Econometric Theory, **33**, 2017, 158–195.
doi:10.1017/S0266466615000328

DYNAMIC LINEAR PANEL REGRESSION MODELS WITH INTERACTIVE FIXED EFFECTS

HYUNGSIK ROGER MOON
University of Southern California
Yonsei University

MARTIN WEIDNER
University College London

We analyze linear panel regression models with interactive fixed effects and predetermined regressors, for example lagged-dependent variables. The first-order asymptotic theory of the least squares (LS) estimator of the regression coefficients is worked out in the limit where both the cross-sectional dimension and the number of time periods become large. We find two sources of asymptotic bias of the LS estimator: bias due to correlation or heteroscedasticity of the idiosyncratic error term, and bias due to predetermined (as opposed to strictly exogenous) regressors. We provide a bias-corrected LS estimator. We also present bias-corrected versions of the three classical test statistics (Wald, LR, and LM test) and show their asymptotic distribution is a χ^2-distribution. Monte Carlo simulations show the bias correction of the LS estimator and of the test statistics also work well for finite sample sizes.

1. INTRODUCTION

In this paper, we study a linear panel regression model in which the individual fixed effects λ_i, called factor loadings, interact with common time-specific effects f_t, called factors. This interactive fixed effect specification contains the conventional individual specific effects and time-specific effects as special cases but is significantly more flexible because it allows the factors f_t to affect each individual with a different loading λ_i.

This paper is based on an unpublished manuscript of the authors that was circulated under the title "Likelihood Expansion for Panel Regression Models with Factors" but is now completely assimilated by the current paper and Moon & Weidner (2015). We greatly appreciate comments from the participants in the Far Eastern Meeting of the Econometric Society 2008, the SITE 2008 Conference, the All-UC-Econometrics Conference 2008, the July 2008 Conference in Honour of Peter Phillips in Singapore, the International Panel Data Conference 2009, the North American Summer Meeting of the Econometric Society 2009, and from seminar participants at Penn State, UCLA, and USC. We are also grateful for the comments and suggestions of Guido Kuersteiner, Peter Phillips, and anonymous referees. Moon is grateful for the financial support from the NSF via grant SES 0920903 and the faculty development award from USC. Weidner acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001. Address correspondence to Hyungsik Roger Moon, Department of Economics and USC Dornsife INET, University of Southern California, Los Angeles, CA 90089-0253. e-mail: moonr@usc.edu.

Factor models have been widely studied in various economics disciplines, for example, in asset pricing, forecasting, empirical macro, and empirical labor economics.[1] The panel literature often uses factor models to represent time-varying individual effects (or heterogenous time effects), so-called interactive fixed effects. For panels with a large cross-sectional dimension (N) but a short time dimension (T), Holtz-Eakin, Newey, & Rosen (1988) (hereafter HNR) study a linear panel regression model with interactive fixed effects and lagged dependent variables. To solve the incidental parameter problem caused by the λ_i's, they estimate a quasidifferenced version of the model using appropriate lagged variables as instruments, and treating f_t's as a fixed number of parameters to estimate. Ahn, Lee, & Schmidt (2001) also consider large N but short T panels. Instead of eliminating the individual effects λ_i by transforming the panel data, they impose various second-moment restrictions including the correlated random effects λ_i, and derive moment conditions to estimate the regression coefficients. The more recent literature considers panels with comparable size of N and T. The interactive fixed effect panel regression model of Pesaran (2006) allows heterogenous regression coefficients. Pesaran's estimator is the common correlated effect (CCE) estimator that uses the cross-sectional averages of the dependent variable and the independent variables as control functions for the interactive fixed effects.[2]

Among the interactive fixed effect panel literature, most closely related to our paper is Bai (2009). Bai assumes the regressors are *strictly* exogenous and the number of factors is known. The estimator he investigates is the least squares (LS) estimator, which minimizes the sum of squared residuals of the model jointly over the regression coefficients and the fixed effect parameters λ_i and f_t.[3] Using alternative asymptotics where $N, T \to \infty$ at the same rate,[4] Bai shows the LS estimator is $\sqrt{NT}$-consistent and asymptotically normal, but may have an asymptotic bias. The bias in the normal limiting distribution occurs when the regression errors are correlated or heteroscedastic. Bai also shows how to estimate the bias, and proposes a bias-corrected estimator.

Following the methodology in Bai (2009), we investigate the LS estimator for a linear panel regression with a known number of interactive fixed effects. The main difference from Bai is that we consider *predetermined* regressors, thus allowing feedback of past outcomes to future regressors. One of the main findings of the present paper is that the limit distribution of the LS estimator has two types of biases: one type of bias due to correlated or heteroscedastic errors (the same bias as in Bai) and the other type of bias due to the predetermined regressors. This additional bias term is analogous to the incidental parameter bias of Nickell (1981) in finite T and the bias in Hahn & Kuersteiner (2002) in large T.

In addition to allowing for predetermined regressors, we also extend Bai's results to models in which both "low-rank regressors" (e.g., time-invariant and common regressors, or interactions of those two) and "high-rank-regressors" (almost all other regressors that vary across individuals and over time) are present simultaneously, whereas Bai (2009) only considers the low-rank regressors separately and in a restrictive setting (in particular, not allowing for regressors

that are obtained by interacting time-invariant and common variables). A general treatment of low-rank regressors is desirable because they often occur in applied work, for example, Gobillon & Magnac (2013). The analysis of those regressors is challenging, however, because the unobserved interactive fixed effects also represent a low-rank $N \times T$ matrix, thus posing a nontrivial identification problem for low-rank regressors, which needs to be addressed. We provide conditions under which the different types of regressors are identified jointly, and under which they can be estimated consistently as N and T grow large.

Another contribution of this paper is to establish the asymptotic theory of the three classical test statistics (Wald test, LR test, and LM (or score) test) for testing restrictions on the regression coefficients in a large N, T panel framework.[5] Regarding testing for coefficient restrictions, Bai (2009) investigates the Wald test based on the bias-corrected LS estimator, and HNR consider the LR test in their 2SLS estimation framework with fixed T.[6] What we show is that the conventional LR and LM test statistics based on the LS profile objective function have noncentral chi-square limits due to incidental parameters in the interactive fixed effects. We therefore propose modified LR and LM tests whose asymptotic distributions are conventional chi-square distributions.

To establish the asymptotic theories of the LS estimator and the three classical tests, we use the quadratic approximation of the profile LS objective function derived in Moon & Weidner (2015). This method is different from Bai (2009), who uses the first-order condition of the LS optimization problem as the starting point of his analysis. One advantage of our methodology is that it can also directly be applied to derive the asymptotic properties of the LR and LM test statistics.

In this paper, we assume the regressors are not endogenous and the number of factors is known, which might be restrictive in some applications. In other papers, we study how to relax these restrictions. Moon & Weidner (2015) investigates the asymptotic properties of the LS estimator of the linear panel regression model with factors when the number of factors is unknown and extra factors are included unnecessarily in the estimation. We find that under suitable conditions,[7] the limit distribution of the LS estimator is unchanged when the number of factors is overestimated. The extension to allow for endogenous regressors is very briefly discussed in Section 6 of the current paper, and is closely related to the results in Moon, Shum, & Weidner (2012) (hereafter MSW). MSW's main purpose is to extend the random coefficient multinomial logit demand model (known as the BLP demand model from Berry, Levinsohn, & Pakes (1995)) by allowing for interactive product and market specific fixed effects. Although the main model of interest is quite different from the linear panel regression model of the current paper, MSW's econometrics framework is directly applicable to the model of the current paper with endogenous regressors.[8]

Comparing the different estimation approaches for interactive fixed effect panel regressions proposed in the literature, it seems fair to say that the LS estimator in Bai (2009) and our paper, the CCE estimator of Pesaran (2006), and the IV estimator based on quasidifferencing in HNR all have their own relative

advantages and disadvantages. These three estimation methods handle the interactive fixed effects quite differently. The LS method concentrates out the interactive fixed effects by taking out the principal components. The CCE method controls the factor (or time effects) using the cross-sectional averages of the dependent and independent variables. The HNR's approach quasidifferences out the individual effects, treating the remaining time effects as parameters to estimate. The IV estimator of HNR should work well when T is short, but is expected to also suffer from an incidental parameter problem when T becomes large, because then many factors need to be estimated as parameters that enter the model nonlinearly. Pesaran's CCE estimation method does not require the number of factors to be known and does not require the strong factor assumption that we will impose below, but for the CCE estimator to work, not only the DGPs of the dependent variable (e.g., the regression model) but also the DGPs of the explanatory variables need to be restricted such that their cross-sectional average can control for unobserved factors. The LS estimator and its bias-corrected version perform well under relatively weak restrictions on the regressors, but requires that T should not be too small and that the factors should be sufficiently strong to be correctly picked up as the leading principal components.

The paper is organized as follows. In Section 2, we introduce the interactive fixed effect model and provide conditions for identifying the regression coefficients in the presence of the interactive fixed effects. In Section 3, we define the LS estimator of the regression parameters and provide a set of assumptions that are sufficient to show consistency of the LS estimator. In Section 4, we work out the asymptotic distribution of the LS estimator under alternative asymptotics. We also provide a consistent estimator for the asymptotic bias and a bias-corrected LS estimator. In Section 5, we consider the Wald, LR, and LM tests for testing restrictions on the regression coefficients of the model. We present bias-corrected versions of these tests and show that they have chi-square limiting distributions. In Section 6, we briefly discuss how to estimate the interactive fixed effect linear panel regression when the regressors are endogenous. In Section 7, we present Monte Carlo simulation results for an AR(1) model with interactive fixed effects. The simulations show the LS estimator for the AR(1) coefficient is biased, and the tests based on it can have severe size distortions and power asymmetries, whereas the bias-corrected LS estimator and test statistics have better properties. We conclude in Section 8. We present all proofs of theorems and some technical details in the appendix or supplementary material.

A few words on notation are due. For a column vector v, the Euclidean norm is defined by $\|v\| = \sqrt{v'v}$. For the n-th largest eigenvalues (counting multiple eigenvalues multiple times) of a symmetric matrix B, we write $\mu_n(B)$. For an $m \times n$ matrix A, the Frobenius norm is $\|A\|_F = \sqrt{\mathrm{Tr}(AA')}$, and the spectral norm is $\|A\| = \max_{0 \neq v \in \mathbb{R}^n} \frac{\|Av\|}{\|v\|}$, or equivalently $\|A\| = \sqrt{\mu_1(A'A)}$. Furthermore, we define $P_A = A(A'A)^{\dagger}A'$ and $M_A = \mathbb{I} - A(A'A)^{\dagger}A'$, where $\mathbb{I}$ is the $m \times m$ identity matrix, and $(A'A)^{\dagger}$ is the Moore–Penrose pseudoinverse, to allow for the case that A is not of full column rank. For square matrices B, C, we write $B > C$ (or $B \geq C$)

to indicate $B - C$ is positive (semi) definite. For a positive definite symmetric matrix A, we write $A^{1/2}$ and $A^{-1/2}$ for the unique symmetric matrices that satisfy $A^{1/2}A^{1/2} = A$ and $A^{-1/2}A^{-1/2} = A^{-1}$. We use ∇ for the gradient of a function; that is, $\nabla f(x)$ is the column vector of partial derivatives of f with respect to each component of x. We use "wpa1" for "with probability approaching one".

2. MODEL AND IDENTIFICATION

We study the following panel regression model with cross-sectional size N, and T time periods:

$$Y_{it} = \beta^{0\prime} X_{it} + \lambda_i^{0\prime} f_t^0 + e_{it}, \qquad i = 1 \dots N, \quad t = 1 \dots T, \tag{1}$$

where X_{it} is a $K \times 1$ vector of observable regressors, β^0 is a $K \times 1$ vector of regression coefficients, λ_i^0 is an $R \times 1$ vector of unobserved factor loadings, f_t^0 is an $R \times 1$ vector of unobserved common factors, and e_{it} are unobserved errors. The superscript zero indicates the true parameters. We write f_{tr}^0 and λ_{ir}^0, where $r = 1, \dots, R$, for the components of λ_i^0 and f_t^0, respectively. R is the number of factors. Note that we can have $f_{tr}^0 = 1$ for all t and a particular r, in which case the corresponding λ_{ir}^0 become standard individual-specific effects. Analogously, we can have $\lambda_{ir}^0 = 1$ for all i and a particular r, so that the corresponding f_{tr}^0 become standard time-specific effects.

Throughout this paper, we assume the true number of factors R is known.[9] We introduce the notation $\beta^0 \cdot X = \sum_{k=1}^K \beta_k^0 X_k$. In matrix notation, the model can then be written as

$$Y = \beta^0 \cdot X + \lambda^0 f^{0\prime} + e\,,$$

where Y, X_k, and e are $N \times T$ matrices, λ^0 is an $N \times R$ matrix, and f^0 is a $T \times R$ matrix. The elements of X_k are denoted by $X_{k,it}$.

We separate the K regressors into K_1 "low-rank regressors" X_l, $l = 1, \dots, K_1$, and $K_2 = K - K_1$ "high-rank regressors" X_m, $m = K_1 + 1, \dots, K$. Each low-rank regressor $l = 1, \dots, L$ is assumed to satisfy $\text{rank}(X_l) = 1$. Therefore, we can write $X_l = w_l v_l'$, where w_l is an N-vector and v_l is a T-vector, and we also define the $N \times K_1$ matrix $w = (w_1, \dots, w_{K_1})$ and the $T \times K_1$ matrix $v = (v_1, \dots, v_{K_1})$.

Let $l = 1, \dots, K_1$. The two most prominent types of low-rank regressors are time-invariant regressors, which satisfy $X_{l,it} = Z_i$ for all i, t, and common (or cross-sectionally invariant) regressors, in which case $X_{l,it} = W_t$ for all i, t. Here, Z_i and W_t are some observed variables, which only vary over i or t, respectively. A more general low-rank regressor can be obtained by interacting Z_i and W_t multiplicatively, namely, $X_{l,it} = Z_i W_t$, an empirical example of which is given in Gobillon & Magnac (2013). In these examples, and probably for the vast majority of applications, the low-rank regressors all satisfy $\text{rank}(X_l) = 1$, but our results can easily be extended to more general low-rank regressors.[10]

High-rank regressors are those whose distribution guarantees they have high rank (usually full rank) when considered as an $N \times T$ matrix. For example,

a regressor whose entries satisfy $X_{m,it} \sim iid\mathcal{N}(\mu,\sigma)$, with $\mu \in \mathbb{R}$ and $\sigma > 0$, satisfies $\text{rank}(X_m) = \min(N,T)$ with probability one.

This separation of the regressors into low- and high-rank regressors is important to formulate our assumptions for identification and consistency, but actually plays no role in the estimation and inference procedures for $\widehat{\beta}$ discussed below.

Assumption ID (Assumptions for Identification).

(i) **Existence of Second Moments:**
The second moments of $X_{k,it}$ and e_{it} conditional on λ^0, f^0, w exist for all i, t, k.

