cemmap
centre for microdata methods and practice

Linear regression for panel with unknown number of factors as interactive fixed effects

Hyungsik Roger Moon
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP49/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects*

Hyungsik Roger Moon[‡] Martin Weidner[§]

October 2, 2013

Abstract

In this paper we study the least squares (LS) estimator in a linear panel regression model with interactive fixed effects for asymptotics where both the number of time periods and the number of cross-sectional units go to infinity. Under appropriate assumptions we show that the limiting distribution of the LS estimator for the regression coefficients is independent of the number of interactive fixed effects used in the estimation, as long as this number does not fall below the true number of interactive fixed effects present in the data. The important practical implication of this result is that for inference on the regression coefficients one does not necessarily need to estimate the number of interactive effects consistently, but can rely on an upper bound of this number to calculate the LS estimator.

Keywords: Panel data, interactive fixed effects, factor models, perturbation theory of linear operators, random matrix theory.

JEL-Classification: C23, C33

1 Introduction

Panel data models typically incorporate individual and time effects to control for heterogeneity in cross-section and over time. While often these individual and time effects enter the model additively, they can also be interacted multiplicatively, thus giving rise to so called interactive effects, which we also refer to as a factor structure. The multiplicative

*We thank the participants of the Cowles Summer Conference "Handling Dependence: Temporal, Cross-sectional, and Spatial" at Yale University, of the 2012 North American Summer Meeting of the Econometric Society at Northwestern University, and of the 18th International Conference on Panel Data at the Banque de France for many interesting comments, and we thank Alexei Onatski for helpful discussions. We are also grateful for the comments and suggestions of James Stock and three anonymous referees. Moon acknowledges financial supports of the NSF via SES 0920903 and from USC. Weidner acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0002.

[‡]Department of Economics, University of Southern California, Los Angeles, CA 90089-0253. Email: moonr@usc.edu. Department of Economics, Yonsei University, Seoul, Korea.

[§]Corresponding author. Department of Economics, University College London, Gower Street, London WC1E 6BT, U.K., and CeMMaP. Email: m.weidner@ucl.ac.uk.

form captures the heterogeneity in the data more flexibly, since it allows for common time-varying shocks (factors) to affect the cross-sectional units with individual specific sensitivities (factor loadings).[1] It is this flexibility that motivated the discussion of interactive effects in the econometrics literature, e.g. Holtz-Eakin, Newey and Rosen (1988), Ahn, Lee and Schmidt (2001; 2007), Pesaran (2006), Bai (2009b; 2009a), Zaffaroni (2009), Moon and Weidner (2010), and Lu and Su (2013).

Let N be the number of cross-sectional units, T be the number of time periods, K be the number of regressors, and R^0 be the true number of interactive fixed effects. We consider a linear regression model with observed outcomes Y, regressors X_k, and unobserved error structure ε, namely

$$Y = \sum_{k=1}^{K} \beta_k^0 \, X_k \, + \varepsilon \, , \qquad \qquad \varepsilon = \lambda^0 \, f^{0\prime} + e \, , \tag{1.1}$$

where Y, X_k, ε and e are $N \times T$ matrices, λ^0 is an $N \times R^0$ matrix, f^0 is a $T \times R^0$ matrix, and the regression parameters β_k^0 are scalars — the superscript zero indicates the true value of the parameters. We write β for the K-vector of regression parameters, and we denote the components of the different matrices by Y_{it}, $X_{k,it}$, e_{it}, λ_{ir}^0 and f_{tr}^0, where $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $r = 1, \ldots, R^0$. It is convenient to introduce the notation $\beta \cdot X \equiv \sum_{k=1}^{K} \beta_k \, X_k$. All matrices, vectors and scalars in this paper are real valued.

We consider the interactive fixed effect specification, i.e. we treat λ^0 and f^0 as nuisance parameters, which are estimated jointly with the parameters of interest β.[2] The advantages of the fixed effects approach are for instance that it is semi-parametric, since no assumption on the distribution of the interactive effects needs to be made, and that the regressors can be arbitrarily correlated with the interactive effect parameters.

The main goal of the paper is to estimate the regression coefficient β when the number of the factors R^0 is *unknown*. The estimator we study is the least squares (LS) estimator.[3]

The LS estimator of model (1.1) minimizes the sum of squared residuals e_{it} to estimate the unknown parameters β, λ and f. To our knowledge, this estimator was first discussed in Kiefer (1980). Under an asymptotic where N and T grow to infinity, the asymptotic properties of the LS estimator were derived in Bai (2009b) for strictly exogeneous regressors, and extended in Moon and Weidner (2010) to the case of pre-determined regressors.

An important restriction of these papers is that the number of factors R^0 is assumed to be known. However, in many empirical applications there is no consensus about the exact number of factors in the data or in the relevant economic model. If R^0 is not known beforehand, then it may be estimated consistently. However, most of the estimators for R^0 were developed for pure factor models that have no regressor. In order to use these

[1] The conventional additive model can be interpreted as a two factor interactive fixed effects model.

[2] When we refer to interactive fixed effects we mean that both factors and factor loadings are treated as non-random parameters. Ahn, Lee and Schmidt (2001; 2007) take a hybrid approach in that they treat the factors as non-random, but the factor loadings as random. The common correlated effects estimator of Pesaran (2006) was introduced in a context, where both the factor loadings and the factors follow certain probability laws, but it exhibits many properties of a fixed effects estimator.

[3] The LS estimator is sometimes called "concentrated" least squares estimator in the literature, and in an earlier version of the paper we referred to it as the "Gaussian Quasi Maximum Likelihood Estimator", since LS estimation is equivalent to maximizing a conditional Gaussian likelihood function. Note also that for fixed β the LS estimator for λ and f is simply the principal components estimator.

existing results, one needs to find out the asymptotic properties of the estimator of β when R ($> R^0$) factors are used, which is the purpose of the current paper.

We investigate the asymptotic properties of the LS estimator when the true number of factors R^0 is unknown and R ($> R^0$) number factors are used in the estimation.[4] We denote this estimator as $\widehat{\beta}_R$.

The main result of the paper, presented in Section 3, is that under appropriate assumptions the LS estimator $\widehat{\beta}_R$ has the same limiting distribution as $\widehat{\beta}_{R^0}$ for any $R \geq R^0$ under an asymptotic where both N and T become large, while R^0 and R are constant. This implies that the LS estimator $\widehat{\beta}_R$ is asymptotically robust towards inclusion of extra interactive effects in the model, and within the LS estimation framework there is no asymptotic efficiency loss from choosing R larger than R^0. The important empirical implication of our result is that the number of factors R^0 need not be known or estimated accurately to apply the LS estimator.

To derive this robustness result, we impose more restrictive conditions than those typically assumed with known R^0. These include that the errors e_{it} are independent and identically (iid) normally distributed and that the regressors are composed of a "low-rank" strictly stationary component, a "high-rank" strictly stationary component, and a "high-rank" pre-determined component.[5] Notice that while some of these restrictions are necessary for our robustness result (see Section 4.3), some of them (e.g. iid normality of e_{it}) are technical regularity conditions required to use certain results from the theory of random matrices (see the discussion in Section 4.5). In the Monte Carlo simulations in Section 5, we consider DGPs that violate some technical conditions to demonstrate robustness of the result.

In Section 4, under less restrictive assumptions, we provide intermediate results that lead to the main result. In Section 4.1 we show $\sqrt{\min(N,T)}$-consistency of the LS estimator $\widehat{\beta}_R$ as $N, T \to \infty$ under very mild regularity condition on X_{it} and e_{it}, and without imposing any assumptions on λ^0 and f^0 apart from $R \geq R^0$. We thus obtain consistency of the LS estimator not only for unknown number factors, but also for weak factors[6], which is a remarkable robustness result.

In Section 4.2 we derive an asymptotic expansion of the LS profile objective function that concentrates out f and λ, for the case $R = R^0$. Given that the profile objective function is a sum of eigenvalues of a covariance matrix, its quadratic approximation is challenging because the derivatives of the eigenvalues with respect to β are not generally known. We thus cannot use a conventional Taylor expansion, but instead apply the perturbation theory of linear operators to derive the approximation. The resulting expansion of the profile objective function can be used to reproduce the asymptotic results in Bai (2009b). In Moon and Weidner (2010) we employ the results derived here to extend Bai's result to the case of pre-determined regressors.

In Section 4.3 we provide an example that satisfies the typical assumptions imposed with known R^0, so that $\widehat{\beta}_{R_0}$ is $\sqrt{NT}$ consistent, but we show that $\widehat{\beta}_R$ with $R > R_0$ is only

[4]For $R < R^0$ the LS estimator can be inconsistent, since then there are interactive fixed effects in the model which can be correlated with the regressors but are not controlled for in the estimation. We therefore restrict attention to the case $R \geq R^0$.

[5]The pre-determined component of the regressors allows for linear feedback of e_{it} into future realizations of $X_{k,it}$.

[6]Onatski (2012) discusses the "weak factor" assumption for the purpose of estimating the number of factors in a pure factor model, and a more general discussion of strong and weak factors is given in Chudik, Pesaran and Tosetti (2011).

$\sqrt{\min(N,T)}$ consistent in that example. This shows that stronger conditions are required to derive our main result.

In Section 4.4 we show faster than $\sqrt{\min(N,T)}$-convergence of $\widehat{\beta}_R$ under assumptions that are less restrictive than those employed for the main result, in particular allowing for either cross-sectional or time-serial correlation of the errors e_{it}.

In Section 4.5 we provide an alternative version of our main result of asymptotic equivalence of $\widehat{\beta}_{R^0}$ and $\widehat{\beta}_R$, $R > R^0$, which is derived under high-level assumptions. Verifying those high-level assumptions requires good knowledge of the properties of the eigenvalues and eigenvectors of the random covariance matrix of the errors e. Further progress in the Random Matrix Theory of real random covariance matrices (see e.g. Bai (1999)) might allow to verify those high-level assumptions for more general error distributions e.

Section 5 contains Monte Carlo simulation results for a static panel model. We consider a DGP that violates the iid normality restriction of the error term. The simulation results confirm our main result of the paper even with a relatively small sample size (e.g. $N = 100, T = 10$) and non-iid-normal errors. In the supplementary appendix, we report the Monte Carlo simulation results of an AR(1) panel model. It also confirms the robust result in large samples, but in finite samples it shows more inefficiency than the static case. In general, one should expect some finite sample inefficiency from overestimating the number of factors when the sample size is small or the number of overfitted factors is large.

A few words on notation. The transpose of a matrix A is denoted by A'. For a column vectors v its Euclidean norm is defined by $\|v\| = \sqrt{v'v}$. For an $m \times n$ matrix A the Frobenius or Hilbert Schmidt norm is $\|A\|_{HS} = \sqrt{\mathrm{Tr}(AA')}$, and the operator or spectral norm is $\|A\| = \max_{0 \neq v \in \mathbb{R}^n} \frac{\|Av\|}{\|v\|}$. Furthermore, we use $P_A = A(A'A)^\dagger A'$ and $M_A = \mathbb{1} - A(A'A)^\dagger A'$, where $\mathbb{1}$ is the $m \times m$ identity matrix, and $(A'A)^\dagger$ denotes some generalized inverse, in case A is not of full column rank. For square matrices B, C, we use $B > C$ (or $B \geq C$) to indicate that $B - C$ is positive (semi) definite. We use "wpa1" for "with probability approaching one".

