2} = \hat{F}'_{1:2}F\left(F'F\right)^{-1}F'\hat{F}_{1:2} = T\hat{\beta}'_{1:2}\hat{\beta}_{1:2}$, where $\hat{\beta}_{1:2}$ is as defined in Theorem 2.

Our calculations give $\hat{\beta}'_{1:2}\hat{\beta}_{1:2} = \begin{pmatrix} 0.916 & 0.024 \\ 0.024 & 0.441 \end{pmatrix}$. This means that the Fama-French factors explain 91.6% of the variation in the PC estimate of the first factor and 44.1% of the variation in the PC estimate of the second factor. Is such an explanatory power consistent with the null hypothesis that the Fama-French factors span the space of the true factors? To answer this question, let us note that $\operatorname{tr}\hat{\beta}'_{1:2}\hat{\beta}_{1:2} = \sum_{i=1}^{3}\sum_{j=1}^{2}\hat{\beta}_{ij}^2 \geq \sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\beta}_{ij}^2$ so that, for any critical value c_p, $\Pr\left(\operatorname{tr}\hat{\beta}'_{1:2}\hat{\beta}_{1:2} \leq c_p\right) \leq \Pr\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\beta}_{ij}^2 \leq c_p\right)$. But the asymptotic distribution of $\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\beta}_{ij}^2$ can be estimated from our data as described by Theorem 3, and the corresponding critical values can be simulated. Hence, rejecting the null when $\operatorname{tr}\hat{\beta}'_{1:2}\hat{\beta}_{1:2} \leq c_p$ provides us with a feasible conservative test of the null hypothesis.

We have simulated 10,000 draws from the asymptotic distribution of $\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\beta}_{ij}^2$ and estimated the 0.5%, 1% and 5% critical values of the conservative test described above at $c_{0.005} = 1.702$, $c_{0.01} = 1.736$ and $c_{0.05} = 1.804$. The computed value of $\operatorname{tr}\hat{\beta}'_{1:2}\hat{\beta}_{1:2}$ is 1.357, which is smaller than any of the above critical values. Hence, we reject the null that the Fama-French factors span the space of the true stock return factors.

There are many possible reasons why the Fama-French factors may be not spanning the true factor space. For example, there might exist additional factors such as Carhart's (1997) momentum factor. Alternatively, the true factor space may be three-dimensional, but a more sophisticated procedure than that proposed by Fama and French should be used to consistently estimate the factor space (see Connor and Linton (2007) for a recent work exploring such a possibility). The analysis of this section does not shed light on these issues. Instead, it is meant to illustrate our theoretical results and to give the reader an idea of how these results can be used in an

empirical analysis.

Beside the empirical finance, our results may potentially be used in the empirical macroeconomics. For example, obtaining $\hat{q}$ for different macroeconomic datasets may give useful information on how many PC estimates to include into diffusion index forecast models or factor-augmented VARs. Interestingly, for the Stock-Watson data, which is often used in the macroeconomic analysis, $\hat{q} = 2$. This finding may explain the fact that, although the number of factors estimated from that data is often much larger than 2, using more than two of the estimated factors for forecasting does not significantly improve the quality of the forecasts (see "k fixed" panels of Tables 1 and 2 in Stock and Watson, 2002).

For the data used by Boivin et al. (2008) to study the effect of Euro on the monetary transmission mechanism, we estimate $\hat{q} = 4$. Such an estimate provides a reassuring answer to Uhlig's (2008) concern that the high explanatory power of the principal components in Boivin et al. (2008) may be an artifact of the large amount of the idiosyncratic serial correlation in their data. According to our result, the first four principal components will have genuine explanatory power despite the fact that the estimated idiosyncratic terms in Boivin et al.'s (2008) are indeed highly serially correlated.

Since typical macroeconomic data contain idiosyncratic terms that are correlated both cross-sectionally and over time, the macroeconomic applications that go beyond estimation of q would call for the extension of Theorems 2 and 3 to the case when both A and B are non-trivial. Such extensions, although very important, require substantial changes in the methods used in our proofs. We, therefore, leave them as exciting topics of future research.

6 Conclusion

In this paper we have introduced a weakly influential factor asymptotics framework which allows us to assess the finite sample properties of the PC estimator in the situation when factors' explanatory power does not strongly dominate the explanatory power of the cross-sectionally and temporally correlated idiosyncratic terms. We have shown that the principal components estimators of factors and factor loadings are inconsistent and found explicit formulae for the amount of the inconsistency. For the special case when there is no temporal correlation in the idiosyncratic terms, we have

shown that coefficients in the OLS regressions of the PC estimates of factors on the true factors are asymptotically normal, and we have found explicit formulae for the corresponding asymptotic covariance matrix. We have shown how to estimate the parameters of the derived asymptotic distributions and how to estimate the number of factors for which the PC estimator does not break down in the sense that the factors are not orthogonal to their PC estimates asymptotically. Our Monte Carlo analysis suggests that our asymptotic formulae and estimators work well even for relatively small n and T. We have applied our theoretical results to the data on the US excess stock returns, and have rejected a hypothesis that the Fama-French (1993) factors span the entire factor space of the stock returns.

Many exciting topics are left for future research. Generalizing our asymptotic distribution results to cases when both cross-sectional and temporal correlation is present in the idiosyncratic terms is one of them. Another interesting direction of research is to study the weakly influential factor asymptotics of the dynamic principal components estimator in the generalized dynamic factor model framework developed by Forni et al. (2000). Perhaps, even more interesting would be developing an alternative estimation method which would improve on the performance of the PC estimator in finite samples with weakly influential factors. One alternative to the PC estimator, which employs maximum likelihood based on Kalman filtering, has been recently introduced in Doz et al. (2006). Whether such an alternative is substantially better than the PC estimator when factors are weakly influential remains to be seen.