(ii) **Mean Zero Errors and Exogeneity:**
$\mathbb{E}\big(e_{it}|\lambda^0, f^0, w\big) = 0$, *and* $\mathbb{E}\big(X_{k,it}e_{it}|\lambda^0, f^0, w\big) = 0$, *a.s., for all i, t, k.*

The following two assumptions only need to be imposed if $K_1 > 0$, that is, if low-rank regressors are present:

(iii) **Noncollinearity of Low-Rank Regressors:**
Consider linear combinations $\alpha \cdot X_{\text{low}} = \sum_{l=1}^{K_1} \alpha_l X_l$ of the low-rank regressors X_l with $\alpha \in \mathbb{R}^{K_1}$. For all $\alpha \neq 0$, we assume

$$\mathbb{E}\Big[(\alpha \cdot X_{\text{low}})M_{f^0}(\alpha \cdot X_{\text{low}})'\big|\lambda^0, f^0, w\Big] \neq 0, \qquad a.s.$$

(iv) **No Collinearity between Factor Loadings and Low-Rank Regressors:**
$\text{rank}\big(M_w\lambda^0\big) = \text{rank}\big(\lambda^0\big)$.[11]

The following assumption only needs to be imposed if $K_2 > 0$, that is, if high-rank regressors are present:

(v) **Noncollinearity of High-Rank Regressors:**
Consider linear combinations $\alpha \cdot X_{\text{high}} = \sum_{m=K_1+1}^{K} \alpha_m X_m$ of the high-rank regressors X_m for $\alpha \in \mathbb{R}^{K_2}$, where the components of the K_2-vector α are denoted by α_{K_1+1} to α_K. For all $\alpha \neq 0$, we assume

$$\text{rank}\left\{\mathbb{E}\Big[(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})'\big|\lambda^0, f^0, w\Big]\right\} > 2R + K_1, \qquad a.s.$$

All expectations in the assumptions are conditional on λ^0, f^0, and w; in particular, e_{it} is not allowed to be correlated with λ^0, f^0, and w. However, e_{it} is allowed to be correlated with v (i.e., predetermined low-rank regressors are allowed). If desired, one can interchange the role of N and T in the assumptions, by using the formal symmetry of the model under exchange of the panel dimensions ($N \leftrightarrow T$, $\lambda^0 \leftrightarrow f^0$, $Y \leftrightarrow Y'$, $X_k \leftrightarrow X_k'$, $w \leftrightarrow v$).

Assumptions ID(i) and (ii) have standard interpretations, but the other assumptions require some further discussion.

Assumption ID(iii) states the low-rank regressors are noncollinear even after projecting out all variation that is explained by the true factors f^0. This assumption would, for example, be violated if $v_l = f_r^0$ for some $l = 1,\ldots,K_1$ and $r = 1,\ldots,R$, because then $X_l M_{f^0} = 0$ and we can choose α such that

$X_{\text{low}} = X_l$. Similarly, Assumption ID(iv) rules out, for example, that $w_l = \lambda_r^0$ for some $l = 1, \ldots, K_1$ and $r = 1, \ldots, R$, because then $\text{rank}(M_w \lambda^0) < \text{rank}(\lambda^0)$, in general. It ought to be expected that λ^0 and f^0 have to feature in the identification conditions for the low-rank regressors, because the interactive fixed effects structure and the low-rank regressors represent similar types of low-rank $N \times T$ structures.

Assumption ID(v) is a generalized noncollinearity assumption for the high-rank regressors, which guarantees any linear combination $\alpha \cdot X_{\text{high}}$ of the high-rank regressors is sufficiently different from the low-rank regressors and from the interactive fixed effects. A standard noncollinearity assumption can be formulated by demanding the $N \times N$ matrix $\mathbb{E}\big[(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})' \big| \lambda^0, f^0, w\big]$ is nonzero for all nonzero $\alpha \in \mathbb{R}^{K_2}$, which can be equivalently expressed as $\text{rank}\big\{\mathbb{E}\big[(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})' \big| \lambda^0, f^0, w\big]\big\} > 0$ for all nonzero $\alpha \in \mathbb{R}^{K_2}$. Assumption ID($v$) strengthens this standard noncollinearity assumption by demanding the rank not only to be positive, but larger than $2R + K_1$. This also explains the name "high-rank regressors," because their rank has to be sufficiently large to satisfy this assumption. Note also that only the number of factors R, but not λ^0 and f^0, features in Assumption ID(v). The sample version of this assumption is given by Assumption 4(ii)(a) below, which is also very closely related to Assumption A in Bai (2009).

THEOREM 2.1 (Identification). *Suppose the Assumptions ID are satisfied. Then, the minima of the expected objective function* $\mathbb{E}\left(\left\| Y - \beta \cdot X - \lambda f' \right\|_F^2 \middle| \lambda^0, f^0, w \right)$ *over* $(\beta, \lambda, f) \in \mathbb{R}^{K+N\times R+T\times R}$ *satisfy* $\beta = \beta^0$ *and* $\lambda f' = \lambda^0 f^{0\prime}$. *This shows that* β^0 *and* $\lambda^0 f^{0\prime}$ *are identified.*

The theorem shows the true parameters are identified as minima of the expected value of $\left\| Y - \beta \cdot X - \lambda f' \right\|_F^2 = \sum_{i,t} (Y_{it}\beta' - X_{it} - \lambda_i' f_t)^2$, which is the sum of squared residuals. We use the same objective function, to define the estimators $\widehat{\beta}$, $\widehat{\lambda}$ and $\widehat{f}$ below. Without further normalization conditions, the parameters λ^0 and f^0 are not separately identified, because the outcome variable Y is invariant under transformations $\lambda^0 \to \lambda^0 A'$ and $f^0 \to f^0 A^{-1}$, where A is a nonsingular $R \times R$ matrix. However, the product $\lambda^0 f^{0\prime}$ is uniquely identified according to the theorem. Because our focus is on identification and estimation of β^0, we do not need to discuss those additional normalization conditions for λ^0 and f^0 in this paper.

3. ESTIMATOR AND CONSISTENCY

The objective function of the model is simply the sum of squared residuals, which in matrix notation can be expressed as

$$\mathcal{L}_{NT}(\beta, \lambda, f) = \frac{1}{NT} \left\| Y - \beta \cdot X - \lambda f' \right\|_F^2 = \frac{1}{NT} \text{Tr}\Big[\big(Y - \beta \cdot X - \lambda f'\big)' \big(Y - \beta \cdot X - \lambda f'\big) \Big]. \tag{2}$$

The estimator we consider is the LS estimator that jointly minimizes $\mathcal{L}_{NT}(\beta, \lambda, f)$ over β, λ and f. Our main objects of interest are the regression parameters $\beta = (\beta_1, ..., \beta_K)'$, whose estimator is given by

$$\widehat{\beta} = \underset{\beta \in \mathbb{B}}{\operatorname{argmin}} \, L_{NT}(\beta) \,, \tag{3}$$

where $\mathbb{B} \subset \mathbb{R}^K$ is a compact parameter set that contains the true parameter, namely, $\beta^0 \in \mathbb{B}$, and the objective function is the profile objective function

$$\begin{aligned} L_{NT}(\beta) &= \min_{\lambda, f} \mathcal{L}_{NT}(\beta, \lambda, f) \\ &= \min_{f} \frac{1}{NT} \operatorname{Tr}\left[(Y - \beta \cdot X) M_f (Y - \beta \cdot X)'\right] \\ &= \frac{1}{NT} \sum_{r=R+1}^{T} \mu_r \left[(Y - \beta \cdot X)'(Y - \beta \cdot X)\right]. \end{aligned} \tag{4}$$

Here, the first expression for $L_{NT}(\beta)$ is its definition as the minimum value of $\mathcal{L}_{NT}(\beta, \lambda, f)$ over λ and f. We denote the minimizing incidental parameters by $\widehat{\lambda}(\beta)$ and $\widehat{f}(\beta)$, and we define the estimators $\widehat{\lambda} = \widehat{\lambda}(\widehat{\beta})$ and $\widehat{f} = \widehat{f}(\widehat{\beta})$. Those minimizing incidental parameters are not uniquely determined – for the same reason that λ^0 and f^0 are nonuniquely identified – but the product $\widehat{\lambda}(\beta)\widehat{f}'(\beta)$ is unique.

The second expression for $L_{NT}(\beta)$ in Equation (4) is obtained by concentrating out λ (analogously, one can concentrate out f to obtain a formulation whereby only the parameter λ remains). The optimal f in the second expression is given by the R eigenvectors that correspond to the R largest eigenvalues of the $T \times T$ matrix $(Y - \beta \cdot X)'(Y - \beta \cdot X)$. This insight leads to the third line that presents the profile objective function as the sum over the $T - R$ smallest eigenvalues of this $T \times T$ matrix. Lemma A.1 in the appendix shows equivalence of the three expressions for $L_{NT}(\beta)$ given above.

Multiple local minima of $L_{NT}(\beta)$ may exist, and one should use multiple starting values for the numerical optimization of β to guarantee the true global minimum $\widehat{\beta}$ is found.

To show consistency of the LS estimator $\widehat{\beta}$ of the interactive fixed effect model, and also later for our first-order asymptotic theory, we consider the limit $N, T \to \infty$. In the following we present assumptions on X_k, e, λ, and f that guarantee consistency.[12]

Assumption 1. (i) $\operatorname{plim}_{N,T\to\infty}\left(\lambda^{0\prime}\lambda^0/N\right) > 0$, (ii) $\operatorname{plim}_{N,T\to\infty}\left(f^{0\prime}f^0/T\right) > 0$.

Assumption 2. $\operatorname{plim}_{N,T\to\infty}\left[(NT)^{-1}\operatorname{Tr}\left(X_k e'\right)\right] = 0$, for all $k = 1, \ldots, K$.

Assumption 3. $\operatorname{plim}_{N,T\to\infty}\left(\|e\|/\sqrt{NT}\right) = 0$.

Assumption 1 guarantees the matrices f^0 and λ^0 have full rank, that is, that R distinct factors and factor loadings exist asymptotically, and that the norm of

each factor and factor loading grows at a rate of $\sqrt{T}$ and $\sqrt{N}$, respectively. Assumption 2 demands the regressors are weakly exogenous. Assumption 3 restricts the spectral norm of the $N \times T$ error matrix e. We discuss this assumption in more detail in Section 4, and we give examples of error distributions that satisfy this condition in Section S.2 of the supplementary material. The final assumption needed for consistency is an assumption on the regressors X_k. We already introduced the distinction between the K_1 "low-rank regressors" $X_l, l = 1, \ldots, K_1$, and the $K_2 = K - K_1$ "high-rank regressors" $X_m, m = K_1 + 1, \ldots, K$ above.

Assumption 4.

(i) $\operatorname{plim}_{N,T\to\infty} \left[(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} X_{it}' \right] > 0.$

(ii) The two types of regressors satisfy:

(a) Consider linear combinations $\alpha \cdot X_{\text{high}} = \sum_{m=K_1+1}^{K} \alpha_m X_m$ of the high-rank regressors X_m for K_2-vectors α with $\|\alpha\| = 1$, where the components of the K_2-vector α are denoted by α_{K_1+1} to α_K. We assume a constant $b > 0$ exists such that

$$\min_{\{\alpha \in \mathbb{R}^{K_2}, \|\alpha\|=1\}} \sum_{r=2R+K_1+1}^{N} \mu_r \left[\frac{(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})'}{NT} \right] \geq b \qquad \text{wpa1.}$$

(b) For the low-rank regressors, we assume $\operatorname{rank}(X_l) = 1, l = 1, \ldots, K_1$; that is, they can be written as $X_l = w_l v_l'$ for N-vectors w_l and T-vectors v_l, and we define the $N \times K_1$ matrix $w = (w_1, \ldots, w_{K_1})$ and the $T \times K_1$ matrix $v = (v_1, \ldots, v_{K_1})$. We assume a constant $B > 0$ exists such that $N^{-1} \lambda^{0\prime} M_w \lambda^0 > B \mathbb{I}_R$ and $T^{-1} f^{0\prime} M_v f^0 > B \mathbb{I}_R$, wpa1.

Assumption $4(i)$ is a standard noncollinearity condition for all the regressors. Assumption $4(ii)(a)$ is an appropriate sample analog of the identification Assumption ID(v). If the sum in Assumption $4(ii)(a)$ were to start from $r = 1$, we would have $\sum_{r=1}^{N} \mu_r \left[\frac{(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})'}{NT} \right] = \frac{1}{NT} \operatorname{Tr}[(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})']$, so that the assumption would become a standard noncollinearity condition. Not including the first $2R + K_1$ eigenvalues in the sum implies the $N \times N$ matrix $(\alpha \cdot X_{\text{high}})(\alpha \cdot X_{\text{high}})'$ needs to have rank larger than $2R + K_1$.

Assumption $4(ii)(b)$ is closely related to the identification Assumptions ID(iii) and (iv). The appearance of the factors and factor loadings in this assumption on the low-rank regressors is inevitable to guarantee consistency. For example, consider a low-rank regressor that is cross-sectionally independent and proportional to the r'th unobserved factor, for example, $X_{l,it} = f_{tr}$. The corresponding regression coefficient β_l is then not identified, because the model is invariant under a shift $\beta_l \mapsto \beta_l + a$, $\lambda_{ir} \mapsto \lambda_{ir} - a$, for an arbitrary $a \in \mathbb{R}$. This phenomenon is well known from ordinary fixed effect models, where the coefficients of time-invariant regressors are not identified. Assumption 4(ii)(b) therefore

guarantees for $X_l = w_l v_l'$ that w_l is sufficiently different from λ^0, and v_l is sufficiently different from f^0.

THEOREM 3.1 (Consistency). *Let Assumptions 1, 2, 3, and 4 be satisfied; let the parameter set $\mathbb{B}$ be compact; and let $\beta^0 \in \mathbb{B}$. In the limit $N, T \to \infty$, we then have*

$$\widehat{\beta} \underset{p}{\longrightarrow} \beta^0.$$

We assume compactness of $\mathbb{B}$ to guarantee existence of the minimizing $\widehat{\beta}$. We also use boundedness of $\mathbb{B}$ in the consistency proof, but only for those parameters $\beta_l, l = 1 \dots K_1$, that correspond to low-rank regressors, that is, if only high-rank regressors ($K_1 = 0$) are present, the compactness assumption can be omitted, as long as existence of $\widehat{\beta}$ is guaranteed (e.g., for $\mathbb{B} = \mathbb{R}^K$).

Bai (2009) also proves consistency of the LS estimator of the interactive fixed effect model, but under somewhat different assumptions. He also employs what we call Assumptions 1 and 2, and he uses a low-level version of Assumption 3. He demands the regressors to be strictly exogenous. Regarding consistency, the main difference between our assumptions and his is the treatment of high- and low-rank regressors. He first gives a condition on the regressors (his Assumption A) that rules out low-rank regressors, and later discusses the case in which all regressors are either time-invariant or common regressors (i.e., are all low rank). By contrast, our Assumption 4 allows for a combination of high- and low-rank regressors, and for low-rank regressors that are more general than time-invariant and common regressors.