2 Identification of $\beta^0, \lambda^0 f^{0\prime}$, and R^0

In this section we provide a set of conditions under which the regression coefficient β^0, the interactive fixed effects $\lambda^0 f^{0\prime}$, and the number of factors R^0 are determined uniquely by the data. Here, and throughout the whole paper, we treat λ and f as non-random parameters, i.e. all stochastics in the following are implicitly conditional on λ and f. Let $x_k = \mathrm{vec}(X_k)$, the NT-vectorization of X_k, and let $x = (x_1, \ldots, x_K)$, which is a $NT \times K$ matrix.

Assumption ID (Assumptions for Identification).

(i) *The second moments of X_{it} and e_{it} exist for all i, t.*

(ii) $\mathbb{E}(e_{it}) = 0$, $\mathbb{E}(X_{it} e_{it}) = 0$, *for all i, t.*

(iii) $\mathbb{E}[x'(M_F \otimes M_{\lambda^0})x] > 0$, *for all $F \in \mathbb{R}^{T \times R}$,*

(iv) $R^0 \equiv \mathrm{rank}(\lambda^0 f^{0\prime}) \leq R$.

Theorem 2.1 (Identification). *Suppose that the Assumptions ID are satisfied. Then, β^0, $\lambda^0 f^{0\prime}$, and R^0 are identified.*[7]

Assumption ID(i) imposes existence of second moments. Assumption ID(ii) is an exogeneity condition, which demands that x_{it} and e_{it} are not correlated contemporaneously, but allows for pre-determined regressors like lagged dependent variables. Assumption ID(iv) imposes that the true number of factors $R^0 = \text{rank}(\lambda^0 f^{0\prime})$ is bounded by a positive integer R, which cannot be too large (e.g. the trivial bound $R = N$ is not possible), since otherwise Assumption ID(iii) cannot be satisfied.

Assumption ID(iii) is a non-collinearity condition, which demands that the regressors have significant variation across i and over t after projecting out all variation that can be explained by the factor loadings λ^0 and by arbitrary factors $F \in \mathbb{R}^{T\times R}$. This generalizes the within variation assumption in the conventional fixed effect panel regression, which in our notation reads $\mathbb{E}[x'(M_{1_T} \otimes \mathbb{1}_N)x] > 0$. This conventional fixed effect assumption rules out time-invariant regressors. Similarly, Assumption ID(iii) rules out more general "low-rank regressors", including both time-invariant and cross-sectional invariant regressors.[8]

3 Main Result

The estimator we investigate in this paper is the least squares (LS) estimator, which for a given choice of R reads[9]

$$\left(\widehat{\beta}_R,\, \widehat{\Lambda}_R,\, \widehat{F}_R\right) \in \underset{\{\beta\in\mathbb{R}^K,\, \Lambda\in\mathbb{R}^{N\times R},\, F\in\mathbb{R}^{T\times R}\}}{\operatorname{argmin}} \left\| Y - \beta\cdot X - \Lambda F' \right\|_{HS}^2 , \tag{3.1}$$

where $\|.\|_{HS}$ refers to the Hilbert Schmidt norm, also called Frobenius norm. The objective function $\|Y - \beta\cdot X - \Lambda F'\|_{HS}^2$ is simply the sum of squared residuals. The estimator for β^0 can equivalently be defined by minimizing the profile objective function that concentrates out the R factors and the R factor loadings, namely

$$\widehat{\beta}_R = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}}\, \mathcal{L}_{NT}^R(\beta) , \tag{3.2}$$

[7]Here, identification means that β^0 and $\lambda^0 f^{0\prime}$ can be uniquely recovered from the distribution of (Y, X) conditional on those parameters. Identification of the number of factors follows since $R^0 = \text{rank}(\lambda^0 f^{0\prime})$. The factor loadings and factors λ^0 and f^0 are not separately identified without further normalization restrictions, but the product $\lambda^0 f^{0\prime}$ is identified.

[8]We do not consider such "low-rank regressors" in this paper. Further discussion can be found in Bai (2009b), whose Assumption A is the sample version of Assumption ID(iii). See also our discussion of Assumption NC below.

[9]The optimal $\widehat{\Lambda}_R$ and $\widehat{F}_R$ in (3.1) are not unique, since the objective function is invariant under right-multiplication of Λ with a non-degenerate $R\times R$ matrix S, and simultaneous right-multiplication of F with $(S^{-1})'$. However, the column spaces of $\widehat{\Lambda}_R$ and $\widehat{F}_R$ are uniquely determined.

with[10]

$$\begin{aligned}
\mathcal{L}^{R}_{NT}(\beta) &= \min_{\{\Lambda\in\mathbb{R}^{N\times R},\, F\in\mathbb{R}^{T\times R}\}} \frac{1}{NT} \left\| Y - \beta\cdot X - \Lambda F' \right\|^2_{HS} \\
&= \min_{F\in\mathbb{R}^{T\times R}} \frac{1}{NT} \operatorname{Tr}\left[(Y-\beta\cdot X)\, M_F\, (Y-\beta\cdot X)' \right] \\
&= \frac{1}{NT} \sum_{r=R+1}^{T} \mu_r \left[(Y-\beta\cdot X)' \, (Y-\beta\cdot X) \right] , \qquad (3.3)
\end{aligned}$$

where, $\mu_r(.)$ is the r'th largest eigenvalue of the matrix argument. Here, we first concentrated out Λ by use of its own first order condition. The resulting optimization problem for F is a principal components problem, so that the the optimal F is given by the R largest principal components of the $T\times T$ matrix $(Y-\beta\cdot X)'\,(Y-\beta\cdot X)$. At the optimum the projector M_F therefore exactly projects out the R largest eigenvalues of this matrix, which gives rise to the final formulation of the profile objective function as the sum over its $T-R$ smallest eigenvalues.[11] Since the model is symmetric under $N\leftrightarrow T$, $\Lambda\leftrightarrow F$, $Y\leftrightarrow Y'$, $X_k\leftrightarrow X'_k$ there also exists a dual formulation of $\mathcal{L}^{R}_{NT}(\beta)$ that involves solving an eigenvalue problem for an $N\times N$ matrix. We write $\mathcal{L}^{0}_{NT}(\beta)$ for $\mathcal{L}^{R^0}_{NT}(\beta)$, the profile objective function obtained for the true number of factors. Notice that we do not impose a compact parameter set for β.

Assumption SF (Strong Factor Assumption).

(i) $0 < \operatorname{plim}_{N,T\to\infty} \frac{1}{N}\, \lambda^{0\prime}\lambda^{0} < \infty$,

(ii) $0 < \operatorname{plim}_{N,T\to\infty} \frac{1}{T}\, f^{0\prime} f^{0} < \infty$.

Assumption NC (Non-Collinearity of X_k). *Consider linear combinations $\alpha\cdot X = \sum_{k=1}^{K} \alpha_k X_k$ of the regressors X_k with K-vector α such that $\|\alpha\| = 1$. We assume that there exists a constant $b > 0$ such that*

$$\min_{\{\alpha\in\mathbb{R}^K,\, \|\alpha\|=1\}} \sum_{r=R+R^0+1}^{T} \mu_r \left[\frac{(\alpha\cdot X)'(\alpha\cdot X)}{NT} \right] \geq b \, , \qquad wpa1.$$

Assumption LL (Low Level Conditions for Main Result).

(i) **Decomposition of Regressors:** $X_k = \overline{X}_k + \widetilde{X}^{\rm str}_k + \widetilde{X}^{\rm weak}_k$*, for* $k = 1,\ldots,K$*, where* $\overline{X}_k$*,* $\widetilde{X}^{\rm str}_k$ *and* $\widetilde{X}^{weak}_k$ *are* $N\times T$ *matrices, and*

(i.a) **Low-Rank (strictly exogenous) Part of Regressors:** $\operatorname{rank}(\overline{X}_k)$ *is bounded as* $N,T\to\infty$*, and* $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{X}^2_{k,it} = \mathcal{O}_P(1)$.

[10] The profile objective function $\mathcal{L}^{R}_{NT}(\beta)$ need not be convex in β and can have multiple local minima. Depending on the dimension of β one should either perform an initial grid search or try multiple starting values for the optimization when calculating the global minimum $\widehat{\beta}_R$ numerically. See also Section S.7 of the supplementary material.

[11] This last formulation of $\mathcal{L}^{R}_{NT}(\beta)$ is very convenient since it does not involve any explicit optimization over nuisance parameters. Numerical calculation of eigenvalues is very fast, so that the numerical evaluation of $\mathcal{L}^{R}_{NT}(\beta)$ is unproblematic for moderately large values of T.

(i.b) **High-Rank (strictly exogenous) Part of Regressors:** $\|\widetilde{X}_k^{str}\| = \mathcal{O}_P(N^{3/4})$, *as can be justified e.g. by Lemma A.1 in the appendix.*

(i.c) **Weakly Exogenous Part of Regressors:** $\widetilde{X}_{k,it}^{\text{weak}} = \sum_{\tau=1}^{t-1} \gamma_\tau e_{i,t-\tau}$, *where the real valued coefficients* γ_τ *satisfy* $\sum_{\tau=1}^{\infty} |\gamma_\tau| < \infty$.

(i.d) **Bounded Moments:** *We assume that* $\mathbb{E}\,|X_{k,it}|^2$, $\mathbb{E}\left|(M_{\lambda^0} X_k M_{f^0})_{it}\right|^{26}$, $\mathbb{E}\,|(M_{\lambda^0} X_k)_{it}|^8$ *and* $\mathbb{E}\left|(X_k M_{f^0})_{it}\right|^8$ *are bounded uniformly over* k, i, j, N *and* T.

(ii) **Errors are iid Normal:** *The error matrix* e *is independent of* λ^0, f^0, $\overline{X}_k$, *and* $\widetilde{X}_k^{\text{str}}$, $k = 1, \ldots, K$, *and its elements* e_{it} *are distributed as iid* $\mathcal{N}(0, \sigma^2)$.

(iii) **Number of Factors not Underestimated:** $R \geq R^0 = \text{rank}(\lambda^0 f^{0\prime})$.

Remarks

(i) Assumption SF assumes that the factor f^0 and the factor loading λ^0 are strong. The strong factor assumption is regularly imposed in the literature on large N and T factor models, including Bai and Ng (2002), Stock and Watson (2002) and Bai (2009b).