References

[1] Bai, J. (2003) "Inferential Theory for Factor Models of Large Dimensions", *Econometrica* 71, 135-171.

[2] Bai, J. and Ng, S (2002). "Determining the number of factors in approximate factor models", Econometrica, 70, pp 191-221

[3] Bai, J. and S. Ng (2005) "Determining the Number of Factors in Approximate Factor Models, Errata", mimeo.

[4] Bai, J. and S. Ng (2008) "Large Dimensional Factor Analysis", Foundations and Trends in Econometrics: Vol. 3: No 2, pp 89-163.

[5] Bekker, P.A. (1994), "Alternative Approximations to the Distribution of Instrumental Variables Estimators", Econometrica, 62, 657-681.

[6] Boivin, J., Giannoni, M.P. and Mojon, B. (2008) "How has the Euro Changed the monetary Transmission?", NBER Macroeconomics Annual 23.

[7] Boivin J. and S. Ng (2006) "Are More Data Always Better for Factor Analysis?", *Journal of Econometrics* 132, p. 169-194.

[8] Carhart, M.M. (1997). "On the persistence in mutual fund performance", The Journal of Finance, 52, pp. 57-82.

[9] Cattell, R. B. (1966) "The Scree Test for the Number of Factors", Multivariate Behavioral Research, vol. 1, 245-76.

[10] Chamberlain, G. and Rothschild, M. (1983) "Arbitrage, factor structure, and mean-variance analysis on large asset markets", Econometrica, 51, pp.1281-1304.

[11] Connor, G., and O. Linton (2007) "Semiparametric estimation of a characteristic-based factor model of common stock returns", Journal of Empirical Finance, 14, No5, 694-717

[12] Davidson, R. and J. G. MacKinnon (2004) Econometric Theory and Methods, Oxford University Press, New York, Oxford.

[13] 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, pp. 318-328

[14] Doz, C., Giannone, D. and L. Reichlin (2006) "A quasi maximum likelihood approach for large approximate dynamic factor models", manuscript, Universite Cergy-Pontoise

[15] Fama, Eugene F, French, Kenneth R. (1993). "Common Risk Factors in the Returns on Stocks and Bonds", Journal of Financial Economics 33 (1): 3–56.

[16] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000) "The generalized dynamic-factor model: identification and estimation", The Review of Economics and Statistics 82, pp 540-554.

[17] Forni, M. and Lippi, M. (1999) "Aggregation of linear dynamic microeconomic models", Journal of Mathematical Economics 31, pp. 131-158

[18] Forni, M. and Lippi, M. (2001) "The generalized dynamic factor model: representation theory", Econometric Theory 17, pp. 1113-1141

[19] Grenander, U. and G. Szegö (1958) Toeplitz Forms and Their Applications. University of California Press, Berkeley.

[20] Harding, M. (2006) "Structural estimation of high-dimensional factor models", unpublished manuscript, Stanford university.

[21] Johnstone, I.M., and A. Y. Lu (2007) "Sparse principal components analysis", Journal of the American Statistical Association. To appear.

[22] Marchenko, V.A., and L.A. Pastur (1967) "Distribution of eigenvalues for some sets of random matrices", Math. USSR-Sbornik, vol. 1, no. 4, 457-483

[23] McManus, D.A. (1991) "Who Invented Local Power Analysis?", Econometric Theory 7, 265-268

[24] Neyman, J. (1937) ""Smooth" Test for Goodness of Fit", Skandinavisk Aktuarietiskrift 20, 149-199.

[25] Onatski, A. (2009) "Determining the number of factors from the empirical distribution of eigenvalues", forthcoming in the Review of Economics and Statistics.

[26] Onatski, A., M. Moreira and M. Hallin (2011) "Local asymptotic power of the eigenvalue-based tests for high-dimensional covariance matrices", manuscript, University of Cambridge.

[27] Paul, D. (2007) "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model", Statistica Sinica 17, pp. 1617-1642

[28] Reichlin, L. (2003) "Factor models in large cross sections of time series", in Dewatripont, M., Hansen, P.L. and S. Turnowsky, editors, Advances in Economics and Econometrics: Theory and Applications, Vol. 11, 8th Wold Congress of the Econometric Society, Cambridge University Press.

[29] Staiger, D., and Stock, J.H. (1997) "Instrumental Variables Regression With Weak Instruments", Econometrica, 65, 557-586.

[30] Stock, J. and Watson, M. (2002) "Macroeconomic Forecasting Using Diffusion Indexes", Journal of Business and Economic Statistics, 20, pp. 147-162.

[31] Stock, J. and Watson, M. (2005) "Implications of Dynamic Factor Models for VAR Analysis", manuscript, Harvard University.

[32] Stock, J. and M. Watson (2006) Macroeconomic Forecasting Using Many Predictors, in Graham Elliott, Clive Granger and Allan Timmerman, editors, Handbook of Economic Forecasting, North Holland.

[33] Uhlig, H. (2008) "Macroeconomic Dynamics in the Euro Area. Discussion by Harald Uhlig", NBER Macroeconomics Annual 23.

[34] 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, pp. 293–314.

[35] Watson, M.W. (2003) "Macroeconomic Forecasting Using Many Predictors", in M. Dewatripont, L. Hansen and S. Turnovsky (eds), Advances in Economics and Econometrics, Theory and Applications, Eight World Congress of the Econometric Society," Vol. III, page 87-115.

[36] Zhang, L. (2006). Spectral Analysis of Large Dimensional Random Matrices. Ph. D. Thesis. National University of Singapore.