4. ASYMPTOTIC DISTRIBUTION AND BIAS CORRECTION

Because we have already shown consistency of the LS estimator $\widehat{\beta}$, it is sufficient to study the local properties of the objective function $L_{NT}(\beta)$ around β^0 to derive the first-order asymptotic theory of $\widehat{\beta}$. Moon and Weidner (2015) derived a useful approximation of $L_{NT}(\beta)$ around β^0, and we briefly summarize the ideas and results of this approximation in the following subsection. We then apply those results to derive the asymptotic distribution of the LS estimator, including working out the asymptotic bias, which was not done previously. Afterward, we discuss bias correction and inference.

4.1. Expansion of the Profile Objective Function

The last expression in Equation (4) for the profile objective function is convenient because it does not involve any minimization over the parameters λ or f. On the other hand, this expression cannot be easily discussed by analytic means, because in general, no explicit formula exists for the eigenvalues of a matrix. The conventional method that involves a Taylor series expansion in the regression parameters β *alone* seems infeasible here. In Moon and Weidner (2015), we showed how to

overcome this problem by expanding the profile objective function *jointly* in β and $\|e\|$. The key idea is the following decomposition:

$$Y - \beta \cdot X = \underbrace{\lambda^0 f^{0\prime}}_{\substack{\text{leading} \\ \text{term}}} \underbrace{- \left(\beta - \beta^0\right) \cdot X + e}_{\text{perturbation term}}.$$

If the perturbation term is zero, the profile objective $L_{NT}(\beta)$ is also zero, because the leading term $\lambda^0 f^{0\prime}$ has rank R, so that the $T - R$ smallest eigenvalues of $f^0 \lambda^{0\prime} \lambda^0 f^{0\prime}$ all vanish. One may thus expect that small values of the perturbation term should correspond to small values of $L_{NT}(\beta)$. This idea can indeed be made mathematically precise. By using the perturbation theory of linear operators (see, e.g., Kato), one can work out an expansion of $L_{NT}(\beta)$ in the perturbation term, and one can show this expansion is convergent as long as the spectral norm of the perturbation term is sufficiently small.

The assumptions on the model made so far are in principle already sufficient to apply this expansion of the profile objective function, but to truncate the expansion at an appropriate order and to provide a bound on the remainder term that is sufficient to derive the first-order asymptotic theory of the LS estimator, we need to strengthen Assumption 3 as follows.

Assumption 3*. $\|e\| = o_p\left(N^{2/3}\right)$.

In the rest of the paper, we only consider asymptotics in which N and T grow at the same rate; that is, we could equivalently write $o_p\left(T^{2/3}\right)$ instead of $o_p\left(N^{2/3}\right)$ in Assumption 3*. In Section S.2 of the supplementary material, we provide examples of error distributions that satisfy Assumption 3*. In fact, for these examples, we have $\|e\| = \mathcal{O}_p\left(\sqrt{\max(N,T)}\right)$. A large literature studies the asymptotic behavior of the spectral norm of random matrices; see, for example, Geman (1980), Silverstein (1989), Bai, Silverstein, & Yin (1988), Yin, Bai, and Krishnaiah (1988), and Latala (2005). Loosely speaking, we expect the result $\|e\| = \mathcal{O}_p\left(\sqrt{\max(N,T)}\right)$ to hold as long as the errors e_{it} have mean zero, uniformly bounded fourth moment, and weak time-serial and cross-sectional correlation (in some well-defined sense, see the examples).

We can now present the quadratic approximation of the profile objective function $L_{NT}(\beta)$ that we derived in Moon and Weidner (2015).

THEOREM 4.1 (Expansion of Profile Objective Function). *Let Assumption 1, 3*, and 4(i) be satisfied, and consider the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Then, the profile objective function satisfies $L_{NT}(\beta) = L_{q,NT}(\beta) + (NT)^{-1} R_{NT}(\beta)$, where the remainder $R_{NT}(\beta)$ is such that for any sequence $\eta_{NT} \to 0$, we have*

$$\sup_{\{\beta : \|\beta - \beta^0\| \leq \eta_{NT}\}} \frac{|R_{NT}(\beta)|}{\left(1 + \sqrt{NT}\, \|\beta - \beta^0\|\right)^2} = o_p(1),$$

and $L_{q,NT}(\beta)$ is a second-order polynomial in β; namely,

$$L_{q,NT}(\beta) = L_{NT}(\beta^0) - \frac{2}{\sqrt{NT}}\left(\beta-\beta^0\right)' C_{NT} + \left(\beta-\beta^0\right)' W_{NT}\left(\beta-\beta^0\right),$$

with $K \times K$ matrix W_{NT} defined by $W_{NT,k_1k_2} = (NT)^{-1}\operatorname{Tr}\left(M_{f^0} X'_{k_1} M_{\lambda^0} X_{k_2}\right)$, and K-vector C_{NT} with entries $C_{NT,k} = C^{(1)}\left(\lambda^0, f^0, X_k e\right) + C^{(2)}\left(\lambda^0, f^0, X_k e\right)$, where

$$C^{(1)}\left(\lambda^0, f^0, X_k, e\right) = \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(M_{f^0} e' M_{\lambda^0} X_k\right),$$

$$\begin{aligned} C^{(2)}\left(\lambda^0, f^0, X_k, e\right) = -\frac{1}{\sqrt{NT}}\Bigg[&\operatorname{Tr}\left(e M_{f^0} e' M_{\lambda^0} X_k f^0 \left(f^{0\prime} f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right) \\ &+\operatorname{Tr}\left(e' M_{\lambda^0} e M_{f^0} X'_k \lambda^0 \left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime} f^0\right)^{-1} f^{0\prime}\right) \\ &+\operatorname{Tr}\left(e' M_{\lambda^0} X_k M_{f^0} e' \lambda^0 \left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime} f^0\right)^{-1} f^{0\prime}\right)\Bigg]. \end{aligned}$$

We refer to W_{NT} and C_{NT} as the approximated Hessian and the approximated score (at the true parameter β^0). The exact Hessian and the exact score (at the true parameter β^0) contain higher-order expansion terms in e, but the expansion up to the particular order above is sufficient to work out the first-order asymptotic theory of the LS estimator, as the following corollary shows.

COROLLARY 4.2. *Let the assumptions of Theorem 3.1 and 4.1 hold; let β^0 be an interior point of the parameter set $\mathbb{B}$; and assume $C_{NT} = \mathcal{O}_p(1)$. We then have $\sqrt{NT}\left(\widehat{\beta}-\beta^0\right) = W_{NT}^{-1} C_{NT} + o_p(1) = \mathcal{O}_p(1)$.*

Combining consistency of the LS estimator and the expansion of the profile objective function in Theorem 4.1, one obtains $\sqrt{NT}\, W_{NT}\left(\widehat{\beta}-\beta^0\right) = C_{NT} + o_p(1)$; see, for example, Andrews (1999). To obtain the corollary, one needs in addition that W_{NT} does not become degenerate as $N, T \to \infty$; that is, the smallest eigenvalue of W_{NT} should be bounded from below by a positive constant. Our assumptions already guarantee existence of such a lower bound, as is shown in the supplementary material.

Analogous to the expansions of the profile objective function $L_{NT}(\beta)$, one can also derive expansions of the projectors $M_{\widehat{\lambda}}$ and $M_{\widehat{f}}$, and those can be used to show consistency of $\widehat{\lambda}$ and $\widehat{f}$, up to normalization; see Lemma S.10.4 in the supplementary material.

4.2. Asymptotic Distribution

We now apply Corollary 4.2 to work out the asymptotic distribution of the LS estimator $\widehat{\beta}$. For this purpose, we need more specific assumptions on λ^0, f^0, X_k, and e.

Assumption 5. A sigma algebra $\mathcal{C} = \mathcal{C}_{NT}$ (which in the following we will refer to as the conditioning set) exists that contains the sigma algebra generated by λ^0 and f^0, such that

(i) $\mathbb{E}\left[e_{it} \,\middle|\, \mathcal{C} \vee \sigma\left(\{(X_{is}, e_{i,s-1}), s \le t\}\right)\right] = 0$, for all i, t.[13]

(ii) e_{it} is independent over t, conditional on $\mathcal{C}$, for all i.

(iii) $\{(X_{it}, e_{it}), t = 1, \ldots, T\}$ is independent across i, conditional on $\mathcal{C}$.

(iv) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t,s=1}^{T}\left|\mathrm{Cov}\left(X_{k,it}, X_{\ell,is} \,\middle|\, \mathcal{C}\right)\right| = \mathcal{O}_p(1)$, for all $k, \ell = 1, \ldots, K$.

(v) $\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t,s,u,v=1}^{T}\left|\mathrm{Cov}\left(e_{it}\widetilde{X}_{k,is}, e_{iu}\widetilde{X}_{\ell,iv} \,\middle|\, \mathcal{C}\right)\right| = \mathcal{O}_p(1)$, where $\widetilde{X}_{k,it} = X_{k,it} - \mathbb{E}\left[X_{k,it} \,\middle|\, \mathcal{C}\right]$, for all $k, \ell = 1, \ldots, K$.

(vi) An $\epsilon > 0$ exists such that $\mathbb{E}\left(e_{it}^8 \,\middle|\, \mathcal{C}\right)$ and $\mathbb{E}\left(\|X_{it}\|^{8+\epsilon} \,\middle|\, \mathcal{C}\right)$ and $\mathbb{E}\|\lambda_i^0\|^4$ and $\mathbb{E}\|f_t^0\|^{4+\epsilon}$ are bounded by a nonrandom constant, uniformly over i, t and N, T.

(vii) β^0 is an interior point of the compact parameter set $\mathbb{B}$.

Remarks on Assumption 5.

(1) Part (i) of Assumption 5 imposes that e_{it} is a martingale difference sequence over time for a particular filtration. Conditioning on $\mathcal{C}$, the time series of e_{it} is independent over time (part (ii) of the assumption) and the error term e_{it} and regressors X_{it} are cross-sectionally independent (part (iii) of the assumption), but unconditional correlation is allowed. Part (iv) imposes weak time-serial correlation of X_{it}. Part (v) demands weak time-serial correlation of $\widetilde{X}_{k,it} = X_{k,it} - \mathbb{E}\left[X_{k,it} \,\middle|\, \mathcal{C}\right]$ and e_{it}. Finally, parts (vi) and (vii) require bounded higher moments of the error term, regressors, factors and factor loadings, and a compact parameter set with an interior true parameter.

(2) Assumption $5(i)$ implies $\mathbb{E}\left(X_{k,it}e_{it} \,\middle|\, \mathcal{C}\right) = 0$ and $\mathbb{E}\left(X_{k,it}e_{it}X_{\ell,is}e_{is} \,\middle|\, \mathcal{C}\right) = 0$ for $t \neq s$. Thus, the assumption guarantees $X_{it}e_{it}$ is mean zero and uncorrelated over t, and independent across i, conditional on $\mathcal{C}$. Notice the conditional mean independence restriction in Assumption $5(i)$ is weaker than Assumption D of Bai (2009), besides sequential exogeneity. Bai imposes independence between e_{it} and $(\{X_{js}, \lambda_j, f_s\}_{j,s})$.

(3) Assumption 5 is sufficient for Assumption 2. To see this, notice $\mathrm{Tr}(X_k e') = \sum_{i,t} X_{k,it}e_{it}$, and also that the sequential exogeneity and the cross-sectional independence assumption imply $\mathbb{E}\left[\left((NT)^{-1}\sum_{i,t} X_{k,it}e_{it}\right)^2 \,\middle|\, \mathcal{C}\right] = (NT)^{-2}\sum_{i,t}\mathbb{E}\left[\left(X_{k,it}e_{it}\right)^2 \,\middle|\, \mathcal{C}\right]$. Then, together with the assumption of bounded moments, we have $(NT)^{-1}\sum_{i,t} X_{k,it}e_{it} = o_p(1)$.

(4) Assumption 5 is also sufficient for Assumption 3^* (and thus for Assumption 3), because e_{it} is assumed independent over t and across i and has a bounded fourth moment, conditional on $\mathcal{C}$, which by using results in

Latala (2005), implies the spectral norm satisfies $\|e\| = \sqrt{\max(N,T)}$ as N and T become large; see the supplementary material.

(5) Examples of regressor processes, which satisfy Assumptions 5(iv) and (v), are discussed in the following. These examples also illuminate the role of the conditioning sigma field $\mathcal{C}$.

Examples of DGPs for X_{it}.

Here we provide examples of the DGPs of the regressors X_{it} that satisfy the conditions in Assumption 5. Proofs for these examples are provided in the supplementary material.

Example 1

The first example is a simple AR(1) interactive fixed effect regression:

$$Y_{it} = \beta^0 Y_{i,t-1} + \lambda_i^{0\prime} f_t^0 + e_{it},$$

where e_{it} is mean zero, independent across i and t, and independent of λ^0 and f^0. Assume $|\beta^0| < 1$ and that e_{it}, λ_i^0, and f_t^0 all possess uniformly bounded moments of order $8+\epsilon$. In this case, the regressor is $X_{it} = Y_{it-1} = \lambda_i^{0\prime} F_t^0 + U_{it}$, where $F_t^0 = \sum_{s=0}^{\infty} (\beta^0)^s f_{t-1-s}^0$ and $U_{it} = \sum_{s=0}^{\infty} (\beta^0)^s e_{i,t-1-s}$. For the conditioning sigma field $\mathcal{C}$ in Assumption 5, we choose $\mathcal{C} = \sigma\left(\{\lambda_i^0 : 1 \le i \le N\}, \{f_t^0 : 1 \le t \le T\}\right)$. Conditional on $\mathcal{C}$, the only variation in X_{it} stems from U_{it}, which is independent across i and weakly correlated over t, so that Assumption 5(iv) holds. Furthermore, we have $\mathbb{E}(X_{it}|\mathcal{C}) = \lambda_i^{0\prime} F_t^0$ and $\widetilde{X}_{it} = U_{it}$, which allows us to verify Assumption 5(v).

This example can be generalized to a VAR(1) model as follows:

$$\begin{pmatrix} Y_{it} \\ Z_{it} \end{pmatrix} = \mathcal{B} \underbrace{\begin{pmatrix} Y_{i,t-1} \\ Z_{i,t-1} \end{pmatrix}}_{=X_{it}} + \begin{pmatrix} \lambda_i^{0\prime} f_t^0 \\ d_{it} \end{pmatrix} + \underbrace{\begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}}_{=E_{it}}, \tag{5}$$

where Z_{it} is an $m \times 1$ vector of additional variables and $\mathcal{B}$ is an $(m+1) \times (m+1)$ matrix of VAR parameters whose eigenvalues lie within the unit circle. The $m \times 1$ vector d_{it} and the factors f_t^0 and factor loadings λ_i^0 are assumed to be independent of the $(m+1) \times 1$ vector of innovations E_{it}. Suppose our interest is to estimate the first row in Equation (5), which corresponds exactly to our interactive fixed effects model with regressors $Y_{i,t-1}$ and $Z_{i,t-1}$. Choosing $\mathcal{C}$ to be the sigma field generated by all f_t^0, λ_i^0, d_{it}, we obtain $\widetilde{X}_{it} = \sum_{s=0}^{\infty} \mathcal{B}^s E_{i,t-1-s}$. Analogous to the AR(1) case, we then find Assumption 5(iv) and (v) are satisfied in this example if the innovations E_{it} are independent across i and over t, and have appropriate bounded moments.