(ii) Assumption NC assumes that there exists significant sampling variation in the regressors after concentrating out $R + R^0$ factors (or factor loadings). It is a sample version of the identification Assumption ID(*iii*), and it is essentially equivalent to Assumption A of Bai (2009b), but avoids mentioning the unobserved loadings λ^0.[12]

(iii) Assumption NC is violated if there exists a linear combination $\alpha \cdot X$ of the regressors with $\alpha \neq 0$ and $\text{rank}(\alpha \cdot X) \leq R + R^0$, i.e. the assumption rules out "low-rank regressors" like time invariant regressors or cross-sectionally invariant regressors. These low-rank regressors require a special treatment in the interactive fixed effect model, see Bai (2009b) and Moon and Weidner (2010), and we do not consider them in the present paper. If one is not interested explicitly in their regression coefficients, then one can always eliminate the low-rank regressors by an appropriate projection of the data, e.g. subtraction of the time (or cross-sectional) means from the data eliminates all time-invariant (or cross-sectionally invariant) regressors.

(iv) The norm restriction in Assumption LL(*i.b*) is a high level assumption. It is satisfied as long as $\widetilde{X}_{k,it}^{\text{weak}}$ is mean zero and weakly correlated across i and over t, for details see Appendix A.1 and Lemma A.1 there.

(v) Assumption LL(*i*) imposes that each regressor consists of three parts: (a) a strictly exogenous low rank component , (b) a strictly exogenous component satisfying a norm restriction, and (c) a weakly exogenous component that follows a linear process with innovation given by the lagged error term e_{it}. For example, if $X_{k,it} \sim iid\,\mathcal{N}(\mu_k, \sigma_k^2)$, independent of e, then we have $\overline{X}_{k,it} = \mu_k$, $\widetilde{X}_{k,it}^{\text{str}} \sim iid\,\mathcal{N}(0, \sigma_k^2)$ and $\widetilde{X}_k^{\text{weak}} = 0$. Assumption LL(*i*) is also satisfied for a stationary panel VAR with interactive fixed effects as in Holtz-Eakin, Newey and Rosen (1988). A special case of this is a dynamic

[12] By dropping the expected value from Assumption ID(*iii*) and replacing the zero lower bound by a positive constant one obtains $\inf_F [x'(M_F \otimes M_{\lambda^0})x/NT] \geq b > 0$, wpa1, which is equivalent to Assumption A of Bai (2009b), and can also be rewritten as $\min_{\|\alpha\|=1} \inf_F \text{Tr}\,[M_{\lambda^0}(\alpha \cdot X)' M_F (\alpha \cdot X)/NT] \geq b$. A slightly stronger version of the Assumption, which avoids mentioning the unobserved factor loading λ^0, reads $\min_{\|\alpha\|=1} \inf_F \inf_\lambda \text{Tr}\,[M_\lambda(\alpha \cdot X)' M_F (\alpha \cdot X)/NT] \geq b$, where $F \in \mathbb{R}^{T \times R}$ and $\lambda \in \mathbb{R}^{N \times R^0}$, and this slightly stronger version is equivalent to Assumption NC.

panel regression with fixed effects, where $Y_{it} = \beta Y_{i,t-1} + \lambda_i^{0\prime} f_t^0 + e_{it}$, with $|\beta| < 1$ and "infinite history". In this case, we have $X_{it} = Y_{i,t-1} = \overline{X}_{it} + \widetilde{X}_{it}^{\rm str} + \widetilde{X}_{it}^{\rm weak}$, where $\overline{X}_{it} = \lambda_i^{0\prime} \sum_{\tau=1}^{\infty} \beta^{\tau-1} f_{t-\tau}^0$, $\widetilde{X}_{it}^{\rm str} = \sum_{\tau=t}^{\infty} \beta^{\tau-1} e_{i,t-\tau}$, and $\widetilde{X}_{it}^{\rm weak} = \sum_{\tau=0}^{t-1} \beta^{\tau-1} e_{i,t-\tau}$.

(vi) Assumption LL(i) is more general than the restriction on the regressors in Pesaran (2006), where – in our notation – the decomposition $X_k = \overline{X}_k + \widetilde{X}_k^{\rm str}$ is imposed, but the lower rank component $\overline{X}_k$ needs to satisfy further assumptions, and the weakly exogenous component $\widetilde{X}_k^{\rm weak}$ is not considered. Bai (2009b) requires no such decomposition, but imposes strict exogeneity of the regressors.

(vii) Among the conditions in Assumption LL, the iid normality condition in Assumption LL(ii) may be the most restrictive. In Section 4 provides more general high-level conditions under which the main result in Theorem 3.1 holds. Verifying those high-level conditions requires results on the eigenvalues and eigenvectors of random covariance matrices, see Assumption EV below. By using results from the random matrix theory literature we are able to verify those high-level condition for the case of iid normal errors. We believe, however, that those high-level conditions and thus our main result hold more generally, and we explore non-normal and serially correlated errors in our Monte Carlo simulations below.

Theorem 3.1 (Main Result). *Let Assumption SF, NC and LL hold and consider a limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Then we have*

$$\sqrt{NT}\big(\widehat{\beta}_R - \beta^0\big) = \sqrt{NT}\big(\widehat{\beta}_{R^0} - \beta^0\big) + o_P(1).$$

Theorem 3.1 follows from Theorem 4.6 and Lemma 4.7 below. Under appropriate conditions the theorem states that the inclusion of unnecessary factors in the estimation does not change the asymptotic distribution of the LS estimator for β^0. From Bai (2009b) and Moon and Weidner (2010) it is known that $\sqrt{NT}\big(\widehat{\beta}_{R^0} - \beta^0\big)$ is asymptotically normally distributed, so the same is true for $\sqrt{NT}\big(\widehat{\beta}_R - \beta^0\big)$, $R > R^0$, and the asymptotic bias and variance of $\widehat{\beta}_{R^0}$ and $\widehat{\beta}_R$ are also shown to be identical by the theorem.[13,14]

In the rest of this section we provide a heuristic discussion of the main result. Intuitively, the inclusion of unnecessary factors in the LS estimation is similar to the inclusion of irrelevant regressors in an OLS regression. In the OLS case it is well known that if those irrelevant extra regressors are uncorrelated with the regressors of interest, then they have no effect on the asymptotic distribution of the regression coefficients of interest. It is thus natural to expect that if the extra estimated factors in $\widehat{F}_R$ are asymptotically uncorrelated with the regressors, then the result of Theorem 3.1 should hold. To explore this, remember that the estimator $\widehat{F}_R$ is given by the first R principal components of the matrix

[13] Bai (2009b) finds asymptotic bias in $\widehat{\beta}_{R^0}$ due to heteroscedasticity and correlation in e_{it}, which in our asymptotic result is ruled out by Assumption LL(ii), but is studies in our Monte Carlo simulations below. Moon and Weidner (2010) work out the additional asymptotic bias in $\widehat{\beta}_{R^0}$ due to pre-determined regressors, which is allowed for in Theorem 3.1.

[14] Interestingly, a result analogous to Theorem 3.1 does not hold when comparing the pooled panel OLS estimator ($\widehat{\beta}_0$) to the within group estimator (the analog of $\widehat{\beta}_1$ for standard fixed effects) in a situation where no fixed effects are present in the true DGP ($R^0 = 0$). In that case one imposes a factor that is redundant (namely $f^0 = (1, 1, \ldots, 1)'$), while in our case we are estimating a redundant factor by principal components.

$(Y - \widehat{\beta}_R \cdot X)'(Y - \widehat{\beta}_R \cdot X)$, and write

$$Y - \widehat{\beta}_R \cdot X = \lambda^0 f^{0\prime} + e - (\widehat{\beta}_R - \beta^0) \cdot X. \tag{3.4}$$

The strong factor assumption guarantees that the first R^0 principal components of $(Y - \widehat{\beta}_R \cdot X)'(Y - \widehat{\beta}_R \cdot X)$ are close to f^0 asymptotically, i.e. the true factors are correctly picked up by the principal component estimator. The additional $R - R^0$ principal components that are estimated for $R > R^0$ cannot pick up anymore true factors and are thus mostly determined by the remaining term $e - (\widehat{\beta}_R - \beta^0) \cdot X$. The key question for the properties of the extra estimated factors, and thus of $\widehat{\beta}_R$, is therefore whether the principal components obtained from $e - (\widehat{\beta}_R - \beta^0) \cdot X$ are dominated by e or by $(\widehat{\beta}_R - \beta^0) \cdot X$. Only if they are dominated by e can we expect the extra factors in $\widehat{F}_R$ to be uncorrelated with X and thus the result in Theorem 3.1 to hold.

Under our assumptions we have $\|e\| = \mathcal{O}_P(\sqrt{N})$ and $\|X_k\| = \mathcal{O}_P(\sqrt{NT})$ as N and T grow at the same rate. Thus, if the convergence rate of $\widehat{\beta}_R$ is faster than $\sqrt{N}$, i.e. $\|\widehat{\beta}_R - \beta^0\| = o_P(\sqrt{N})$, then we have $\|e\| \gg \left\|(\widehat{\beta}_R - \beta^0) \cdot X\right\|$ asymptotically, and we expect the extra $\widehat{F}_R$ to be dominated by e and thus the main result to hold. A key step in the derivation of the main result is therefore to show faster than $\sqrt{N}$ convergence of $\widehat{\beta}_R$. Conversely, we expect counter examples to the main result to be such that the convergence rate of the estimator $\widehat{\beta}_R$ is not faster than $\sqrt{N}$, and we provide such a counter example – which, however, violates Assumptions LL – in Section 4.3 below. Whether the nice OLS intuition about "inclusion of irrelevant regressors" carries over to the "inclusion of irrelevant factors" thus crucially depends on the convergence rate of $\widehat{\beta}_R$.

4 Asymptotic Theory of $\widehat{\beta}_R$

In this section we introduce the key intermediate results that lead to the main Theorem 3.1 stated above. These intermediate results may be useful independently of the main result, e.g. Moon and Weidner (2010) and Moon, Shum, and Weidner (2012) use the results established here for the case of known $R = R^0$. The assumptions SN, EX, DX-1 and EV that are sequentially introduced in this Section are all implied by the low-level Assumptions LL above, see Lemma 4.7 below.

4.1 Consistency of $\widehat{\beta}_R$

Here we present a consistency result for $\widehat{\beta}_R$ under an arbitrary asymptotic $N, T \to \infty$, i.e. without the assumption that N and T grow at the same rate, which is imposed everywhere else in the paper. In addition to Assumption NC we require the following high level assumptions to obtain the result.

Assumption SN (Spectral Norm of X_k and e).

(i) $\|X_k\| = \mathcal{O}_P(\sqrt{NT})$, $\quad k = 1, \ldots, K$,

(ii) $\|e\| = \mathcal{O}_P(\sqrt{\max(N,T)})$.

Assumption EX (Weak Exogeneity of X_k). $\frac{1}{\sqrt{NT}} \mathrm{Tr}(X_k e') = \mathcal{O}_P(1)$, $\quad k = 1, \ldots, K$.