Example 2

Consider a scalar X_{it} for simplicity, and let $X_{it} = g(v_{it}, \delta_i, h_t)$. We assume (i) $\left\{(e_{it}, v_{it})_{i=1,\ldots,N;t=1,\ldots,T}\right\} \perp \left\{\left(\lambda_i^0, \delta_i\right)_{i=1,\ldots,N}, \left(f_t^0, h_t\right)_{t=1,\ldots,T}\right\}$, (ii) $(e_{it}, v_{it}, \delta_i)$

are independent across i for all t, and (iii) $v_{is} \perp e_{it}$ for $s \leq t$ and all i. Furthermore, assume $\sup_{it} \mathbb{E}|X_{it}|^{8+\epsilon} < \infty$ for some positive ϵ. For the conditioning sigma field $\mathcal{C}$ in Assumption 5, we choose $\mathcal{C} = \sigma\left(\left\{\lambda_i^0 : 1 \leq i \leq N\right\}, \{\delta_i : 1 \leq i \leq N\}, \left\{f_t^0 : -\infty \leq t \leq \infty\right\}, \{h_t : -\infty \leq t \leq \infty\}\right)$. Furthermore, as in Hahn & Kuersteiner (2011), let $\mathcal{F}_\tau^t(i) = \mathcal{C} \vee \sigma\left(\{(e_{is}, v_{is}) : \tau \leq s \leq t\}\right)$, and define the conditional α-mixing coefficient on $\mathcal{C}$:

$$\alpha_m(i) = \sup_{A \in \mathcal{F}_{-\infty}^t(i), B \in \mathcal{F}_{t+m}^\infty(i)} \left[\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \,|\mathcal{C}\right].$$

Let $\alpha_m = \sup_i \alpha_m(i)$, and assume $\alpha_m = O\left(m^{-\zeta}\right)$, where $\zeta > \frac{12p}{4p-1}$ for $p > 4$. Then, Assumptions $5(iv)$ and (v) are satisfied.

In this example, the shocks h_t (which may contain the factors f_t^0), δ_i (which may contain the factor loadings λ_i^0), and v_{it} (which may contain past values of e_{it}) can enter in a general nonlinear way into the regressor X_{it}.

The following assumption guarantees the limiting variance and the asymptotic bias converge to constant values.

Assumption 6. Let $\mathcal{X}_k = M_{\lambda^0} X_k M_{f^0}$, which is an $N \times T$ matrix with entries $\mathcal{X}_{k,it}$. For each i and t, define the K-vector $\mathcal{X}_{it} = (\mathcal{X}_{1,it}, \ldots, \mathcal{X}_{K,it})'$. We assume existence of the following probability limits for all $k = 1, \ldots, K$:

$$W = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathcal{X}_{it} \mathcal{X}_{it}',$$

$$\Omega = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T e_{it}^2 \mathcal{X}_{it} \mathcal{X}_{it}',$$

$$B_{1,k} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{N} \operatorname{Tr}\left[P_{f^0} \mathbb{E}\left(e' X_k \,|\mathcal{C}\right)\right],$$

$$B_{2,k} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{T} \operatorname{Tr}\left[\mathbb{E}\left(ee' \,|\mathcal{C}\right) M_{\lambda^0} X_k f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime}\right],$$

$$B_{3,k} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{N} \operatorname{Tr}\left[\mathbb{E}\left(e'e \,|\mathcal{C}\right) M_{f^0} X_k' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime}\right],$$

where $\mathcal{C}$ is the same conditioning set that appears in Assumption 5.

Here, W and Ω are $K \times K$ matrices, and we define the K-vectors B_1, B_2, and B_3 with components $B_{1,k}$, $B_{2,k}$ and $B_{3,k}$, $k = 1, \ldots, K$.

THEOREM 4.3 (Asymptotic Distribution). *Let Assumptions 1, 4, 5, and 6 be satisfied,*[14] *and consider the limit $N, T \to \infty$ with $N/T \to \kappa^2$, where $0 < \kappa < \infty$. Then we have*

$$\sqrt{NT}\left(\widehat{\beta} - \beta^0\right) \underset{d}{\to} \mathcal{N}\left(W^{-1} B,\ W^{-1} \Omega W^{-1}\right),$$

where $B = -\kappa B_1 - \kappa^{-1} B_2 - \kappa B_3$.

From Corollary 4.2, we already know the limiting distribution of $\widehat{\beta}$ is given by the limiting distribution of $W_{NT}^{-1} C_{NT}$. Note $W_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathcal{X}_{it} \mathcal{X}_{it}'$; that is, W is simply defined as the probability limit of W_{NT}. Assumption 4 guarantees W is positive definite, as shown in the supplementary material.

Thus, the main task in proving Theorem 4.3 is to show the approximated score at the true parameter satisfies $C_{NT} \to_d \mathcal{N}(B, \Omega)$. We find the asymptotic variance Ω and the asymptotic bias B_1 originate from the $C^{(1)}$ term, whereas the two further bias terms B_2 and B_3 originate from the $C^{(2)}$ term of C_{NT}.

The bias B_1 is due to correlation of the errors e_{it} and the regressors $X_{k,i\tau}$ in the time direction (for $\tau > t$). This bias term generalizes the Nickell (1981) bias that occurs in dynamic models with standard fixed effects, and it is not present in Bai (2009), where only strictly exogenous regressors are considered.

The other two bias terms B_2 and B_3 are already described in Bai (2009). If e_{it} is homoscedastic, that is, if $\mathbb{E}(e_{it}|\mathcal{C}) = \sigma^2$, then $\mathbb{E}\left(ee'|\mathcal{C}\right) = \sigma^2 \mathbb{I}_N$ and $\mathbb{E}\left(e'e|\mathcal{C}\right) = \sigma^2 \mathbb{I}_T$, so that $B_2 = 0$ and $B_3 = 0$ (because the trace is cyclical and $f^{0\prime} M_{f^0} = 0$ and $\lambda^{0\prime} M_{\lambda^0} = 0$). Thus, B_2 is only nonzero if e_{it} is heteroscedastic across i, and B_3 is only nonzero if e_{it} is heteroscedastic over t. Correlation in e_{it} across i or over t would also generate nonzero bias terms of exactly the form B_2 and B_3, but is ruled out by our assumptions.

4.3. Bias Correction

Estimators for W, Ω, B_1, B_2, and B_3 are obtained by forming suitable sample analogs and replacing the unobserved λ^0, f^0, and e by the estimates $\widehat{\lambda}$, $\widehat{f}$, and the residuals $\widehat{e}$.

DEFINITION 1. *Let $\widehat{\mathcal{X}}_k = M_{\widehat{\lambda}} X_k M_{\widehat{f}}$. For each i and t, define the K-vector $\widehat{\mathcal{X}}_{it} = (\widehat{\mathcal{X}}_{1,it}, \ldots, \widehat{\mathcal{X}}_{K,it})'$. Let $\Gamma : \mathbb{R} \to \mathbb{R}$ be the truncation kernel defined by $\Gamma(x) = 1$ for $|x| \leq 1$, and $\Gamma(x) = 0$ otherwise. Let M be a bandwidth parameter that depends on N and T. We define the $K \times K$ matrices $\widehat{W}$ and $\widehat{\Omega}$, and the K-vectors $\widehat{B}_1$, $\widehat{B}_2$, and $\widehat{B}_3$ as follows:*

$$\widehat{W} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\mathcal{X}}_{it} \widehat{\mathcal{X}}_{it}',$$

$$\widehat{\Omega} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{e}_{it})^2 \widehat{\mathcal{X}}_{it} \widehat{\mathcal{X}}_{it}',$$

$$\widehat{B}_{1,k} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \Gamma\left(\frac{s-t}{M}\right) \left[P_{\widehat{f}}\right]_{ts} \widehat{e}_{it} X_{k,is},$$

$$\widehat{B}_{2,k} = \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{e}_{it})^2 \left[M_{\widehat{\lambda}} X_k \widehat{f} (\widehat{f}' \widehat{f})^{-1} (\widehat{\lambda}' \widehat{\lambda})^{-1} \widehat{\lambda}'\right]_{ii},$$

$$\widehat{B}_{3,k} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{e}_{it})^2\left[M_{\widehat{f}}X_k'\widehat{\lambda}\,(\widehat{\lambda}'\widehat{\lambda})^{-1}(\widehat{f}'\widehat{f})^{-1}\,\widehat{f}'\right]_{tt},$$

where $\widehat{e} = Y - \widehat{\beta}\cdot X - \widehat{\lambda}\,\widehat{f}'$, *and* $\widehat{e}_{it}$ *denotes the elements of* $\widehat{e}$, $[A]_{ts}$ *denotes the (t,s)th element of the matrix A.*

Notice the estimators $\widehat{\Omega}$, $\widehat{B}_2$, and $\widehat{B}_3$ are similar to White's standard error estimator under heteroskedasticity, and the estimator $\widehat{B}_1$ is similar to the HAC estimator with a kernel. To show consistency of these estimators, we impose some additional assumptions.

Assumption 7.

(i) $\|\lambda_i^0\|$ and $\|f_t^0\|$ are uniformly bounded over i, t, and N, T.

(ii) There exist $c > 0$ and $\epsilon > 0$ such that for all i, t, m, N, and T, we have $\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(e_{it}X_{k,it+m}\,|\,\mathcal{C})\right| \le c\,m^{-(1+\epsilon)}$.

Assumption $7(i)$ is made for convenience to simplify the consistency proof for the estimators in Definition 1. Weakening this assumption is possible by only assuming suitable bounded moments of $\|\lambda_i^0\|$ and $\|f_t^0\|$. To show consistency of $\widehat{B}_1$, we need to control how strongly e_{it} and $X_{k,i\tau}$, $t < \tau$, are allowed to be correlated, which is done by Assumption $7(ii)$. It is straightforward to verify Assumption $7(ii)$ is satisfied in the two examples of regressor processes presented after Assumption 5.

THEOREM 4.4 (Consistency of Bias and Variance Estimators). *Let Assumptions 1, 4, 5, 6, and 7 hold, and consider a limit* $N, T \to \infty$ *with* $N/T \to \kappa^2$, $0 < \kappa < \infty$, *such that the bandwidth* $M = M_{NT}$ *satisfies* $M \to \infty$ *and* $M^5/T \to 0$. *We then have* $\widehat{W} = W + o_p(1)$, $\widehat{\Omega} = \Omega + o_p(1)$, $\widehat{B}_1 = B_1 + o_p(1)$, $\widehat{B}_2 = B_2 + o_p(1)$, *and* $\widehat{B}_3 = B_3 + o_p(1)$.

The assumption $M^5/T \to 0$ can be relaxed if additional higher- moment restrictions on e_{it} and $X_{k,it}$ are imposed. Note also that for the construction of the estimators $\widehat{W}$, $\widehat{\Omega}$, and $\widehat{B}_i$, $i = 1, 2, 3$, knowing whether the regressors are strictly exogenous or predetermined is unnecessary; in both cases, the estimators for W, Ω, and B_i, $i = 1, 2, 3$, are consistent. We can now present our bias-corrected estimator and its limiting distribution.

COROLLARY 4.5. *Under the assumptions of Theorem 4.4, the bias-corrected estimator*

$$\widehat{\beta}^* = \widehat{\beta} + \widehat{W}^{-1}\left(T^{-1}\widehat{B}_1 + N^{-1}\widehat{B}_2 + T^{-1}\widehat{B}_3\right)$$

satisfies $\sqrt{NT}\left(\widehat{\beta}^* - \beta^0\right) \to_d \mathcal{N}\left(0,\ W^{-1}\Omega W^{-1}\right)$.

According to Theorem 4.4, a consistent estimator of the asymptotic variance of $\widehat{\beta}^*$ is given by $\widehat{W}^{-1}\widehat{\Omega}\widehat{W}^{-1}$.

An alternative to the analytical bias-correction result given by Corollary 4.5 is to use Jackknife bias correction to eliminate the asymptotic bias. For panel models with incidental parameters only in the cross-sectional dimensions, one typical finds a large N, T leading incidental parameter bias of order $1/T$ for the parameters of interest. To correct for this $1/T$ bias, one can use the delete-one Jackknife bias correction if observations are iid over t Hahn & Newey (2004) and the split-panel Jackknife bias-correction if observations are correlated over t Dhaene & Jochmans (2015). In our current model, we have incidental parameters in both panel dimensions (λ_i^0 and f_t^0), resulting in leading bias terms of order $1/T$ (bias term B_1 and B_3) and of order $1/N$ (bias term B_2). Fernández-Val & Weidner (2013) discuss the generalizations of the split-panel Jackknife bias-correction to that case.

The corresponding bias-corrected split-panel Jackknife estimator reads $\widehat{\beta}^J = 3\widehat{\beta}_{NT} - \overline{\beta}_{N,T/2} - \overline{\beta}_{N/2,T}$, where $\widehat{\beta}_{NT} = \widehat{\beta}$ is the LS estimator obtained from the full sample, $\overline{\beta}_{N,T/2}$ is the average of the two LS estimators that leave out the first and second halves of the time periods, and $\overline{\beta}_{N/2,T}$ is the average of the two LS estimators that leave out half of the individuals. Jackknife bias correction is convenient because only the order of the bias, and not the structure of the terms B_1, B_2, and B_3, needs not be known in detail. However, one requires additional stationarity assumptions over t and homogeneity assumptions across i to justify the Jackknife correction and to show that $\widehat{\beta}^J$ has the same limiting distribution as $\widehat{\beta}^*$ in Corollary 4.5; see Fernández-Val & Weidner (2013) for more details. They also observe through Monte Carlo simulations that the finite sample variance of the Jackknife-corrected estimator is often larger than of the analytically corrected estimator. We do not explore Jackknife bias-correction further in this paper.

5. TESTING RESTRICTIONS ON β^0

In this section, we discuss the three classical test statistics for testing linear restrictions on β^0. The null hypothesis is $H_0: \ H\beta^0 = h$, and the alternative is $H_a: \ H\beta^0 \neq h$, where H is an $r \times K$ matrix of rank $r \leq K$, and h is an $r \times 1$ vector. We restrict the presentation to testing a linear hypothesis for ease of exposition. One can generalize the discussion to the testing of nonlinear hypotheses, under conventional regularity conditions. Throughout this subsection, we assume β^0 is an interior point of $\mathbb{B}$; that is, no local restrictions are on β as long as the null hypothesis is not imposed. Using the expansion of $L_{NT}(\beta)$, one could also discuss testing when the true parameter is on the boundary, as shown in Andrews (2001).