Theorem 4.1. *Let Assumptions SN, EX and NC be satisfied and let $R \geq R^0$. For $N, T \to \infty$ we then have* $\sqrt{\min(N,T)}\left(\widehat{\beta}_R - \beta^0\right) = \mathcal{O}_P(1)$.

Remarks

(i) One can justify Assumption SN(i) by use of the norm inequality $\|X_k\| \leq \|X_k\|_{HS}$ and the fact that $\|X_k\|_{HS}^2 = \sum_{i,t} X_{k,it}^2 = \mathcal{O}_P(NT)$, where the last step follows e.g. if $X_{k,it}$ has a uniformly bounded second moment.

(ii) Assumption SN(ii) is a condition on the largest eigenvalue of the random covariance matrix $e'e$, which is often studied in the literature on random matrix theory, e.g. Geman (1980), Bai, Silverstein, Yin (1988), Yin, Bai, and Krishnaiah (1988), Silverstein (1989). The results in Latala (2005) show that $\|e\| = \mathcal{O}_P(\sqrt{\max(N,T)})$ if e has independent entries with mean zero and uniformly bounded fourth moment. Weak dependence of the entries e_{it} across i and over t is also permissible, see Appendix A.1

(iii) Assumption EX requires exogeneity of the regressors X_k that is not necessarily strict and some weak dependence of $X_{k,it}e_{it}$ across i and over t.[15]

(iv) The theorem imposes no restriction at all on f^0 and λ^0, apart from the condition $R \geq \mathrm{rank}(\lambda^0 f^{0\prime})$.[16] In particular, the strong factor Assumption SF is not imposed here, i.e. consistency of $\widehat{\beta}_R$ holds independently of whether the factors are strong, weak, or not present at all. This is a remarkable robustness result, which is new in the literature.

(v) Under an asymptotic where N and T grow at the same rate, which is imposed everywhere else in the paper, Theorem 4.1 shows $\sqrt{N}$ (or equivalently $\sqrt{T}$) consistency of the estimator $\widehat{\beta}_R$.

4.2 Quadratic Approximation of $\mathcal{L}^0_{NT}(\beta) (\equiv \mathcal{L}^{R^0}_{NT}(\beta))$

To derive the limiting distribution of $\widehat{\beta}_R$, we study the asymptotic properties of the profile objective function $\mathcal{L}^R_{NT}(\beta)$ around β^0. The expression in (3.3) cannot easily be discussed by analytic means, since no explicit formula for the eigenvalues of a matrix is available. In particular, a standard Taylor expansion of $\mathcal{L}^R_{NT}(\beta)$ around β^0 cannot easily be derived. Here, we consider the case of known $R = R^0$ and we perform a joint expansion of the corresponding profile objective function $\mathcal{L}^0_{NT}(\beta)$ in the regression parameters β and in the idiosyncratic error terms e. To perform this joint expansion we apply the perturbation theory of linear operators (e.g., Kato (1980)). We thereby obtain an approximate quadratic expansion of $\mathcal{L}^0_{NT}(\beta)$ in β, which can be used to derive the first order asymptotic theory

[15] Note that $\frac{1}{\sqrt{NT}}\mathrm{Tr}(X_k e') = \frac{1}{\sqrt{NT}}\sum_i \sum_t X_{k,it}e_{it}$.

[16] This is the main reason why we use a slightly different non-collinearity Assumption NC, which avoids mentioning λ^0, compared to Bai (2009b).

of the LS estimator $\widehat{\beta}_{R^0}$, see also Appendix A.2. Define

$$
\begin{aligned}
W_{k_1 k_2} &= \frac{1}{NT}\, \mathrm{Tr}(M_{\lambda^0}\, X_{k_1}\, M_{f^0}\, X'_{k_2})\ , \\
C^{(1)}_k &= \frac{1}{\sqrt{NT}}\, \mathrm{Tr}(M_{\lambda^0}\, X_k\, M_{f^0}\, e')\ , \\
C^{(2)}_k &= -\frac{1}{\sqrt{NT}} \Big[\mathrm{Tr}\left(e M_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) \\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, e\, M_{f^0}\, X'_k\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \Big] . \qquad (4.1)
\end{aligned}
$$

Let W be the $K \times K$ matrix with elements $W_{k_1 k_2}$, and let $C^{(1)}$ and $C^{(2)}$ be K-vectors with elements $C^{(1)}_k$ and $C^{(2)}_k$, respectively.

Theorem 4.2. *Let Assumptions SF and SN be satisfied. Suppose that $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Then we have*

$$
\mathcal{L}^0_{NT}(\beta) = \mathcal{L}^0_{NT}(\beta^0) - \tfrac{2}{\sqrt{NT}} \left(\beta - \beta^0\right)' \left(C^{(1)} + C^{(2)}\right) + \left(\beta - \beta^0\right)' W \left(\beta - \beta^0\right) + \mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta),
$$

where the remainder term $\mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta)$ satisfies for any sequence $c_{NT} \to 0$

$$
\sup_{\{\beta : \|\beta - \beta^0\| \le c_{NT}\}} \frac{\left|\mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta)\right|}{\left(1 + \sqrt{NT}\, \|\beta - \beta^0\|\right)^2} = o_p\left(\frac{1}{NT}\right).
$$

The bound on remainder[17] in Theorem 4.2 is such that it has no effect on the first order asymptotic theory of $\widehat{\beta}_{R^0}$, as stated in the following corollary (see also Andrews (1999)).

Corollary 4.3. *Let Assumptions SF, SN, EX and NC be satisfied. In the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have $\sqrt{NT}\left(\widehat{\beta}_{R^0} - \beta^0\right) = W^{-1}\left(C^{(1)} + C^{(2)}\right) + o_P\left(1 + \|C^{(1)}\|\right)$. If we furthermore assume that $C^{(1)} = \mathcal{O}_P(1)$, then we obtain*

$$
\sqrt{NT}\left(\widehat{\beta}_{R^0} - \beta^0\right) = W^{-1}\left(C^{(1)} + C^{(2)}\right) + o_P(1) = \mathcal{O}_P(1).
$$

Note that our assumptions already guarantee $C^{(2)} = \mathcal{O}_P(1)$ and that W is invertible with $W^{-1} = \mathcal{O}_P(1)$, so this need not be explicitly assumed in Corollary 4.3.

Remarks

(i) Corollary 4.3 allows to replicate the result in Bai (2009b) on the asymptotic distri-

[17] The expansion in Theorem 4.2 contains a term that is linear in β and linear in e ($C^{(1)}$ term), a term that is linear in β and quadratic in e ($C^{(2)}$ term), and a term that is quadratic in β (W term). All higher order terms of the expansion are contained in the remainder term $\mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta)$.

bution of $\widehat{\beta}_{R^0}$.[18] Furthermore, the assumptions in the corollary do not restrict the regressor to be strictly exogenous and do not impose Assumption LL. The techniques developed here are applied in Moon and Weidner (2010) to discuss pre-determined regressors in the linear factor regression model with $R = R^0$.

(ii) If one weakens Assumption SN(ii) to $\|e\| = o_P(N^{2/3})$, then Theorem 4.2 still continues to hold. If we assume that $C^{(2)} = \mathcal{O}_P(1)$, then Corollary 4.3 also holds under this weaker condition on $\|e\|$.

More details on the expansion of $\mathcal{L}^0_{NT}(\beta)$ are provided in Appendix A.2 and the formal proofs can be found in in Section S.2 of the supplementary appendix.

4.3 Remarks On the Convergence Rate of $\widehat{\beta}_R$ for $R > R^0$

The results in Bai (2009b) and Corollary 4.3 above show that under appropriate assumptions the estimator $\widehat{\beta}_R$ is $\sqrt{NT}$-consistent for $R = R^0$. For $R > R^0$ we know from Theorem 4.1 that $\widehat{\beta}_R$ is $\sqrt{N}$ consistent as N and T grow at the same rate, but we have not shown faster than $\sqrt{N}$ converge of $\widehat{\beta}_R$ for $R > R^0$, yet, which according to the heuristic discussion at the end of Section 3 is a very important step.[19] However, one might not obtain a faster than $\sqrt{N}$ convergence rate of $\widehat{\beta}_R$ for $R > R^0$ without imposing further restrictions, as the following example shows:

Proposition 4.4. *Let $R^0 = 0$ (no true factors) and $K = 1$ (one regressor). The model reads $Y_{it} = \beta^0 X_{it} + e_{it}$, and we consider the following data generating process (DGP)*

$$X_{it} = a\widetilde{X}_{it} + \lambda_{x,i} f_{x,t}, \qquad e = \left(\mathbb{1}_N + c\,\frac{\lambda_x\lambda_x'}{N}\right) u \left(\mathbb{1}_T + c\,\frac{f_x f_x'}{T}\right),$$

where e and u are $N \times T$ matrices with entries e_{it} and u_{it}, respectively, and λ_x is an N-vector with entries $\lambda_{x,i}$, and f_x is a T-vector with entries $f_{x,t}$. Let $\widetilde{X}_{it}$ and u_{it} be mutually independent iid standard normally distributed random variables. Let $\lambda_{x,i} \in \mathcal{B}$ and $f_{x,t} \in \mathcal{B}$ be non-random sequences with bounded range $\mathcal{B} \subset \mathbb{R}$ such that $\frac{1}{N}\sum_{i=1}^N \lambda_{x,i}^2 \to 1$ and $\frac{1}{T}\sum_{t=1}^T f_{x,t}^2 \to 1$ asymptotically.[20] *Consider $N, T \to \infty$ such that $N/T \to \kappa^2$, $0 < \kappa < \infty$,*

[18] Let ρ, $D(.)$, D_0, D_Z, B_0 and C_0 be the notation used in Assumption A and Theorem 3 of Bai (2009b), and let Bai's assumptions be satisfied. Then, our κ, W, $C^{(1)}$ and $C^{(2)}$ satisfy $\kappa = \rho^{-1/2}$, $W = D(f^0) \to_p D > 0$, $C^{(1)} \to_d \mathcal{N}(0, D_Z)$ and $W^{-1}C^{(2)} \to_p \rho^{1/2}B_0 + \rho^{-1/2}C_0$. Corollary 4.3 can therefore be used to replicate Theorem 3 in Bai (2009b). For more details and extensions of this we refer to Moon and Weidner (2010).

[19] One reason why $\widehat{\beta}_R$ might only converge at $\sqrt{N}$ rate, but not faster, are weak factors (both for $R > R^0$ and for $R = R^0$). A weak factor (see e.g. Onatski (2012) and Chudik, Pesaran and Tosetti (2011)) might not be picked up at all or might only be estimated very inaccurately by the principal components estimator $\widehat{F}_R$, in which case that factor is not properly accounted for in the LS estimation procedure. If this happens and the weak factor is correlated with the regressors, then there is some uncorrected weak endogeneity problem, and $\widehat{\beta}_R$ will only converge at $\sqrt{N}$ rate. The problem will become more severe if there are many weak factors, but we restrict attention to $R^0 \leq R$ here, in which case weak factors can only "spoil" the convergence rate of $\widehat{\beta}_R$, but not its consistency. We do not consider the issue of weak factors any further in this paper.

[20] The proposition also holds if λ_x and f_x are random (but independent of e and $\widetilde{X}$) and also if the range $\mathcal{B}$ is unbounded. We assume non-random λ_x and f_x to guarantee that the DGP satisfies Assumption D of Bai (2009b), namely that X and e are independent (otherwise we only have mean-independence, i.e. $\mathbb{E}(e|X) = 0$). Similarly, we assume bounded $\mathcal{B}$ to satisfy the restrictions on e_{it} imposed in Assumption C of Bai (2009b).

and let $0 < a < (1/2)^{2/3} \min(\kappa^2, \kappa^{-2})$ *and* $c \geq \frac{(2+\sqrt{2})(1+\kappa)(1+\sqrt{3}a^{-1/4})}{\min(1,\kappa)[1/2 - a^{3/2}\max(\kappa,\kappa^{-1})]}$.[21] *Let* $\mathcal{L}^1_{NT}(\beta)$ *be the profile objective function for* $R = 1$*, defined in* (3.3)*. Then, for any sequence* $\Delta_{NT} > 0$ *with* $\Delta_{NT} = o(N^{-1/2})$ *we have*

$$\min_{\beta \in \mathbb{R}} \mathcal{L}^1_{NT}(\beta) < \min_{\beta \in [\beta^0 - \Delta_{NT}, \beta^0 + \Delta_{NT}]} \mathcal{L}^1_{NT}(\beta), \qquad \textit{wpa1}.$$

This implies that $\left\| \widehat{\beta}_1 - \beta^0 \right\|$ *cannot converge to zero at a faster than* $\sqrt{N}$ *rate.*

Remarks

(i) In Proposition 4.4 we have $R^0 = 0$, i.e. the estimator $\widehat{\beta}_0$ based on the correct number of factors is just the OLS estimator, which is $\sqrt{NT}$ consistent for this DGP. However, for $R = 1$ the proposition together with Theorem 4.1 shows that $\widehat{\beta}_1$ only converges at a $\sqrt{N}$ rate to the true parameter β^0, i.e. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are certainly not asymptotically equivalent here.

(ii) The DGP in Proposition 4.4 satisfies all the assumptions imposed in Corollary 4.3 to derive the limiting distribution of the LS-estimator for $R = R^0$. It also satisfies all the regularity conditions imposed in Bai (2009b) — see Section S.8 in the supplementary material for more details.

(iii) The regressor X_{it} in Proposition 4.4 is strictly exogenous and it is a "high-rank regressor" that satisfied the generalized non-collinearity Assumption NC for any values of R and R^0. The errors e_{it} are mean zero and correlated across i and over t, namely $\mathbb{E}(e_{it}e_{js}) = \Sigma_{ij}\Omega_{ts}$, where $\Sigma = \left(\mathbb{1}_N + c\,\frac{\lambda_x \lambda_x'}{N}\right)^2$ and $\Omega = \left(\mathbb{1}_T + c\,\frac{f_x f_x'}{T}\right)^2$. The cross-sectional and time-serial covariance matrices Σ and Ω have diagonal elements of order one and off-diagonal elements of order $1/N$ and $1/T$, respectively, i.e. both types of correlation become weak as N and T become large.

(iv) The aspect that is special about this DGP is that λ_x and f_x feature both in X_{it} and in the second moment structure of e_{it}, characterized by Σ and Ω. The heuristic discussion at the end of Section 3 provides some intuition why this can be problematic, because the leading principal components obtained from only the error matrix e will have a strong sample correlation with X_{it} for this DGP. It appears rather unlikely to us that something like this would happen in a practical application, and we believe that the example in Proposition 4.4 is really quite artificial in that regard. From a theoretical perspective, however, the example is quite instructive, since it shows that the main result in Theorem 3.1 does not necessarily hold under the assumptions we have imposed so far in this section, and that stronger assumptions are necessary to study $\widehat{\beta}_R$ for $R > R^0$ than for $R = R^0$.

4.4 $N^{3/4}$-Convergence Rate of $\widehat{\beta}_R$ for $R > R^0$

The discussion at the end of Section 3 and in the last subsection reveals that showing faster than $\sqrt{N}$ convergence of $\widehat{\beta}_R$ is a very important step on the way to the main result.

[21] The bounds on the constants a and c imposed in the proposition are sufficient, but not necessary. Simulation evidence suggests that the result in the proposition holds for a much larger range of a, c values.

For purely technical reasons we show $N^{3/4}$-convergence first, but it will usually be the case that if $\widehat{\beta}_R$ is $N^{3/4}$-consistent, then it is also $\sqrt{NT}$-consistent as N and T grow at the same rate. We require one of the following two alternative assumptions.

Assumption DX-1 (Decomposition of X_k and Distribution of e_{it}, Version 1).

(i) *For $k = 1, \ldots, K$ we have $X_k = \overline{X}_k + \widetilde{X}_k$, where $\text{rank}(\overline{X}_k)$ is bounded as $N, T \to \infty$, and $\|\overline{X}_k\| = \mathcal{O}_P(\sqrt{NT})$, and $\|\widetilde{X}_k\| = \mathcal{O}_P(N^{3/4})$.*

(ii) *Let u be an $N \times T$ matrix whose elements are distributed as i.i.d. $\mathcal{N}(0,1)$, independent of λ^0, f^0 and $\overline{X}_k$, $k = 1, \ldots, K$, and let one of the following hold*

(a) *either: $e = \Sigma^{1/2} u$, where Σ is an $N \times N$ covariance matrix, independent of u, which satisfies $\|\Sigma\| = \mathcal{O}_P(1)$. In that case, define g to be an $N \times Q$ matrix, independent of u, for some $Q \leq \sum_{k=1}^{K} \text{rank}(\overline{X}_k)$, such that $g'g = \mathbb{1}_Q$ and $\text{span}(M_{\lambda^0}\overline{X}_k) \subset \text{span}(g)$ for all $k = 1, \ldots, K$.*[22]

(b) *or: $e = u\,\Sigma^{1/2}$, where Σ is a $T \times T$ covariance matrix, independent of u, which satisfies $\|\Sigma\| = \mathcal{O}_P(1)$. In that case, define g to be a $T \times Q$ matrix, independent of u, for some $Q \leq \sum_{k=1}^{K} \text{rank}(\overline{X}_k)$, such that $g'g = \mathbb{1}_Q$ and $\text{span}(M_{f^0}\overline{X}'_k) \subset \text{span}(g)$ for all $k = 1, \ldots, K$.*

In addition, we assume that there exist a (potentially random) integer sequence $n = n_{NT} > 0$ with $1/n = \mathcal{O}_P(1/N)$ such that $\mu_n(\Sigma) \geq \|g'\Sigma g\|$. Finally, assume that either $R \geq Q$ or that $g'\Sigma g = \|g'\Sigma g\| \mathbb{1}_Q + \mathcal{O}_P(N^{-1/2})$.

Assumption DX-2 (Decomposition of X_k and Distribution of e_{it}, Version 2).

(i) *For $k = 1, \ldots, K$ we have $X_k = \overline{X}_k + \widetilde{X}_k$, such that $M_{\lambda^0}\overline{X}_k M_{f^0} = 0$, and $\|\overline{X}_k\| = \mathcal{O}_P(\sqrt{NT})$, and $\|\widetilde{X}_k\| = \mathcal{O}_P(N^{3/4})$.*

(ii) *$\|e\| = \mathcal{O}_P(\sqrt{\max(N,T)})$. (same as Assumption SN(ii))*

Theorem 4.5. *Let $R > R^0$. Let Assumptions SF, NC and EX hold, and let either Assumption DX-1 or DX-2 be satisfied. Consider $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Then we have $N^{3/4}\left(\widehat{\beta}_R - \beta^0\right) = \mathcal{O}_P(1)$.*

Remarks

(i) Assumption SN is not explicitly imposed in Theorem 4.5, because it is already implied by both Assumption DX-1 and DX-2, see also Lemma 4.7 below.

(ii) The restrictions that Assumption DX-1 imposes on X_k are weaker than those imposed in Assumption LL above. The regressors are decomposed into a low-rank strictly exogenous part $\overline{X}_k$ and a term $\widetilde{X}_k$, which can be both strictly or weakly exogenous. The spectral norm bound $\|\widetilde{X}_k\| = \mathcal{O}_P(N^{3/4})$ is satisfied as long as $\widetilde{X}_{k,it}$ is mean zero

[22] The column space of g thus contains the column space of all $M_{\lambda^0}\overline{X}_k$. $g'g = \mathbb{1}_Q$ is just a normalization.

and weakly correlated across i and over t, see Appendix A.1. We can always write $\overline{X}_k = \ell h'$ for some appropriate $\ell \in \mathbb{R}^{N \times \mathrm{rank}(\overline{X}_k)}$ and $h \in \mathbb{R}^{T \times \mathrm{rank}(\overline{X}_k)}$. Thus, the decomposition $X_k = \overline{X}_k + \widetilde{X}_k = \ell h' + \widetilde{X}_k$ essentially imposes an approximate factor structure on X_k, with factor part $\overline{X}_k$ and idiosyncratic part $\widetilde{X}_k$.

(iii) The restrictions that Assumption DX-1 imposes on e are also weaker than those imposed in Assumption LL above. Normality is imposed, but either cross-sectional correlation and heteroscedasticity (case (a)) or time-serial correlation and heteroscedasticity (case (b)), described by Σ, are still allowed. The condition $\|\Sigma\| = \mathcal{O}_P(1)$ requires the correlation of e_{it} to be weak.[23]

(iv) The additional restrictions on Σ in Assumption DX-1 rule out the type of correlation of the low-rank regressor part $\overline{X}_k$ with the second moment structure of e_{it} that was the key feature of the counter example in Proposition 4.4 above.[24] Firstly, the condition $\mu_n(\Sigma) \geq \|g'\Sigma g\|$ guarantees that the eigenvectors corresponding to the largest few eigenvectors of Σ (the eigenvectors ν_r of Σ when normalized satisfy $\mu_r(\Sigma) = \nu_r'\Sigma\nu_r$) are not strongly correlated with g (and thus with $\overline{X}_k$). Secondly, the condition $g'\Sigma g = \|g'\Sigma g\| \mathbb{1}_Q + \mathcal{O}_P(N^{-1/2})$ guarantees that Σ behaves almost as an identity matrix when projected with g, thus not possessing special structure in the "direction of $\overline{X}_k$". Both of these assumption are obviously satisfied when Σ is proportional to the identity matrix.