The restricted estimator is defined by

$$\widetilde{\beta} = \underset{\beta \in \widetilde{\mathbb{B}}}{\operatorname{argmin}} \, L_{NT}(\beta), \tag{6}$$

where $\widetilde{\mathbb{B}} = \{\beta \in \mathbb{B} | \, H\beta = h\}$ is the restricted parameter set. Analogous to Theorem 4.3 for the unrestricted estimator $\widehat{\beta}$, we can use the expansion of the profile

objective function to derive the limiting distribution of the restricted estimator. Under the assumptions of Theorem 4.3, we have

$$\sqrt{NT}(\widetilde{\beta}-\beta^0) \underset{d}{\longrightarrow} \mathcal{N}\left(\mathfrak{W}^{-1}B,\, \mathfrak{W}^{-1}\Omega\mathfrak{W}^{-1}\right),$$

where $\mathfrak{W}^{-1} = W^{-1} - W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}$. The $K \times K$ covariance matrix in the limiting distribution of $\widetilde{\beta}$ is not full rank, but satisfies $\operatorname{rank}(\mathfrak{W}^{-1}\Omega\mathfrak{W}^{-1}) = K-r$, because $H\mathfrak{W}^{-1} = 0$ and thus $\operatorname{rank}(\mathfrak{W}^{-1}) = K-r$. The asymptotic distribution of $\sqrt{NT}(\widetilde{\beta}-\beta^0)$ is therefore $K-r$ dimensional, as it should be for the restricted estimator.

Wald Test

Using the result of Theorem 4.3, we find that under the null hypothesis, $\sqrt{NT}\left(H\widehat{\beta}-h\right)$ is asymptotically distributed as $\mathcal{N}\left(HW^{-1}B,\, HW^{-1}\Omega W^{-1}H'\right)$. Thus, due to the presence of the bias B, the standard Wald test statistic $WD_{NT} = NT\left(H\widehat{\beta}-h\right)'\left(H\widehat{W}^{-1}\widehat{\Omega}\widehat{W}^{-1}H'\right)^{-1}\left(H\widehat{\beta}-h\right)$ is not asymptotically χ^2_r distributed. Using the estimator $\widehat{B} = -\sqrt{\frac{N}{T}}\widehat{B}_1 - \sqrt{\frac{T}{N}}\widehat{B}_2 - \sqrt{\frac{N}{T}}\widehat{B}_3$ for the bias, we can define the bias-corrected Wald test statistic as

$$WD^*_{NT} = \left[\sqrt{NT}\left(H\widehat{\beta}^*-h\right)\right]'\left(H\widehat{W}^{-1}\widehat{\Omega}\widehat{W}^{-1}H'\right)^{-1}\left[\sqrt{NT}\left(H\widehat{\beta}^*-h\right)\right], \quad (7)$$

where $\widehat{\beta}^* = \widehat{\beta} - \widehat{W}^{-1}\widehat{B}$ is the bias-corrected estimator. WD^*_{NT} is just the standard Wald test statistics applied to $\widehat{\beta}^*$. Under the null hypothesis and the Assumptions of Theorem 4.4, we find $WD^*_{NT} \to_d \chi^2_r$.

Likelihood Ratio Test

To implement the LR test, we need the relationship between the asymptotic Hessian W and the asymptotic score variance Ω of the profile objective function to be of the form $\Omega = cW$, where $c > 0$ is a scalar constant. This condition is satisfied in our interactive fixed effect model if $\mathbb{E}(e_{it}^2|\mathcal{C}) = c$, that is, if the error is homoskedastic. A consistent estimator for c is then given by $\widehat{c} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{e}_{it}^2$, where $\widehat{e} = Y - \widehat{\beta}\cdot X - \widehat{\lambda}\,\widehat{f}'$. Because the likelihood function for the interactive fixed effect model is just the sum of squared residuals, we have $\widehat{c} = L_{NT}(\widehat{\beta})$. The likelihood ratio test statistic is defined by

$$LR_{NT} = \widehat{c}^{-1}\,NT\left[L_{NT}\left(\widetilde{\beta}\right) - L_{NT}\left(\widehat{\beta}\right)\right].$$

Under the assumption of Theorem 4.3, we then have

$$LR_{NT} \underset{d}{\longrightarrow} c^{-1}C'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}C,$$

where $C \sim \mathcal{N}(B,\Omega)$, i.e. $C_{NT} \to_d C$. It is the same limiting distribution that one finds for the Wald test if $\Omega = cW$ (in fact, one can show $WD_{NT} = LR_{NT} + o_p(1)$).

Therefore, we need to do a bias-correction for the LR test to achieve a χ^2 limiting distribution. We define

$$LR^*_{NT} = \widehat{c}^{-1} NT \left[\min_{\{\beta\in\mathbb{B}|\, H\beta=h\}} L_{NT}\left(\beta + (NT)^{-1/2}\widehat{W}^{-1}\widehat{B}\right) - \min_{\beta\in\mathbb{B}} L_{NT}\left(\beta + (NT)^{-1/2}\widehat{W}^{-1}\widehat{B}\right)\right], \tag{8}$$

where $\widehat{B}$ and $\widehat{W}$ do not depend on the parameter β in the minimization problem.[15] Asymptotically, we have $\min_{\beta\in\mathbb{B}} L_{NT}\left(\beta + (NT)^{-1/2}\widehat{W}^{-1}\widehat{B}\right) = L_{NT}(\widehat{\beta})$, because $\beta \in \mathbb{B}$ does not impose local constraints; in other words, close to β^0, whether one minimizes over β or over $\beta + (NT)^{-1/2}\widehat{W}^{-1}\widehat{B}$ does not matter for the value of the minimum. The correction to the LR test therefore originates from the first term in LR^*_{NT}. For the minimization over the restricted parameter set, whether the argument of L_{NT} is β or $\beta + (NT)^{-1/2}\widehat{W}^{-1}\widehat{B}$ matters, because generically, we have $HW^{-1}B \neq 0$ (otherwise, no correction would be necessary for the LR statistics). One can show that

$$LR^*_{NT} \underset{d}{\longrightarrow} c^{-1}(C-B)'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}(C-B)\,;$$

that is, we obtain the same formula as for LR_{NT}, but the bias-corrected term $C - B$ replaces the limit of the score C. Under the Assumptions of Theorem 4.4, if H_0 is satisfied, and for homoscedastic errors e_{it}, we have $LR^*_{NT} \to_d \chi^2_r$. In fact, one can show $LR^*_{NT} = WD^*_{NT} + o_p(1)$.

Lagrange Multiplier Test

Let $\widetilde{\nabla}\mathcal{L}_{NT}$ be the gradient of the LS objective function (2) with respect to β, evaluated at the restricted parameter estimates; that is,

$$\widetilde{\nabla}\mathcal{L}_{NT} = \nabla\mathcal{L}_{NT}(\widetilde{\beta},\widetilde{\lambda},\widetilde{f}) = \left(\left.\frac{\partial\mathcal{L}_{NT}(\beta,\widetilde{\lambda},\widetilde{f})}{\partial\beta_1}\right|_{\beta=\widetilde{\beta}},\ldots,\left.\frac{\partial\mathcal{L}_{NT}(\beta,\widetilde{\lambda},\widetilde{f})}{\partial\beta_K}\right|_{\beta=\widetilde{\beta}}\right)' = -\frac{2}{NT}\left(\mathrm{Tr}\left(X_1'\widetilde{e}\right),\ldots,\mathrm{Tr}\left(X_K'\widetilde{e}\right)\right)',$$

where $\widetilde{\lambda} = \widehat{\lambda}(\widetilde{\beta})$, $\widetilde{f} = \widehat{f}(\widetilde{\beta})$, and $\widetilde{e} = Y - \widetilde{\beta}\cdot X - \widetilde{\lambda}\,\widetilde{f}'$. Under the assumptions of Theorem 4.3, and if the null hypothesis $H_0:\ H\beta^0 = h$ is satisfied, one finds that[16]

$$\sqrt{NT}\,\widetilde{\nabla}\mathcal{L}_{NT} = \sqrt{NT}\,\nabla L_{NT}(\widetilde{\beta}) + o_p(1). \tag{9}$$

Due to this equation, one can base the Lagrange multiplier test on the gradient of $\mathcal{L}_{NT}(\widetilde{\beta},\widetilde{\lambda},\widetilde{f})$, or on the gradient of the profile quasilikelihood function $L_{NT}(\widetilde{\beta})$, and obtain the same limiting distribution.

Using the bound on the remainder $R_{NT}(\beta)$ given in Theorem 4.1, one cannot infer any properties of the score function, that is, of the gradient $\nabla L_{NT}(\beta)$,

because nothing is said about $\nabla R_{NT}(\beta)$. The following theorem gives a bound on $\nabla R_{NT}(\beta)$ that is sufficient to derive the limiting distribution of the Lagrange multiplier.

THEOREM 5.1. *Under the assumptions of Theorem 4.1, and with W_{NT} and C_{NT} as defined there, the score function satisfies*

$$\nabla L_{NT}(\beta) = 2\,W_{NT}\,(\beta-\beta^0) - \frac{2}{\sqrt{NT}}\,C_{NT} + \frac{1}{NT}\,\nabla R_{NT}(\beta),$$

where the remainder $\nabla R_{NT}(\beta)$ satisfies for any sequence $\eta_{NT} \to 0$:

$$\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}} \frac{\|\nabla R_{NT}(\beta)\|}{\sqrt{NT}\left(1+\sqrt{NT}\,\|\beta-\beta^0\|\right)} = o_p(1).$$

From this theorem, and the fact that $\widetilde{\beta}$ is $\sqrt{NT}$-consistent under H_0, we obtain

$$\begin{aligned}\sqrt{NT}\,\widetilde{\nabla}\mathcal{L}_{NT} &= \sqrt{NT}\,\nabla L_{q,NT}(\widetilde{\beta}) + o_p(1)\\ &= 2\sqrt{NT}\,W_{NT}\,(\widetilde{\beta}-\beta^0) - 2\,C_{NT} + o_p(1).\end{aligned}$$

Remember $\widetilde{\beta}$ is the restricted estimator defined in Equation (6). Using this result and the known limiting distribution of $\widetilde{\beta}$, we now find

$$\sqrt{NT}\,\widetilde{\nabla}\mathcal{L}_{NT} \underset{d}{\longrightarrow} -2H'(HW^{-1}H')^{-1}HW^{-1}C. \tag{10}$$

Note also that $\sqrt{NT}\,HW^{-1}\nabla L_{NT}(\widetilde{\beta}) \to_d -2HW^{-1}C$. We define $\widetilde{B}$, $\widetilde{W}$, and $\widetilde{\Omega}$, analogous to $\widehat{B}$, $\widehat{W}$, and $\widehat{\Omega}$, but with unrestricted parameter estimates replaced by restricted parameter estimates. The LM test statistic is then given by

$$LM_{NT} = \frac{NT}{4}\,(\widetilde{\nabla}\mathcal{L}_{NT})'\widetilde{W}^{-1}H'(H\widetilde{W}^{-1}\widetilde{\Omega}\widetilde{W}^{-1}H')^{-1}H\widetilde{W}^{-1}\widetilde{\nabla}\mathcal{L}_{NT}.$$

One can show the LM test is asymptotically equivalent to the Wald test: $LM_{NT} = WD_{NT} + o_p(1)$; that is, again, bias-correction is necessary. We define the bias-corrected LM test statistic as

$$\begin{aligned}LM^*_{NT} = \frac{1}{4}&\left(\sqrt{NT}\,\widetilde{\nabla}\mathcal{L}_{NT} + 2\widetilde{B}\right)'\widetilde{W}^{-1}H'\left(H\widetilde{W}^{-1}\widetilde{\Omega}\widetilde{W}^{-1}H'\right)^{-1}H\widetilde{W}^{-1}\\ &\times\left(\sqrt{NT}\,\widetilde{\nabla}\mathcal{L}_{NT} + 2\widetilde{B}\right).\end{aligned} \tag{11}$$

The following theorem summarizes the main results of the present subsection.

THEOREM 5.2 (Chi-Square Limit of Bias-Corrected Test Statistics). *Let the assumptions of Theorem 4.4 and the null hypothesis $H_0: H\beta^0 = h$ be satisfied. For the bias-corrected Wald and LM test statistics introduced in Equation (7) and (11), we then have*

$$WD^*_{NT} \underset{d}{\longrightarrow} \chi^2_r, \qquad LM^*_{NT} \underset{d}{\longrightarrow} \chi^2_r.$$

If, in addition, we assume $\mathbb{E}(e_{it}^2|\mathcal{C}) = c$*, that is, the idiosyncratic errors are homoscedastic, and we use* $\widehat{c} = L_{NT}(\widehat{\beta})$ *as an estimator for c, the LR test statistic defined in Equation (8) satisfies*

$$LR_{NT}^{*} \xrightarrow[d]{} \chi_r^2.$$

6. EXTENSION TO ENDOGENOUS REGRESSORS

In this section, we briefly discuss how to estimate the regression coefficient β^0 of Model (1) when some of the regressors in X_{it} are endogenous with respect to the regression error e_{it}. The question is how instrumental variables can be used to estimate the regression coefficients of the endogenous regressor in the presence of the interactive fixed effects $\lambda_i^{0\prime} f_t^0$.

The existing literature has already investigated similar questions under various setups. Harding & Lamarche (2009; 2011) investigate the problem of estimating an endogenous panel (quantile) regression with interactive fixed effects, and show how to use IVs in the CCE estimation framework. Moon, Shum, and Weidner (2012) (hereafter MSW) estimate a random coefficient multinomial demand model (as in Berry, Levinsohn, & Pakes (1995)) when the unobserved product-market characteristics have interactive fixed effects. The IVs are required to identify the parameters of the random coefficient distribution and to control for price endogeneity. They suggested a multi-step "least squares-minimum distance" (LS-MD) estimator.[17] The LS-MD approach is also applicable to linear panel regression models with endogenous regressors and interactive fixed effects, as demonstrated in Lee, Moon, & Weidner (2012) for the case of a dynamic linear panel regression model with interactive fixed effects and measurement error.