(v) Instead of Assumption DX-1 we can also impose Assumption DX-2 to obtain $N^{3/4}$-consistency in Theorem 4.5. The Assumption on e imposed in Assumption DX-2 is the same as in Assumption SN, and as already discussed above, this assumption is quite weak (see also Appendix A.1). However, Assumption DX-2 imposes a much stronger assumption on the regressors by requiring that $M_{\lambda^0}\overline{X}_k M_{f^0} = 0$. This condition implies that $\overline{X}_k = \lambda^0 h' + \ell f^{0\prime}$ for some $\ell \in \mathbb{R}^{N \times R^0}$ and $h \in \mathbb{R}^{T \times R^0}$, i.e. the factor structure of the regressors is severely restricted. The AR(1) model discussed in Remark (v) of Section 3 does satisfy $M_{\lambda^0}\overline{X}_k = 0$, and the same is true for a stationary AR(p) model without additional regressors, i.e. for such AR(p) models with factors we obtain $N^{3/4}$-consistency of $\widehat{\beta}_R$ without imposing strong assumptions (like normality) of e_{it}. Assumption DX-2(i) is furthermore satisfied if $\overline{X}_k = 0$, i.e. if the regressors $X_k = \widetilde{X}_k$ satisfy $\|X_k\| = \mathcal{O}_P(N^{3/4})$, which is true for zero mean weakly correlated processes (see Appendix A.1).

(vi) Theorem S.5 in the supplementary material provides an alternative $N^{3/4}$-consistency result, in which Assumptions DX-1 and DX-2 are replaced by a high-level condition, which is more general, but not easy to verify in terms of low-level assumptions.

4.5 Asymptotic Equivalence of $\widehat{\beta}_{R^0}$ and $\widehat{\beta}_R$ for $R > R^0$

In this section, we provide high level conditions on the singular values and singular vectors of the error matrix (or equivalently on the eigenvalues and eigenvectors of the corresponding random covariance matrix). Under those assumptions we then establish the main result of

[23] A sufficient condition for $\|\Sigma\| = \mathcal{O}_P(1)$ is, for example, $\max_i \sum_j |\Sigma_{ij}| = \mathcal{O}_P(1)$, formulated here for case (a). Note that Σ is symmetric.

[24] However, in the example of Proposition 4.4 we have both time-serial and cross-sectional correlation in e_{it}, one of which is already ruled out by Assumption DX-1.

the paper that $\widehat{\beta}_{R^0}$ and $\widehat{\beta}_R$ with $R > R^0$ are asymptotically equivalent, that is, $\sqrt{NT}(\widehat{\beta}_R - \widehat{\beta}_{R^0}) = o_P(1)$.

Assumption EV. (Eigenvalues and Eigenvectors of Random Cov. Matrix) *Let the singular value decomposition of $M_{\lambda^0} e M_{f^0}$ be given by $M_{\lambda^0} e M_{f^0} = \sum_{r=1}^{Q} \sqrt{\rho_r}\, v_r\, w_r'$, where $Q = \min(N,T) - R^0$, and $\sqrt{\rho_r}$ are the singular values, and v_r and w_r are normalized N- and T-vectors, respectively.*[25] *Let $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_Q \geq 0$. We assume that there exists a constant $c > 0$ and a series of integers $q_{NT} > R - R^0$ with $q_{NT} = o(N^{1/4})$ such that as $N, T \to \infty$ we have*

(i) $\dfrac{\rho_{R-R^0}}{N} > c$, *wpa1.*

(ii) $\dfrac{1}{q_{NT}} \displaystyle\sum_{r=q_{NT}}^{Q} \dfrac{1}{\rho_{R-R^0} - \rho_r} = \mathcal{O}_P(1)$.

(iii)
$$\max_r \|v_r' e P_{f^0}\| = o_P\left(N^{1/4}\, q_{NT}^{-1}\right), \qquad \max_r \|w_r' e' P_{\lambda^0}\| = o_P\left(N^{1/4}\, q_{NT}^{-1}\right),$$
$$\max_r \|v_r' X_k P_{f^0}\| = o_P\left(N\, q_{NT}^{-1}\right), \qquad \max_r \|w_r' X_k' P_{\lambda^0}\| = o_P\left(N\, q_{NT}^{-1}\right),$$
$$\max_{r,s,k} |v_r' X_k w_s| = o_P\left(N^{1/4}\, q_{NT}^{-1}\right), \qquad \text{where } r, s = 1, \ldots, Q, \text{ and } k = 1, \ldots, K.$$

Theorem 4.6. *Let $R > R^0$. Let Assumptions SF, NC, EX, and EV hold, and let either Assumption DX-1 or DX-2 hold, and assume that $C^{(1)} = \mathcal{O}_P(1)$. In the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have*

$$\sqrt{NT}\left(\widehat{\beta}_R - \beta^0\right) = \sqrt{NT}\left(\widehat{\beta}_{R^0} - \beta^0\right) + o_P(1) = \mathcal{O}_P(1).$$

Remarks

(i) Theorem 4.6 also holds if we replace the Assumptions EX, DX-1, DX-2 by any other condition that guarantees that Assumption SN holds and that $N^{3/4}\left(\widehat{\beta}_R - \beta^0\right) = \mathcal{O}_P(1)$.

(ii) Consider Assumption EV(iii). Since v_r and w_r are the normalized singular vectors of $M_{\lambda^0} e M_{f^0}$ we expect them to be essentially uncorrelated with X_k and eP_{f^0}, and therefore we expect $v_r' X_k w_s = \mathcal{O}_P(1)$, $\|v_r' e P_{f^0}\| = \mathcal{O}_P(1)$, $\|w_r' e' P_{\lambda^0}\| = \mathcal{O}_P(1)$. We also expect $\|v_r' X_k P_{f^0}\| = \mathcal{O}_P(\sqrt{T})$ and $\|w_r' X_k' P_{\lambda^0}\| = \mathcal{O}_P(\sqrt{N})$, which is different to the analogous expressions with e, since X_k can be correlated with f^0 and λ^0. The key to making this discussion rigorous is a good knowledge of the properties of the eigenvectors v_r and w_r. If the entries e_{it} are iid normal, then the distribution of v_r and w_r can be characterized as follows: Let $\widetilde{v}$ be an N-vector with $iid\mathcal{N}(0,1)$ entries and let $\widetilde{w}$ be an T-vector with $iid\mathcal{N}(0,1)$ entries. Then we have $v_r =_d \|M_{\lambda^0}\widetilde{v}\|^{-1} M_{\lambda^0}\widetilde{v}$ and $w_r =_d \|M_{f^0}\widetilde{w}\|^{-1} M_{f^0}\widetilde{w}$, see also Lemma S.13 in the supplementary material.

[25] Thus, w_r is the normalized eigenvector corresponding to the eigenvalue ρ_r of $M_{f^0} e' M_{\lambda^0} e M_{f^0}$, while v_r is the normalized eigenvector corresponding to the eigenvalue ρ_r of $M_{\lambda^0} e M_{f^0} e' M_{\lambda^0}$. We use a convention were eigenvalues with non-trivial multiplicity appear multiple times in the list of eigenvalues ρ_r, but under standard distributional assumptions on e all eigenvalues are simple with probability one anyways.

Here, $=_d$ refers to "equal in distribution". Thus, if $R^0 = 0$, then v_r and w_r are distributed as $iid\mathcal{N}(0,1)$ vectors, normalized to satisfy $\|v_r\| = \|w_r\| = 1$. This follows from the rotational invariance of the distribution of e when e_{it} is *iid* normally distributed. Using this characterization of v_r and w_r one can formally show that Assumption EV(*iii*) holds, see Lemma 4.7 below. The conjecture in the random matrix theory literature is that the limiting distribution of the eigenvectors of a random covariance matrix is "distribution free", i.e. is independent of the particular distribution of e_{it} (see, e.g., Silverstein (1990), Bai (1999)). However, we are not aware of a formulation and corresponding proof of this conjecture that is sufficient for our purposes, which is one reason why we have to impose *iid* normality of e_{it}.

(iii) Assumption EV(*ii*) imposes a condition on the eigenvalues ρ_r of the random covariance matrix $M_{f^0}e'M_{\lambda^0}eM_{f^0}$. Eigenvalues are studied more intensely than eigenvectors in the random matrix theory literature, and it is well-known that the properly normalized empirical distribution of the eigenvalues (the so called empirical spectral distribution) of an *iid* sample covariance matrix converges to the Marčenko-Pastur-law (Marčenko and Pastur (1967)) for asymptotics where N and T grow at the same rate. This means that the sum over the function of the eigenvalues ρ_s in Assumption EV(*ii*) can be approximated by an integral over the Marčenko-Pastur limiting spectral distribution. To bound the asymptotic error of this approximation one needs to know the convergence rate of the empirical spectral distribution to its limit law, which is an ongoing research subject in the literature, e.g. Bai (1993), Bai, Miao and Yao (2004), Götze and Tikhomirov (2010). This literature usually considers either *iid* or *iid* normal distributions of e_{it}.

(iv) For random covariance matrices from *iid* normal errors, it is known from Johnstone (2001) and Soshnikov (2002) that the properly normalized few largest eigenvalues converge to the Tracy-Widom law.[26] This result can be used to verify Assumption EV(*i*) in the case of *iid* normal e_{it}.

(v) An additional subtlety when justifying Assumption EV is that we consider the eigenvalues and eigenvectors of the random covariance matrix $M_{f^0}e'M_{\lambda^0}eM_{f^0}$, not just $e'e$. The additional projections with M_{f^0} and M_{λ^0} stem from integrating out the true factors and factor loadings of the model. Those projectors are not normally present in the literature on large dimensional random covariance matrices. If the idiosyncratic error distribution is *iid* normal these projections are unproblematic, since the distribution of e is rotationally invariant from the left and right in that case, i.e. the projections are mathematically equivalent to a reduction of the sample size by R^0 in both directions.

(vi) Details on how to derive Theorem 4.6 are given in Section S.4 of the supplementary material.

The following Lemma provides the connection between Theorem 4.6 and our main result Theorem 3.1.

Lemma 4.7. *Let Assumption LL hold, let* $R^0 = \text{rank}(\lambda^0) = \text{rank}(f^0)$, *and consider a limit* $N, T \to \infty$ *with* $N/T \to \kappa^2$, $0 < \kappa < \infty$. *Then Assumptions SN, EX, DX-1 and EV are satisfied, and we have* $C^{(1)} = \mathcal{O}_P(1)$.

[26] To our knowledge this result is not established for error distributions that are not normal. Soshnikov (2002) has a result under non-normality but only for asymptotics with $N/T \to 1$.

5 Monte Carlo Simulations

In this section we investigate the finite sample properties of $\widehat{\beta}_R$ through a small scale Monte Carlo simulation. The model is a static panel model with one regressor ($K = 1$), two factors ($R^0 = 2$), and the following data generating process (DGP):

$$\begin{aligned} Y_{it} &= \beta^0 X_{it} + \sum_{r=1}^{2} \lambda_{ir} f_{tr} + e_{it}, \\ X_{it} &= 1 + \widetilde{X}_{it} + \sum_{r=1}^{2} (\lambda_{ir} + \chi_{ir})(f_{tr} + f_{t-1,r}), \\ e_{it} &= \frac{1}{\sqrt{2}}(v_{it} + v_{i,t-1}). \end{aligned} \tag{5.1}$$

The random variables $\widetilde{X}_{it}$, λ_{ir}, f_{tr}, χ_{ir} and v_{it} are mutually independent; with $\widetilde{X}_{it}$ and $f_{tr} \sim iid\mathcal{N}(0,1)$; λ_{ir} and $\chi_{ir} \sim iid\mathcal{N}(1,1)$; and $v_{it} \sim iid\, t(5)$, i.e. v_{it} has a Student's t-distribution with 5 degrees of freedom.

Note that this model satisfies Assumptions SF, NC, and LL(i), but not LL(ii). The error term e_{it} is *not* distributed as *iid* normal. The time series of e_{it} is serially correlated and generated by the sum of two independent random variables whose distribution is $t(5)$. The purpose of this design is to demonstrate that the iid normality restriction on e_{it} in Assumption LL(ii) is a technical assumption as mentioned in Section 2 and may be relaxed.

We choose $\beta^0 = 1$, and use $10,000$ repetitions in our simulation. The true number of factors is chosen to be $R^0 = 2$. For each draw of Y and X we compute the LS estimator $\widehat{\beta}_R$ according to equation (3.1) for different values of R, namely $R \in \{0, 1, 2, 3, 4, 5\}$.

Table 1 reports bias and standard deviation of the estimator $\widehat{\beta}_R$ for different combinations of R, N and T. For $R < R^0 = 2$ the model is mis-specified and $\widehat{\beta}_R$ turns out to be severely biased. There is also bias in $\widehat{\beta}_R$ for $R \geq R^0$, due to time-serial correlation of e_{it}. This bias was discussed in Bai (2009b), and bias correction is also discussed there. We have purposefully chosen a DGP where $\widehat{\beta}_R$ exhibits such a bias to illustrate that all features of the asymptotic distribution of $\widehat{\beta}_{R^0}$ are replicated by $\widehat{\beta}_R$, $R > R^0$, including the bias.

Table 2 reports various quantiles of the distribution of $\sqrt{NT}(\widehat{\beta}_R - \beta^0)$ for $N = T = 100$ and $N = T = 300$, and different values of $R \geq R^0$. From these tables, we see that as N, T increases the distribution of $\widehat{\beta}_R$ gets closer to that of $\widehat{\beta}_{R^0}$.

Monte Carlo Simulation results for an AR(1) model with factors can be found in Section S.6 of the supplementary material. Those additional simulations show that the finite sample properties (e.g. for $T = 30$) of $\widehat{\beta}_{R^0}$ and $\widehat{\beta}_R$, $R > R^0$, can be quite different, but those differences vanish as T becomes large, as predicted by our asymptotic theory. In general, we always expect some finite sample inefficiency from overestimating the number of factors.

6 Conclusions

In this paper we showed that under certain regularity conditions the limiting distribution of the LS estimator of a linear panel regression with interactive fixed effects does not change

	N=100							
	T=10		T=30		T=100		T=300	
R	Bias	SD	Bias	SD	Bias	SD	Bias	SD
0	0.22860	0.03212	0.23005	0.01667	0.23052	0.01173	0.23053	0.01028
1	0.10607	0.05524	0.11548	0.02959	0.11906	0.01949	0.12001	0.01604
2	-0.03851	0.03425	-0.01664	0.01425	-0.00531	0.00711	-0.00191	0.00399
3	-0.04268	0.03417	-0.01702	0.01418	-0.00532	0.00716	-0.00191	0.00402
4	-0.04499	0.03563	-0.01723	0.01440	-0.00533	0.00720	-0.00192	0.00404
5	-0.04606	0.03698	-0.01751	0.01458	-0.00534	0.00727	-0.00192	0.00407

	N=300							
	T=10		T=30		T=100		T=300	
R	Bias	SD	Bias	SD	Bias	SD	Bias	SD
0	0.22909	0.02980	0.22994	0.01361	0.23063	0.00823	0.23067	0.00647
1	0.10539	0.05004	0.11586	0.02633	0.11927	0.01481	0.12029	0.01049
2	-0.04080	0.02366	-0.01723	0.00850	-0.00542	0.00406	-0.00181	0.00229
3	-0.04418	0.02444	-0.01754	0.00858	-0.00544	0.00409	-0.00182	0.00229
4	-0.04617	0.02577	-0.01789	0.00867	-0.00547	0.00410	-0.00182	0.00229
5	-0.04676	0.02748	-0.01815	0.00877	-0.00548	0.00412	-0.00182	0.00230

Table 1: For different combinations of sample sizes N and T we report the bias and standard deviation of the estimator $\widehat{\beta}_R$, for $R = 0, 1, \ldots, 5$, based on simulations with $10,000$ repetition of design (5.1), where the true number of factors is $R^0 = 2$. See the main text for further explanations.

				N=100	T=100				
R	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
2	-1.951	-1.704	-1.443	-1.012	-0.519	-0.049	0.355	0.614	0.858
3	-1.939	-1.725	-1.465	-1.007	-0.518	-0.047	0.375	0.638	0.869
4	-1.969	-1.729	-1.473	-1.009	-0.521	-0.044	0.387	0.635	0.850
5	-1.966	-1.745	-1.481	-1.017	-0.521	-0.041	0.388	0.645	0.876

				N=300	T=300				
R	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
2	-1.906	-1.678	-1.432	-0.999	-0.541	-0.070	0.331	0.565	0.781
3	-1.914	-1.679	-1.435	-1.000	-0.548	-0.076	0.321	0.568	0.781
4	-1.924	-1.681	-1.437	-1.001	-0.546	-0.076	0.328	0.566	0.782
5	-1.913	-1.684	-1.442	-1.001	-0.541	-0.074	0.335	0.582	0.790

Table 2: For $N = T = 100$ and $N = T = 300$ we report certain quantiles of the distribution of $\sqrt{NT}(\widehat{\beta}_R - \beta^0)$, for $R = 2, 3, 4, 5$, based on simulations with $10,000$ repetition of design (5.1), where the true number of factors is $R^0 = 2$. See the main text for further explanations.

when we include redundant factors in the estimation. The important empirical implication of this result is that one can use an upper bound of the number of factors in the estimation without asymptotic efficiency loss. We impose *iid* normality of the regression errors to derive this result, because we require certain results on the eigenvalues and eigenvectors of random covariance matrices that are only known in that case. We expect that progress in the literature on large dimensional random covariance matrices will allow verification of our high-level assumptions under more general error distributions, and our simulation results suggest that the result also holds for non-normal and correlated errors. We also provide multiple intermediate asymptotic results under more general conditions.

References

Ahn, S., Lee, Y., and Schmidt, P. (2007). Panel data models with multiple time-varying individual effects. *Journal of Productivity Analysis*.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101(2):219–255.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1384.

Bai, J. (2009a). Likelihood approach to small T dynamic panel models with interactive effects. Manuscript.

Bai, J. (2009b). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z. (1993). Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices. *The Annals of Probability*, 21(2):649–672.

Bai, Z. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, 9:611–677.

Bai, Z., Miao, B., and Yao, J. (2004). Convergence rates of spectral distributions of large sample covariance matrices. *SIAM journal on matrix analysis and applications*, 25(1):105–127.

Bai, Z. D., Silverstein, J. W., and Yin, Y. Q. (1988). A note on the largest eigenvalue of a large dimensional sample covariance matrix. *J. Multivar. Anal.*, 26(2):166–168.

Chudik, A., Pesaran, M. H., and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal*, 14(1):C45–C90.

Cox, D. D. and Kim, T. Y. (1995). Moment bounds for mixing random variables useful in nonparametric function estimation. *Stochastic processes and their applications*, 56(1):151–158.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Annals of Probability*, 8(2):252–261.

Götze, F. and Tikhomirov, A. (2010). The Rate of Convergence of Spectra of Sample Covariance Matrices. *Theory of Probability and its Applications*, 54:129.

Holtz-Eakin, D., Newey, W., and Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56(6):1371–95.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

Kato, T. (1980). *Perturbation Theory for Linear Operators*. Springer-Verlag.

Kiefer, N. (1980). A time series-cross section model with fixed effects with an intertemporal factor structure. *Unpublished manuscript, Department of Economics, Cornell University*.

Latala, R. (2005). Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133:1273–1282.

Lu, X. and Su, L. (2013). Shrinkage estimation of dynamic panel data models with interactive fixed effects. Technical report, Working paper, Hong Kong University of Science & Technology.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

Moon, H., Shum, M., and Weidner, M. (2012). Interactive fixed effects in the blp random coefficients demand model. *CeMMAP working paper series*.

Moon, H. and Weidner, M. (2010). Dynamic Linear Panel Regression Models with Interactive Fixed Effects. Manuscript.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Onatski, A. (2013). Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models. Manuscript.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Silverstein, J. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *The Annals of Probability*, 18(3):1174–1194.

Silverstein, J. W. (1989). On the eigenvectors of large dimensional sample covariance matrices. *J. Multivar. Anal.*, 30(1):1–16.

Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics*, 108(5):1033–1056.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probability Theory Related Fields*, 78:509–521.

Zaffaroni, P. (2009). Generalized least squares estimation of panel with common shocks. Manuscript.

A Appendix

A.1 Spectral Norm of Random Matrices

Consider an $N \times T$ matrix u whose entries u_{it} have uniformly bounded second moments. Then we have $\|u\| \leq \|u\|_{HS} = \sqrt{\sum_{i,t} u_{it}^2} = \mathcal{O}_P(\sqrt{NT})$. However, in Assumption LL($i.b$) and Assumption DX-1(i) and Assumption DX-2(i) we impose $\|\widetilde{X}_k^{str}\| = \mathcal{O}_P(N^{3/4})$ and $\|\widetilde{X}_k\| = \mathcal{O}_P(N^{3/4})$, respectively, as N and T grow at the same rate, and in Assumption SN(ii) we impose $\|e\| = \mathcal{O}_P(\sqrt{\max(N,T)})$ under an arbitrary asymptotic $N, T \to \infty$. Those smaller asymptotic rates for the spectral norms of $\widetilde{X}_k^{str}$, $\widetilde{X}_k$ and e can be justified by firstly assuming that the entries of these matrices are mean zero and have certain bounded moments, and secondly imposing weak cross-sectional and time-serial correlation. The purpose of this appendix section is to provide some examples of matrix distributions that make the last statement more precise. We consider the $N \times T$ matrix u, which can represent either e, $\widetilde{X}_k^{str}$ or $\widetilde{X}_k$.

Example 1: If we assume that $\mathbb{E}u_{it} = 0$, that $\mathbb{E}u_{it}^4$ is uniformly bounded, and that the u_{it} are independently distributed across i and over t, then the results in Latala (2005) show that $\|u\| = \mathcal{O}_P(\sqrt{\max(N,T)})$.