We now discuss how to implement the LS-MD estimation in our setup. Let X_{it}^{end} be the vectors of endogenous regressors, and let X_{it}^{exo} be the vector of exogenous regressors, with respect to e_{it}, such that $X_{it} = (X_{it}^{\text{end}\prime}, X_{it}^{\text{exo}\prime})'$. The model then reads

$$Y_{it} = \beta_{\text{end}}^{0\prime} X_{it}^{\text{end}} + \beta_{\text{exo}}^{0\prime} X_{it}^{\text{exo}} + \lambda_i^{0\prime} f_t^0 + e_{it},$$

where X_{it}^{exo} denotes the exogenous and X_{it}^{end} denotes the endogenous regressors (wrt to e_{it}). Suppose Z_{it} is an additional L-vector of exogenous instrumental variables (IVs), but Z_{it} may be correlated with λ_i^0 and f_t^0. The LS-MD estimator of $\beta^0 = \left(\beta_{\text{end}}^{0\prime}, \beta_{\text{exo}}^{0\prime}\right)'$ can then be calculated by the following three steps:

(1) For given β_{end}, we run the least squares regression of $Y_{it} - \beta_{\text{end}}' X_{it}^{\text{end}}$ on the included exogeneous regressors X_{it}^{exo}, the interactive fixed effects $\lambda_i' f_t$, and the IVs Z_{it} :

$$\left(\widetilde{\beta}_{\text{exo}}(\beta_{\text{end}}), \widetilde{\gamma}(\beta_{\text{end}}), \widetilde{\lambda}(\beta_{\text{end}}), \widetilde{f}(\beta_{\text{end}})\right)$$
$$= \underset{\{\beta_{\text{exo}}, \gamma, \lambda, f\}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - \beta_{\text{end}}' X_{it}^{\text{end}} - \beta_{\text{exo}}' X_{it}^{\text{exo}} - \gamma' Z_{it} - \lambda_i' f_t\right)^2.$$

(2) We estimate β_{end} by finding $\widetilde{\gamma}(\beta_{\text{end}})$, obtained by step (1), that is closest to zero. To do so, we choose a symmetric positive definite $L \times L$ weight matrix W_{NT}^{γ} and compute

$$\widehat{\beta}_{\text{end}} = \underset{\beta_{\text{end}}}{\text{argmin}}\, \widetilde{\gamma}(\beta_{\text{end}})' \, W_{NT}^{\gamma} \, \widetilde{\gamma}(\beta_{\text{end}}) .$$

(3) We estimate β_{exo} (and λ, f) by running the least squares regression of $Y_{it} - \widehat{\beta}_{\text{end}}' X_{it}^{\text{end}}$ on the included exogeneous regressors X_{it}^{exo} and the interactive fixed effects $\lambda_i' f_t$:

$$\left(\widehat{\beta}_{\text{exo}}, \widehat{\lambda}, \widehat{f}\right) = \underset{\{\beta_{\text{exo}}, \gamma, \lambda, f\}}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \widehat{\beta}_{\text{end}}' X_{it}^{\text{end}} - \beta_{\text{exo}}' X_{it}^{\text{exo}} - \lambda_i' f_t \right)^2 .$$

The idea behind this estimation procedure is that valid instruments are excluded from the model for Y_{it}, so that their first-step regression coefficients $\widetilde{\gamma}(\beta_{\text{end}})$ should be close to zero if β_{end} is close to its true value β_{end}^0. Thus, as long as X_{it}^{exo} and Z_{it} jointly satisfy the assumptions of the current paper, we obtain $\widetilde{\gamma}(\beta_{\text{end}}^0) = o_p(1)$ for the first-step LS estimator, and we also obtain the asymptotic distribution of $\widetilde{\gamma}(\beta_{\text{end}}^0)$ from the results derived in Section 4.

However, to justify the second-step minimization formally, one needs to study the properties of $\widetilde{\gamma}(\beta_{\text{end}})$ also for $\beta_{\text{end}} \neq \beta_{\text{end}}^0$. To do so, we refer to MSW. Our $\beta_{\text{end}}, \beta_{\text{exo}}$, and $Y_{it} - \beta_{\text{end}}' X_{it}^{\text{end}}$ correspond to their α, β, and $\delta_{jt}(\alpha)$, respectively. Assumptions 1–5 in MSW can be translated accordingly, and the results in MSW show large N, T consistency and asymptotic normality of the LS-MD estimator.

The final step of the LS-MD estimation procedure is essentially a repetition of the first step, but without including Z_{it} in the set of regressors, which results in some efficiency gains for $\widehat{\beta}_{\text{exo}}$ compared to the first step.

7. MONTE CARLO SIMULATIONS

We consider an AR(1) model with $R = 1$ factors:

$$Y_{it} = \rho^0 Y_{i,t-1} + \lambda_i^0 f_t^0 + e_{it}.$$

We estimate the model as an interactive fixed effect model; that is, no distributional assumptions on λ_i^0 and f_t^0 are made in estimation. The parameter of interest is ρ^0. The estimators we consider are the OLS estimator (which completely ignores the presence of the factors), the least squares estimator with interactive fixed effects (denoted FLS in this section to differentiate from OLS) defined in Equation (3),[18] and its bias-corrected version (denoted BC-FLS), defined in Theorem 4.5.

For the simulation, we draw the e_{it} independently and identically distributed from a t-distribution with five degrees of freedom, the λ_i^0 independently distributed from $\mathcal{N}(1,1)$, and we generate the factors from an AR(1) specification, namely, $f_t^0 = \rho_f f_{t-1}^0 + u_t$, where $u_t \sim \text{iid}\mathcal{N}(0, (1-\rho_f^2)\sigma_f^2)$, and σ_f is the standard deviation of f_t^0. For all simulations, we generate 1,000 initial time

periods for f_t^0 and Y_{it} that are not used for estimation. This approach guarantees the simulated data used for estimation are distributed according to the stationary distribution of the model.

This setup contains no correlation and heteroscedasticity in e_{it}; that is, only the bias term B_1 of the FLS estimator is nonzero, but we ignore this information in the estimation; that is, we correct for all three bias terms (B_1, B_2, and B_3, as introduced in Assumption 6) in the bias-corrected FLS estimator.

Table 1 shows the simulation results for the bias, standard error, and root mean square error of the three different estimators for the case $N = 100$, $\rho_f = 0.5$, and $\sigma_f = 0.5$, and different values of ρ^0 and T. The OLS estimator, the FLS estimator (computed with correct $R = 1$), and the corresponding bias-corrected FLS estimator with factors (BC-FLS) were computed for 10,000 simulation runs. The table lists the mean bias, the standard deviation (std), and the square root of the mean square error (rmse) for the three estimators. As expected, the OLS estimator is biased because of the factor structure and its bias does not vanish (it actually increases) as T increases. The FLS estimator is also biased, but as theory predicts its bias vanishes as T increases. The bias-corrected FLS estimator performs better than the noncorrected FLS estimator, in particular, its bias vanishes faster. Because we only correct for the first-order bias of the FLS estimator, we could

TABLE 1. Simulation results for the AR(1) model described in the main text with $N = 100$, $\rho_f = 0.5$, $\sigma_f = 0.5$, and different values of T (with corresponding bandwidth M) and true AR(1) coefficient ρ^0.

		$\rho^0 = 0.3$			$\rho^0 = 0.9$		
		OLS	FLS	BC-FLS	OLS	FLS	BC-FLS
$T = 5$	bias	0.1232	−0.1419	−0.0713	0.0200	−0.3686	−0.2330
$(M = 2)$	std	0.1444	0.1480	0.0982	0.0723	0.1718	0.1301
	rmse	0.1898	0.2050	0.1213	0.0750	0.4067	0.2669
$T = 10$	bias	0.1339	−0.0542	−0.0201	0.0218	−0.1019	−0.0623
$(M = 3)$	std	0.1148	0.0596	0.0423	0.0513	0.1094	0.0747
	rmse	0.1764	0.0806	0.0469	0.0557	0.1495	0.0973
$T = 20$	bias	0.1441	−0.0264	−0.0070	0.0254	−0.0173	−0.0085
$(M = 4)$	std	0.0879	0.0284	0.0240	0.0353	0.0299	0.0219
	rmse	0.1687	0.0388	0.0250	0.0434	0.0345	0.0235
$T = 40$	bias	0.1517	−0.0130	−0.0021	0.0294	−0.0057	−0.0019
$(M = 5)$	std	0.0657	0.0170	0.0160	0.0250	0.0105	0.0089
	rmse	0.1654	0.0214	0.0161	0.0386	0.0119	0.0091
$T = 80$	bias	0.1552	−0.0066	−0.0007	0.0326	−0.0026	−0.0006
$(M = 6)$	std	0.0487	0.0112	0.0109	0.0179	0.0056	0.0053
	rmse	0.1627	0.0130	0.0109	0.0372	0.0062	0.0053

TABLE 2. Same DGP as Table 1, but misspecification in number of factors R is present. The true number of factors is $R = 1$, but the FLS and BC-FLS are calculated with $R = 2$.

		$\rho^0 = 0.3$			$\rho^0 = 0.9$		
		OLS	FLS	BC-FLS	OLS	FLS	BC-FLS
$T = 5$	bias	0.1239	−0.5467	−0.3721	0.0218	−0.9716	−0.7490
$(M = 2)$	std	0.1454	0.1528	0.1299	0.0731	0.1216	0.1341
	rmse	0.1910	0.5676	0.3942	0.0763	0.9792	0.7609
$T = 10$	bias	0.1343	−0.1874	−0.1001	0.0210	−0.4923	−0.3271
$(M = 3)$	std	0.1145	0.1159	0.0758	0.0518	0.1159	0.0970
	rmse	0.1765	0.2203	0.1256	0.0559	0.5058	0.3412
$T = 20$	bias	0.1451	−0.0448	−0.0168	0.0255	−0.1822	−0.1085
$(M = 4)$	std	0.0879	0.0469	0.0320	0.0354	0.0820	0.0528
	rmse	0.1696	0.0648	0.0362	0.0436	0.1999	0.1207
$T = 40$	bias	0.1511	−0.0161	−0.0038	0.0300	−0.0227	−0.0128
$(M = 5)$	std	0.0663	0.0209	0.0177	0.0250	0.0342	0.0225
	rmse	0.1650	0.0264	0.0181	0.0390	0.0410	0.0258
$T = 80$	bias	0.1550	−0.0072	−0.0011	0.0325	−0.0030	−0.0010
$(M = 6)$	std	0.0488	0.0123	0.0115	0.0182	0.0064	0.0057
	rmse	0.1625	0.0143	0.0116	0.0372	0.0071	0.0058

not expect the bias-corrected FLS estimator to be unbiased. However, as T gets larger, more and more of the FLS estimator bias is corrected for; for example, for $\rho^0 = 0.3$, we find that at $T = 5$, the bias correction only corrects for about half of the bias, whereas at $T = 80$, it already corrects for about 90% of it.

Table 2 is similar to Table 1, with the only difference being that we allow for misspecification in the number of factors R, namely, the true number of factors is assumed to be $R = 1$ (i.e., same DGP as for Table 1), but we incorrectly use $R = 2$ factors when calculating the FLS and BC-FLS estimator. By comparing Table 2 with Table 1, we find this type of misspecification of the number of factors increases the bias and the standard deviation of both the FLS and the BC-FLS estimator in finite samples. That increase, however, is comparatively small once both N and T are large. According to the results in Moon and Weidner (2015), we expect the limiting distribution of the correctly specified ($R = 1$) and incorrectly specified ($R = 2$) FLS estimator to be identical when N and T grow at the same rate. Our simulations suggest the same is true for the BC-FLS estimator. The remaining simulation all assume correctly specified $R = 1$.

An import issue is the choice of bandwidth M for the bias correction. Table 3 gives the fraction of the FLS estimator bias that is captured by the estimator for the bias in a model with $N = 100$, $T = 20$, $\rho_f = 0.5$, $\sigma_f = 0.5$ and different values for ρ and M. The table shows the optimal bandwidth (in the sense that

TABLE 3. Simulation results for the AR(1) model with $N = 100$, $T = 20$, $\rho_f = 0.5$, and $\sigma_f = 0.5$. For different values of the AR(1) coefficient ρ^0 and of the bandwidth M, we give the fraction of the LS estimator bias that is accounted for by the bias correction, i.e. the fraction $\sqrt{NT}\,\mathbb{E}(\widehat{\beta}-\beta)/\mathbb{E}(\widehat{W}^{-1}\widehat{B})$, computed over 10,000 simulation runs. Here and in all following tables it is assumed that $R = 1$ is correctly specified.

	$M=1$	$M=2$	$M=3$	$M=4$	$M=5$	$M=6$	$M=7$	$M=8$
$\rho^0=0$	0.889	0.832	0.791	0.754	0.720	0.689	0.660	0.633
$\rho^0=0.3$	0.752	0.806	0.778	0.742	0.708	0.677	0.648	0.621
$\rho^0=0.6$	0.589	0.718	0.728	0.704	0.674	0.644	0.616	0.590
$\rho^0=0.9$	0.299	0.428	0.486	0.510	0.519	0.516	0.508	0.495

most of the bias is corrected for) depends on ρ^0: it is $M = 1$ for $\rho = 0$, $M = 2$ for $\rho = 0.3$, $M = 3$ and $\rho = 0.6$, and $M = 5$ for $\rho = 0.9$. Choosing too large or too small a bandwidth results in a smaller fraction of the bias to be corrected. Table 4 also reports the properties of the BC-FLS estimator for different values of ρ^0, T, and M. It shows the effect of the bandwidth choice on the standard deviation of the BC-FLS estimator is relatively small at $T = 40$, but is more pronounced at $T = 20$. The issue of optimal bandwidth choice is therefore an important topic for future research. In the simulation results presented here, we tried to choose reasonable values for M, but made no attempt to optimize the bandwidth.

In our setup, we have $\|\lambda^0 f^{0\prime}\| \approx \sqrt{2NT\sigma_f}$ and $\|e\| \approx \sqrt{N}+\sqrt{T}$.[19] Assumptions 1 and 3 imply $\|\lambda^0 f^{0\prime}\| \gg \|e\|$ asymptotically. We can therefore only be sure our asymptotic results for the FLS estimator distribution are a good approximation of the finite sample properties if $\|\lambda^0 f^{0\prime}\| \gtrsim \|e\|$, that is, if $\sqrt{2NT\sigma_f} \gtrsim \sqrt{N}+\sqrt{T}$. To explore this further, we present in Table 5 simulation results for $N = 100$,

TABLE 4. Same specification as Table 1. We only report the properties of the bias-corrected LS estimator, but for multiple values of the bandwidth parameter M and two different values for T. Results were obtained using 10,000 simulation runs.

		BC-FLS for $\rho^0 = 0.3$			BC-FLS for $\rho^0 = 0.9$		
		M=2	M=5	M=8	M=2	M=5	M=8
$T = 20$	bias	−0.0056	−0.0082	−0.0100	−0.0100	−0.0083	−0.0089
	std	0.0239	0.0241	0.0247	0.0253	0.0212	0.0208
	rmse	0.0245	0.0255	0.0266	0.0272	0.0228	0.0227
$T = 40$	bias	−0.0017	−0.0023	−0.0030	−0.0024	−0.0019	−0.0018
	std	0.0159	0.0159	0.0159	0.0095	0.0089	0.0085
	rmse	0.0160	0.0161	0.0162	0.0098	0.0091	0.0087

TABLE 5. Simulation results for the AR(1) model with $N = 100$, $T = 20$, $M = 4$, and $\rho^0 = 0.6$. The three different estimators were computed for 10,000 simulation runs, and the mean bias, standard deviation (std), and root mean square error (rmse) are reported.