Example 2: Onatski (2013) provides the following example, which allows for both cross-sectional and time-serial dependence: Let ϵ be an $N \times T$ matrix with mean zero, independent entries that have uniformly bounded fourth moment, let ϵ_t denote the columns of ϵ, and also define past ϵ_t, $t \leq 0$, satisfying the same distributional assumptions. Let $u_t = \sum_{j=0}^{m} \Psi_{N,j} \epsilon_{t-j}$, where m is a fixed integer, and $\Psi_{N,j}$ are $N \times N$ matrices such that $\max_j \|\Psi_{N,j}\|$ is uniformly bounded. Then, the $N \times T$ matrix u with columns u_t satisfies $\|u\| = \mathcal{O}_P(\sqrt{\max(N,T)})$.

More examples of matrix distributions that satisfy $\|u\| = \mathcal{O}_P(\sqrt{\max(N,T)})$ are discussed in Onatski (2013) and Moon and Weidner (2010). Theorem 5.48 and Remark 5.49 in Vershynin (2010) can also be used to obtain a slightly weaker bound on $\|u\|$ under very general correlation of u in one of its dimensions.

Note that the random matrix theory literature often only discusses asymptotics where N and T grow at the same rate and shows $\|u\| = \mathcal{O}_P(\sqrt{N})$ under that asymptotic. Those results can easily be extended to more general asymptotics with $N, T \to \infty$ by considering u as a submatrix of a $\max(N,T) \times \max(N,T)$ matrix $u^{\rm big}$, and using that $\|u\| \leq \|u^{\rm big}\|$.

Example 3: The following Lemma provides a justification for the bounds on $\|\widetilde{X}_k^{str}\|$ and $\|\widetilde{X}_k\|$, allowing for a quite general type of correlation in both panel dimensions.

Lemma A.1. *Let u be an $N \times T$ matrix with entries u_{it}. Let $\Sigma_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(u_{it}u_{jt})$, and let Σ be the $N \times N$ matrix with entries Σ_{ij}. Let $\eta_{ij} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [u_{it}u_{jt} - \mathbb{E}(u_{it}u_{jt})]$, $\Psi_{ij} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}(\eta_{ik}\eta_{jk})$, and $\chi_{ij} = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} [\eta_{ik}\eta_{jk} - \mathbb{E}(\eta_{ik}\eta_{jk})]$. Consider an asymptotic where $N, T \to \infty$ such that N/T converges to a finite positive constant, and assume that*

(i) $\|\Sigma\| = \mathcal{O}(1)$.

(ii) $\frac{1}{N^2}\sum_{i,j=1}^{N} \mathbb{E}(\eta_{ij}^2) = \mathcal{O}(1)$.

(iii) $\frac{1}{N}\sum_{i,j=1}^{N} \Psi_{ij}^2 = \mathcal{O}(1)$.

(iv) $\frac{1}{N^2}\sum_{i,j=1}^{N} \mathbb{E}(\chi_{ij}^2) = \mathcal{O}(1)$.

Then we have $\|u\| = \mathcal{O}_P(N^{5/8})$.

The Lemma does not impose $\mathbb{E}u_{it} = 0$ explicitly, but justification of assumption (i) in the lemma usually requires $\mathbb{E}u_{it} = 0$. The assumptions (ii), (iii) and (iv) in the lemma can e.g. be justified by assuming appropriate mixing conditions in both panel dimensions, see e.g. Cox and Kim (1995) for the time-series case.

As pointed out above, our results in Section 4.2 can be obtained under the weaker condition $\|e\| = o_P(N^{2/3})$, and Lemma A.1 can also be applied with $u = e$ then. In that case, the assumptions in Lemma A.1 are not the same, but are similar to those imposed in Bai (2009b).

A.2 Expansion of Objective Function when $R = R^0$

Here we provide a heuristic derivation of the expansion of $\mathcal{L}_{NT}^0(\beta)$ in Theorem 4.2. We expand the profile objective function $\mathcal{L}_{NT}^0(\beta)$ simultaneously in β and in the spectral norm of e. Let the $K+1$ expansion parameters be defined by $\epsilon_0 = \|e\|/\sqrt{NT}$ and $\epsilon_k = \beta_k^0 - \beta_k$, $k = 1, \ldots, K$, and define the $N \times T$ matrix $X_0 = (\sqrt{NT}/\|e\|)e$. With these definitions we obtain

$$\frac{1}{\sqrt{NT}}\left(Y-\beta\cdot X\right) \;=\; \frac{1}{\sqrt{NT}}\left[\lambda^0 f^{0\prime}+(\beta^0-\beta)\cdot X+e\right] \;=\; \frac{\lambda^0 f^{0\prime}}{\sqrt{NT}}+\sum_{k=0}^{K}\epsilon_k\,\frac{X_k}{\sqrt{NT}}\,. \tag{A.1}$$

According to equation (3.3) the profile objective function $\mathcal{L}^0_{NT}(\beta)$ can be written as the sum over the $T-R^0$ smallest eigenvalues of the matrix in (A.1) multiplied by its transposed. We consider $\sum_{k=0}^K \epsilon_k\, X_k/\sqrt{NT}$ as a small perturbation of the unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$, and thus expand $\mathcal{L}^0_{NT}(\beta)$ in the perturbation parameters $\epsilon=(\epsilon_0,\ldots,\epsilon_K)$ around $\epsilon=0$, namely

$$\mathcal{L}^0_{NT}(\beta)=\frac{1}{NT}\sum_{g=0}^{\infty}\;\sum_{k_1,\ldots,k_g=0}^{K}\epsilon_{k_1}\,\epsilon_{k_2}\,\ldots\,\epsilon_{k_g}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_{k_1},\,X_{k_2},\ldots,X_{k_g}\right)\,, \tag{A.2}$$

where $L^{(g)}=L^{(g)}\left(\lambda^0,\,f^0,\,X_{k_1},\,X_{k_2},\ldots,X_{k_g}\right)$ are the expansion coefficients.

The unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$ has rank R^0, so that the $T-R^0$ smallest eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ are all zero, i.e. $\mathcal{L}^0_{NT}(\beta)=0$ for $\epsilon=0$ and thus $L^{(0)}\left(\lambda^0,\,f^0\right)=0$. Due to Assumption SF the R^0 non-zero eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ converge to positive constants as $N,T\to\infty$. This means that the "separating distance" of the $T-R^0$ zero-eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ converges to a positive constant, i.e. the next largest eigenvalue is well separated. This is exactly the technical condition under which the perturbation theory of linear operators guarantees that the above expansion of $\mathcal{L}^0_{NT}$ in ϵ exists and is convergent as long as the spectral norm of the perturbation $\sum_{k=0}^K \epsilon_k\, X_k/\sqrt{NT}$ is smaller than a particular convergence radius $r_0(\lambda^0,f^0)$, which is closely related to the separating distance of the zero-eigenvalues. For details on that see Kato (1980) and Section S.2 of the supplementary appendix, where we define $r_0(\lambda^0,f^0)$ and show that it converges to a positive constant as $N,T\to\infty$. Note that for the expansion (A.2) it is crucial that we have $R=R^0$, since the perturbation theory of linear operators describes the perturbation of the sum of *all* zero-eigenvalues of the unperturbed matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$. For $R>R^0$ the sum in $\mathcal{L}^R_{NT}(\beta)$ leaves out the $R-R^0$ largest of these perturbed zero-eigenvalues, which results in a much more complicated mathematical problem, since the structure and ranking among these perturbed zero-eigenvalues needs to be discussed.

The above expansion of $\mathcal{L}^0_{NT}(\beta)$ is applicable whenever the operator norm of the perturbation matrix $\sum_{k=0}^K \epsilon_k\, X_k/\sqrt{NT}$ is smaller than $r_0(\lambda^0,f^0)$. Since our assumptions guarantee that $\|X_k/\sqrt{NT}\|=\mathcal{O}_P(1)$, for $k=0,\ldots,K$, and $\epsilon_0=\mathcal{O}_P(\min(N,T)^{-1/2})=o_P(1)$, we have $\left\|\sum_{k=0}^K \epsilon_k\, X_k/\sqrt{NT}\right\|=\mathcal{O}_P(\|\beta-\beta^0\|)+o_P(1)$, i.e. the above expansion is always applicable asymptotically within a shrinking neighborhood of β^0 — which is sufficient since we already know that $\widehat{\beta}_R$ is consistent for $R\geq R^0$.

In addition, to guaranteeing converge of the series expansion, the perturbation theory of linear operators also provides explicit formulas for the expansion coefficients $L^{(g)}$, namely for $g=1,2,3$ we have $L^{(1)}\left(\lambda^0,f^0,X_k\right)=0$, $L^{(2)}\left(\lambda^0,f^0,X_{k_1},X_{k_2}\right)=\mathrm{Tr}(M_{\lambda^0}X_{k_1}M_{f^0}X'_{k_2})$, $L^{(3)}\left(\lambda^0,f^0,X_{k_1},X_{k_2},X_{k_3}\right)=-\frac{1}{3}[\mathrm{Tr}\left(M_{\lambda^0}X_{k_1}M_f X'_{k_2}\lambda^0(\lambda^{0\prime}\lambda^0)^{-1}(f^{0\prime}f^0)^{-1}f^{0\prime}X'_{k_3}\right)+\ldots]$, where the dots refer to 5 additional terms obtained from the first one by permutation of k_1, k_2 and k_3, so that the expression becomes totally symmetric in these indices. A general expression for the coefficients for all orders in g is given in Lemma S.1 in the appendix. One can show that for $g \geq 3$ the coefficients $L^{(g)}$ are bounded as follows

$$\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right)\right| \;\leq\; a_{NT}\; (b_{NT})^g\; \frac{\|X_{k_1}\|}{\sqrt{NT}}\, \frac{\|X_{k_2}\|}{\sqrt{NT}} \,\cdots\, \frac{\|X_{k_g}\|}{\sqrt{NT}}\,, \quad \text{(A.3)}$$

where a_{NT} and b_{NT} are functions of λ^0 and f^0 that converge to finite positive constants in probability. This bound on the coefficients $L^{(g)}$ allows us to derive a bound on the remainder term, when the profile objective expansion is truncated at a particular order. The expansion can be applied under more general asymptotics, but here we only consider the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, i.e. N and T grow at the same rate. Then, apart from the constant $\mathcal{L}^0_{NT}(\beta^0)$, the relevant coefficients of the expansion, which are not treated as part of the remainder term turn out to be $W_{k_1 k_2} = \frac{1}{NT} L^{(2)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}\right)$, $C^{(1)}_k = \frac{1}{\sqrt{NT}} L^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = \frac{1}{\sqrt{NT}} \operatorname{Tr}(M_{\lambda^0}\, X_k\, M_{f^0}\, e')$, and $C^{(2)}_k = \frac{3}{2\sqrt{NT}} L^{(3)}\left(\lambda^0,\, f^0,\, X_k,\, e,\, e\right)$, which corresponds exactly to the definitions in (4.1) above. From the expansion (A.2) and the bound (A.3) we obtain Theorem 4.2. For a more rigorous derivation we refer to Section S.2 in the supplementary appendix.