		$\rho_f = 0.3$			$\rho_f = 0.7$		
		OLS	FLS	BC-FLS	OLS	FLS	BC-FLS
$\sigma_f = 0$	bias	−0.0007	−0.0076	−0.0043	−0.0004	−0.0074	−0.0041
	std	0.0182	0.0332	0.0243	0.0178	0.0331	0.0242
	rmse	0.0182	0.0340	0.0247	0.0178	0.0339	0.0245
$\sigma_f = 0.2$	bias	0.0153	−0.0113	−0.0032	0.0474	−0.0291	−0.0071
	std	0.0251	0.0303	0.0229	0.0382	0.0387	0.0272
	rmse	0.0294	0.0323	0.0231	0.0609	0.0484	0.0281
$\sigma_f = 0.5$	bias	0.0567	−0.0137	−0.0041	0.1491	−0.0403	−0.0126
	std	0.0633	0.0260	0.0207	0.0763	0.0298	0.0226
	rmse	0.0850	0.0294	0.0211	0.1675	0.0501	0.0259

$T = 20$, $\rho^0 = 0.6$, and different values of ρ_f and σ_f. For $\sigma_f = 0$, we have $0 = \|\lambda^0 f^{0\prime}\| \ll \|e\|$, and this case is equivalent to $R = 0$ (no factor at all). In this case, the OLS estimator estimates the true model and is almost unbiased, and correspondingly, the FLS estimator and the bias-corrected FLS estimator perform worse than OLS in finite samples (though we expect all three estimators are asymptotically equivalent), but the bias-corrected FLS estimator has a lower bias and a lower variance than the noncorrected FLS estimator. The case $\sigma_f = 0.2$ corresponds to $\|\lambda^0 f^{0\prime}\| \approx \|e\|$, and one finds the bias and the variance of the OLS estimator and of the FLS estimator are of comparable size. However, the bias-corrected FLS estimator already has much smaller bias and a bit smaller variance in this case. Finally, in the case $\sigma_f = 0.5$, we have $\|\lambda^0 f^{0\prime}\| > \|e\|$, and we expect our asymptotic results to be a good approximation of this situation. Indeed, one finds that for $\sigma_f = 0.5$, the OLS estimator is heavily biased and very inefficient compared to the FLS estimator, whereas the bias-corrected FLS estimator performs even better in terms of bias and variance.

In Table 6, we present simulation results for the size of the various tests discussed in the last section when testing the null hypothesis $H_0 : \rho = \rho^0$. We choose a nominal size of 5%, $\rho_f = 0.5$, $\sigma_f = 0.5$, and different values for ρ^0, N, and T. In all cases, the size distortions of the uncorrected Wald, LR, and LM test are rather large, and the size distortion of these tests do not vanish as N and T increase: the size for $N = 100$ and $T = 20$ is about the same as for $N = 400$ and $T = 80$, and the size for $N = 400$ and $T = 20$ is about the same as for $N = 1600$ and $T = 80$. By contrast, the size distortions for the bias-corrected Wald, LR, and LM test are much smaller, and tend toward zero (i.e., the size becomes closer to 5%) as N, T increase, holding the ratio N/T constant. For fixed T, an increase in N results in

TABLE 6. Simulation results for the AR(1) model with $\rho_f = 0.5$ and $\sigma_f = 0.5$. For the different values of ρ^0, N, T, and M, we test the hypothesis $H_0 : \rho = \rho^0$ using the uncorrected and bias-corrected Wald, LR, and LM test, and nominal size 5%. The bias-corrected tests are indicated by an asterisk superscript. The size of the different tests is reported, based on 10,000 simulation runs.

	size			size		
	WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$						
$N = 100, T = 20, M = 4$	0.219	0.214	0.192	0.066	0.062	0.056
$N = 400, T = 80, M = 6$	0.199	0.198	0.195	0.055	0.054	0.054
$N = 400, T = 20, M = 4$	0.560	0.556	0.532	0.089	0.088	0.076
$N = 1600, T = 80, M = 6$	0.593	0.591	0.586	0.056	0.055	0.055
$\rho^0 = 0.6$						
$N = 100, T = 20, M = 4$	0.326	0.311	0.272	0.098	0.091	0.077
$N = 400, T = 80, M = 6$	0.260	0.255	0.248	0.056	0.053	0.057
$N = 400, T = 20, M = 4$	0.591	0.582	0.552	0.174	0.167	0.136
$N = 1600, T = 80, M = 6$	0.666	0.663	0.656	0.060	0.058	0.059

a larger size distortion, whereas for fixed N, an increase in T results in a smaller size distortion (both for the noncorrected and for the bias-corrected tests).

In Table 7 and 8, we present the power and the size-corrected power when testing the left-sided alternative $H_a^{\text{left}} : \rho = \rho^0 - (NT)^{-1/2}$ and the right-sided alternative $H_a^{\text{right}} : \rho = \rho^0 + (NT)^{-1/2}$. The model specifications are the same as for the size results in Table 4. Because both the FLS estimator and the bias-corrected FLS estimator for ρ have a negative bias, one finds the power for the left-sided alternative to be much smaller than the power for the right-sided alternative. For the uncorrected tests, this effect can be extreme and the size-corrected power of these tests for the left-sided alternative is below 2% in all cases and does not improve as N and T become large, holding N/T fixed. By contrast, the power for the bias-corrected tests becomes more symmetric as N and T become large, and the size-corrected power for the left-sided alternative is much larger than for the uncorrected tests, whereas the size-corrected power for the right-sided alternative is about the same.

8. CONCLUSIONS

This paper studies the least squares estimator for dynamic linear panel regression models with interactive fixed effects. We provide conditions under which the estimator is consistent, allowing for predetermined regressors and for a general combination of "low-rank" and "high-rank" regressors. We then show how a quadratic approximation of the profile objective function $L_{NT}(\beta)$ can be used

TABLE 7. As Table 6, but we report the power for testing the alternatives $H_a^{\text{left}}: \rho = \rho^0 - (NT)^{-1/2}$ and $H_a^{\text{right}}: \rho = \rho^0 + (NT)^{-1/2}$. The bias-corrected tests are indicated by an asterisk superscript.

		power			power		
		WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$							
$N = 100, T = 20, M = 4$	H_a^{left}	0.094	0.089	0.076	0.128	0.123	0.121
	H_a^{right}	0.526	0.515	0.487	0.235	0.227	0.206
$N = 400, T = 80, M = 6$	H_a^{left}	0.066	0.064	0.063	0.154	0.151	0.153
	H_a^{right}	0.549	0.545	0.540	0.194	0.191	0.190
$N = 400, T = 20, M = 4$	H_a^{left}	0.306	0.305	0.284	0.100	0.097	0.096
	H_a^{right}	0.791	0.787	0.769	0.309	0.305	0.279
$N = 1600, T = 80, M = 6$	H_a^{left}	0.254	0.253	0.248	0.128	0.127	0.129
	H_a^{right}	0.871	0.869	0.866	0.225	0.224	0.224
$\rho^0 = 0.6$							
$N = 100, T = 20, M = 4$	H_a^{left}	0.192	0.180	0.147	0.184	0.171	0.171
	H_a^{right}	0.619	0.605	0.563	0.335	0.318	0.294
$N = 400, T = 80, M = 6$	H_a^{left}	0.081	0.079	0.076	0.184	0.195	0.200
	H_a^{right}	0.680	0.675	0.668	0.335	0.262	0.267
$N = 400, T = 20, M = 4$	H_a^{left}	0.421	0.412	0.378	0.184	0.160	0.150
	H_a^{right}	0.792	0.787	0.765	0.335	0.426	0.399
$N = 1600, T = 80, M = 6$	H_a^{left}	0.318	0.314	0.307	0.200	0.169	0.172
	H_a^{right}	0.912	0.911	0.908	0.268	0.316	0.320

to derive the first-order asymptotic theory of the LS estimator of β under the alternative asymptotic $N, T \to \infty$. We find the asymptotic distribution of the LS estimator can be asymptotically biased (i) because of weak exogeneity of the regressors and (ii) because of heteroscedasticity (and correlation) of the idiosyncratic errors e_{it}. Consistent estimators for the asymptotic covariance matrix and for the asymptotic bias of the LS estimator are provided, and thus a bias-corrected LS estimator is given. We furthermore study the asymptotic distributions of the Wald, LR, and LM test statistics for testing a general linear hypothesis on β. The uncorrected test statistics are not asymptotically chi-square because of the asymptotic bias of the score and of the LS estimator, but bias-corrected test statistics that are asymptotically chi-square distributed can be constructed. We also discussed a possible extension of the estimation procedure to the case of endogeneous regressors. The findings of our Monte Carlo simulations show our asymptotic results on the distribution of the (bias-corrected) LS estimator and of the (bias-corrected) test statistics provide a good approximation of their finite sample properties.

TABLE 8. As Table 7, but we report the size-corrected power.

		size-corrected power			size-corrected power		
		WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$							
$N = 100, T = 20, M = 4$	H_a^{left}	0.010	0.011	0.010	0.105	0.104	0.112
	H_a^{right}	0.211	0.208	0.206	0.199	0.197	0.193
$N = 400, T = 80, M = 6$	H_a^{left}	0.008	0.008	0.008	0.143	0.143	0.145
	H_a^{right}	0.236	0.237	0.235	0.181	0.182	0.181
$N = 400, T = 20, M = 4$	H_a^{left}	0.008	0.008	0.009	0.055	0.052	0.062
	H_a^{right}	0.187	0.185	0.181	0.210	0.208	0.208
$N = 1600, T = 80, M = 6$	H_a^{left}	0.005	0.005	0.005	0.119	0.119	0.120
	H_a^{right}	0.226	0.227	0.225	0.213	0.213	0.212
$\rho^0 = 0.6$							
$N = 100, T = 20, M = 4$	H_a^{left}	0.014	0.014	0.016	0.114	0.115	0.127
	H_a^{right}	0.196	0.193	0.196	0.233	0.234	0.231
$N = 400, T = 80, M = 6$	H_a^{left}	0.005	0.005	0.005	0.114	0.187	0.184
	H_a^{right}	0.288	0.288	0.288	0.233	0.252	0.247
$N = 400, T = 20, M = 4$	H_a^{left}	0.013	0.016	0.015	0.114	0.039	0.051
	H_a^{right}	0.128	0.127	0.126	0.233	0.201	0.209
$N = 1600, T = 80, M = 6$	H_a^{left}	0.005	0.005	0.005	0.185	0.153	0.154
	H_a^{right}	0.236	0.236	0.238	0.248	0.291	0.291

Although the bias-corrected LS estimator has a nonzero bias in finite samples, this bias is much smaller than that of the LS estimator. Analogously, the size distortions and power asymmetries of the bias-corrected Wald, LR, and LM test are much smaller than for the nonbias-corrected versions.

NOTES

1. See, e.g., Chamberlain & Rothschild (1983), Ross (1976), and Fama & French (1993) for asset pricing; Stock & Watson (2002) and Bai & Ng (2006) for forecasting; Bernanke, Boivin, & Eliasz (2005) for empirical macro; and Holtz-Eakin, Newey, & Rosen (1988) for empirical labor economics.

2. The theory of the CCE estimator was further developed in, e.g., Harding & Lamarche (2009; 2011), Kapetanios, Pesaran, & Yamagata (2011), Pesaran & Tosetti (2011), Chudik, Pesaran, & Tosetti (2011), and Chudik & Pesaran (2015).

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", because LS estimation is equivalent to maximizing a conditional Gaussian likelihood function.

4. Hahn & Kuersteiner (2002) introduced the alternative asymptotics to characterize the asymptotic bias due to incidental parameter problems in fixed effect dynamic panel data models. See also Arellano & Hahn (2007) and Moon, Perron, & Phillips (2014) and references therein.

5. The "likelihood ratio" and the score used in the tests are based on the LS objective function, which can be interpreted as the (misspecified) conditional Gaussian likelihood function.

6. Another type of widely studied tests in the interactive fixed effect panel literature are panel unit root tests, e.g., Bai & Ng (2004), Moon & Perron (2004), and Phillips & Sul (2003).

7. In Moon & Weidner (2015) we do not consider low-rank regressors or testing problems, and we impose more restrictive assumptions on the error term of the model implying that some leading bias terms of the LS estimator are not present.

8. Lee, Moon, & Weidner (2012) also apply the MSW estimation method to estimate a simple dynamic panel regression with interactive fixed effect and classical measurement errors.

9. To remove this restriction, one could estimate R consistently in the presence of the regressors. In the literature so far, however, consistent estimation procedures for R are established mostly in pure factor models (e.g., Bai & Ng (2002), Onatski (2010) & Harding (2007)). Alternatively, one could rely on Moon & Weidner (2015) who consider a regression model with interactive fixed effects when only an upper bound on the number of factors is known — but extending those results to the more general setup considered here is mathematically challenging.

10. If we have low-rank regressors with rank larger than one, then we write $X_l = w_l v_l'$, where w_l is an $N \times \text{rank}(X_l)$ matrix and v_l is a $T \times \text{rank}(X_l)$ matrix, and we define $w = (w_1, \ldots, w_{K_1})$ as a $N \times \sum_{l=1}^{L} \text{rank}(X_l)$ matrix, and $v = (v_1, \ldots, v_{K_1})$ ae a $T \times \sum_{l=1}^{L} \text{rank}(X_l)$ matrix. All our results are then unchanged, as long as $\text{rank}(X_l)$ is a finite constant for all $l = 1, \ldots, K_1$, and we replace $2R + K_1$ by $2R + \text{rank}\,(w)$ in Assumption ID(v) and Assumption 4(ii)(a).

11. Note that $\text{rank}(\lambda^0) = R$ if R factors are present. Our identification results are consistent with the possibility that $\text{rank}(\lambda^0) < R$, i.e., that R only represents an upper bound on the number of factors, but later we assume $\text{rank}(\lambda^0) = R$ to show consistency.

12. We could write $X_k^{(N,T)}$, $e^{(N,T)}$, $\lambda^{(N,T)}$, and $f^{(N,T)}$, because all these matrices, and even their dimensions, are functions on N and T, but we suppress this dependence throughout the paper.

13. Here and in the following, we write $\sigma(A)$ for the sigma algebra generated by the (collection of) random variable(s) A, and we write $\mathcal{A} \vee \mathcal{B}$ for the sigma algebra generated by the unions of all elements in the sigma algebra $\mathcal{A}$ and $\mathcal{B}$, so that in the conditional expectation in AssumptionS 5(ii), we condition jointly on $\mathcal{C}$ and $\{(X_{is}, e_{i,s-1}), s \le t\}$.

14. Assumption 2 and 3* are implied by Assumption 5 and therefore need not be explicitly assumed here.

15. Alternatively, one could use $\widehat{B}(\widetilde{\beta})$ and $\widehat{W}(\widetilde{\beta})$ as estimates for B and W, and would obtain the same limiting distribution of LR_{NT}^* under the null hypothesis H_0. These alternative estimators are not consistent if H_0 is false, i.e. the power-properties of the test would be different. The question of which specification should be preferred is left for future research.

16. The proof of the statement is given in the supplementary material as part of the proof of Theorem 5.2.

17. Chernazhukov & Hansen (2005) also used a similar method for estimating endogenous quantile regression models.

18. Here we can either use $\mathbb{B} = (-1, 1)$, or $\mathbb{B} = \mathbb{R}$. In the present model, we only have high-rank regressors; i.e., the parameter space need not be bounded to show consistency.

19. To be precise, we have $\|\lambda^0 f^{0\prime}\|/(\sqrt{2NT}\sigma_f) \to_p 1$, and $\|e\|/(\sqrt{N} + \sqrt{T}) \to_p 1$.

REFERENCES

Ahn, S.C., Y.H. Lee, & P. Schmidt (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.

Andrews, D.W.K. (2001) Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* 69(3), 683–734.

Arellano, M. & J. Hahn (2007) Understanding bias in nonlinear panel models: Some recent developments. *Econometric Society Monographs* 43, 381.

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

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

Bai, J. & S. Ng (2004) A panic attack on unit roots and cointegration. *Econometrica* 72(4), 1127–1177.

Bai, J. & S. Ng (2006) Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74(4), 1133–1150.

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

Bernanke, B.S., J. Boivin, & P. Eliasz (2005) Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics* 120(1), 387–422.

Berry, S., J. Levinsohn, & A. Pakes (1995) Automobile prices in market equilibrium. *Econometrica* pp. 841–890.

Chamberlain, G. & M. Rothschild (1983) Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51(5), 1281–1304.

Chernozhukov, V. & C. Hansen (2005) An iv model of quantile treatment effects. *Econometrica* 73(1), 245–261.

Chudik, A. & M.H. Pesaran (2015) Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*.

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

Dhaene, G. & K. Jochmans (2015) Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*.

Fama, E.F. & K.R. French (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33(1), 3–56.

Fernández-Val, I. & M. Weidner (2013) Individual and time effects in nonlinear panel data models with large N, T. *CeMMAP working paper series*.

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

Gobillon, L. & T. Magnac (2013) Regional policy evaluation: Interactive fixed effects and synthetic controls. *IZA Discussion Paper No. 7493*.

Hahn, J. & G. Kuersteiner (2002) Asymptotically unbiased inference for a dynamic panel model with fixed effects when both "n" and "T" are large. *Econometrica* 70(4), 1639–1657.

Hahn, J. & G. Kuersteiner (2011) Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory* 27(06), 1152–1191.

Hahn, J. & W. Newey (2004) Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72(4), 1295–1319.

Harding, M. (2007) Structural estimation of high-dimensional factor models. *unpublished manuscript*.

Harding, M. & C. Lamarche (2009) A quantile regression approach for estimating panel data models using instrumental variables. *Economics Letters* 104(3), 133–135.

Harding, M. & C. Lamarche (2011) Least squares estimation of a panel data model with multifactor error structure and endogenous covariates. *Economics Letters* 111(3), 197–199.

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

Kapetanios, G., M.H. Pesaran, & T. Yamagata (2011) Panels with non-stationary multifactor error structures. *Journal of Econometrics* 160(2), 326–348.

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

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

Lee, N., H.R. Moon, & M. Weidner (2012) Analysis of interactive fixed effects dynamic linear panel regression with measurement error. *Economics Letters* 117(1), 239–242.

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

Moon, H.R. & B. Perron (2004) Testing for a unit root in panels with dynamic factors. *Journal of Econometrics* 122(1), 81–126.

Moon, H.R. & M. Weidner (2015) Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* 83(4), 1543–1579.

Moon, R., B. Perron, & P.C. Phillips (2014) Incidental parameters and dynamic panel modeling. *forthcoming in The Oxford Handbook of Panel Data, Chapter 4*.

Nickell, S. (1981) Biases in dynamic models with fixed effects. *Econometrica* 49(6), 1417–1426.

Onatski, A. (2010) Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* 92(4), 1004–1016.

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

Pesaran, M.H. & E. Tosetti (2011) Large panels with common factors and spatial correlation. *Journal of Econometrics* 161(2), 182–202.

Phillips, P.C.B. & D. Sul (2003) Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* 6(1), 217–259.

Ross, S.A. (1976) The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3), 341–360.

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

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

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

APPENDIX

A. Proof of Consistency (Theorem 3.1)

The following theorem is useful for the consistency proof and beyond.

LEMMA A.1. *Let N, T, R, R_1, and R_2 be positive integers such that $R \leq N$, $R \leq T$, and $R = R_1 + R_2$. Let Z be an $N \times T$ matrix, λ be an $N \times R$, f be a $T \times R$ matrix, $\widetilde{\lambda}$ be an $N \times R_1$ matrix, and $\widetilde{f}$ be a $T \times R_2$ matrix. Then the following six expressions (that are functions of Z only) are equivalent:*

$$\min_{f,\lambda} \operatorname{Tr}\left[\left(Z-\lambda f'\right)\left(Z'-f\lambda'\right)\right] = \min_{f} \operatorname{Tr}\left(Z\,M_f\,Z'\right) = \min_{\lambda} \operatorname{Tr}\left(Z'\,M_\lambda\,Z\right)$$

$$= \min_{\tilde{\lambda},\tilde{f}} \operatorname{Tr}\left(M_{\tilde{\lambda}}\,Z\,M_{\tilde{f}}\,Z'\right) = \sum_{i=R+1}^{T} \mu_i\left(Z'Z\right) = \sum_{i=R+1}^{N} \mu_i\left(ZZ'\right).$$

In the above minimization problems, we do not have to restrict the matrices λ, f, $\widetilde{\lambda}$, and $\widetilde{f}$ to be of full rank. If, for example, λ is not of full rank, the generalized inverse $(\lambda'\lambda)^{\dagger}$ is still well defined, and the projector M_λ still satisfies $M_\lambda \lambda = 0$ and $\operatorname{rank}(M_\lambda) = N - \operatorname{rank}(\lambda)$. If $\operatorname{rank}(Z) \geq R$, the optimal λ, f, $\widetilde{\lambda}$, and $\widetilde{f}$ always have full rank.

Lemma A.1 shows the equivalence of the three different versions of the profile objective function in Equation (4). It also considers minimization of $\operatorname{Tr}(M_{\widetilde{\lambda}}\,Z\,M_{\widetilde{f}}\,Z')$ over $\widetilde{\lambda}$ and $\widetilde{f}$, which will be used in the consistency proof below. The proof of the theorem is given in the supplementary material. The following lemma is due to Bai (2009).

LEMMA A.2. *Under the assumptions of Theorem 3.1 we have*

$$\sup_f \left| \frac{\mathrm{Tr}\left(X_k M_f e'\right)}{NT} \right| = o_p(1), \quad \sup_f \left| \frac{\mathrm{Tr}\left(\lambda^0 f^{0\prime} M_f e'\right)}{NT} \right| = o_p(1), \quad \sup_f \left| \frac{\mathrm{Tr}\left(e P_f e'\right)}{NT} \right| = o_p(1),$$

where the parameters f are $T \times R$ matrices with $\mathrm{rank}(f) = R$.

Proof. By Assumption 2, we know the first equation in Lemma A.2 is satisfied when replacing M_f by the identity matrix. So we are left to show $\max_f \left| \frac{1}{NT} \mathrm{Tr}(\Xi e') \right| = o_p(1)$, where Ξ is either $X_k P_f$, $\lambda^0 f^{0\prime} M_f$, or $e P_f$. In all three cases, we have $\|\Xi\|/\sqrt{NT} = \mathcal{O}_p(1)$ by Assumption 1, 3, and 4, respectively, and we have $\mathrm{rank}(\Xi) \leq R$. We therefore find

$$\sup_f \left| \frac{1}{NT} \mathrm{Tr}(\Xi P_f e') \right| \leq R \frac{\|e\|}{\sqrt{NT}} \frac{\|\Xi\|}{\sqrt{NT}} = o_p(1).$$

Here, we used $|\mathrm{Tr}(C)| \leq \|C\| \mathrm{rank}(C)$, which holds for all square matrices C; see the supplementary material. ■

Proof of Theorem 3.1. For the second version of the profile objective function in Equation (4), we write $L_{NT}(\beta) = \min_f S_{NT}(\beta, f)$, where

$$S_{NT}(\beta, f) = \frac{1}{NT} \mathrm{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k \right) X_k + e \right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k \right) X_k + e \right)' \right].$$

We have $S_{NT}(\beta^0, f^0) = \frac{1}{NT} \mathrm{Tr}\left(e M_{f^0} e' \right)$. Using Lemma (A.2), we find

$$\begin{aligned} S_{NT}(\beta, f) &= S_{NT}(\beta^0, f^0) + \widetilde{S}_{NT}(\beta, f) \\ &\quad + \frac{2}{NT} \mathrm{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k \right) X_k \right) M_f e' \right] + \frac{1}{NT} \mathrm{Tr}\left(e \left(P_{f^0} - P_f \right) e' \right) \\ &= S_{NT}(\beta^0, f^0) + \widetilde{S}_{NT}(\beta, f) + o_p(\|\beta - \beta^0\|) + o_p(1), \end{aligned} \tag{A.1}$$

where we defined

$$\widetilde{S}_{NT}(\beta, f) = \frac{1}{NT} \mathrm{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k \right) X_k \right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k \right) X_k \right)' \right].$$

Up to this point, the consistency proof is almost equivalent to the one given in Bai (2009), but the remainder of the proof differs from Bai, because we allow for more general low-rank regressors, and because we allow for high-rank and low-rank regressors simultaneously. We split $\widetilde{S}_{NT}(\beta, f) = \widetilde{S}_{NT}^{(1)}(\beta, f) + \widetilde{S}_{NT}^{(2)}(\beta, f)$, where

$$\widetilde{S}_{NT}^{(1)}(\beta, f) = \frac{1}{NT} \mathrm{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)' M_{(\lambda^0, w)} \right]$$

$$= \frac{1}{NT} \operatorname{Tr}\left[\left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right) M_f \left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right)' M_{(\lambda^0, w)} \right],$$

$$\widetilde{S}_{NT}^{(2)}(\beta, f)$$

$$= \frac{1}{NT} \operatorname{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)' P_{(\lambda^0, w)} \right],$$

and (λ^0, w) is the $N \times (R + K_1)$ matrix that is composed out of λ^0 and the $N \times K_1$ matrix w defined in Assumption 4. For $\widetilde{S}_{NT}^{(1)}(\beta, f)$, we can apply Lemma A.1 with $\widetilde{f} = f$ and $\widetilde{\lambda} = (\lambda^0, w)$ (the R in the theorem is now $2R + K_1$) to find

$$\begin{aligned} \widetilde{S}_{NT}^{(1)}(\beta, f) &\geq \frac{1}{NT} \\ &\quad \times \sum_{i=2R+K_1+1}^{N} \mu_i \left[\left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right) \left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right)' \right] \\ &\geq b \left\| \beta^{\text{high}} - \beta^{0,\text{high}} \right\|^2, \qquad \text{wpa1}, \end{aligned} \tag{A.2}$$

where in the last step, we used the existence of a constant $b > 0$ guaranteed by Assumption 4(ii)(a), and we introduced $\beta^{\text{high}} = (\beta_{K_1+1}, \ldots, \beta_K)'$, which refers to the $K_2 \times 1$ parameter vector corresponding to the high-rank regressors. Similarly, we define $\beta^{\text{low}} = (\beta_1, \ldots, \beta_{K_1})'$ for the $K_1 \times 1$ parameter vector of low-rank regressors.

Using $P_{(\lambda^0, w)} = P_{(\lambda^0, w)} P_{(\lambda^0, w)}$ and the cyclicality of the trace, we see $\widetilde{S}_{NT}^{(2)}(\beta, f)$ can be written as the trace of a positive definite matrix, and therefore $\widetilde{S}_{NT}^{(2)}(\beta, f) \geq 0$. Note also that we can choose $\beta = \beta^0$ and $f = f^0$ in the minimization problem over $S_{NT}(\beta, f)$; that is, the optimal $\beta = \widehat{\beta}$ and $f = \widehat{f}$ must satisfy $S_{NT}(\widehat{\beta}, \widehat{f}) \leq S_{NT}(\beta^0, f^0)$. Using this result, Equation (A.1), $\widetilde{S}_{NT}^{(2)}(\beta, f) \geq 0$, and the bound in (A.2), we find

$$0 \geq b \left\| \widehat{\beta}^{\text{high}} - \beta^{0,\text{high}} \right\|^2 + o_p\left(\left\| \widehat{\beta}^{\text{high}} - \beta^{0,\text{high}} \right\| \right) + o_p\left(\left\| \widehat{\beta}^{\text{low}} - \beta^{0,\text{low}} \right\| \right) + o_p(1).$$

Because we assume $\widehat{\beta}^{\text{low}}$ is bounded, the last equation implies $\left\| \widehat{\beta}^{\text{high}} - \beta^{0,\text{high}} \right\| = o_p(1)$; that is, $\widehat{\beta}^{\text{high}}$ is consistent. What is left to show is that $\widehat{\beta}^{\text{low}}$ is consistent, too. In the supplementary material, we show Assumption 4(ii)(b) guarantees that finite positive constants a_0, a_1, a_2, a_3, and a_4 exist such that

$$\begin{aligned} \widetilde{S}_{NT}^{(2)}(\beta, f) \geq & \frac{a_0 \left\| \beta^{\text{low}} - \beta^{0,\text{low}} \right\|^2}{\left\| \beta^{\text{low}} - \beta^{0,\text{low}} \right\|^2 + a_1 \left\| \beta^{\text{low}} - \beta^{0,\text{low}} \right\| + a_2} \\ & - a_3 \left\| \beta^{\text{high}} - \beta^{0,\text{high}} \right\| - a_4 \left\| \beta^{\text{high}} - \beta^{0,\text{high}} \right\| \left\| \beta^{\text{low}} - \beta^{0,\text{low}} \right\|, \quad \text{wpa1}. \end{aligned}$$

Using consistency of $\widehat{\beta}^{\text{high}}$ and again boundedness of β^{low}, the previous inequality implies $a > 0$ exists such that $\widetilde{S}_{NT}^{(2)}(\widehat{\beta}, f) \geq a \left\| \widehat{\beta}^{\text{low}} - \beta^{0,\text{low}} \right\|^2 + o_p(1)$. With the same argument as for $\widehat{\beta}^{\text{high}}$, we therefore find $\left\| \widehat{\beta}^{\text{low}} - \beta^{0,\text{low}} \right\| = o_p(1)$; that is, $\widehat{\beta}^{\text{low}}$ is consistent. ■

B. Proof of Limiting Distribution (Theorem 4.3)

Theorem 4.1 is from Moon and Weidner (2015), and the proof can be found there. Note Assumption 4(i) implies $\|X_k\| = \mathcal{O}_p(\sqrt{NT})$, which we assume in Moon and Weidner (2015). There, we also assume that $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)}) = \mathcal{O}_p(\sqrt{N})$, whereas in the current paper we assume $\|e\| = o_p(\|N^{2/3}\|)$. It is, however, straightforward to verify that the proof of Theorem 4.1 is also valid under this weaker assumption.

Moon and Weidner (2015) also includes the proof of Corollary 4.2. The proof requires consistency of $\widehat{\beta}$, which in the current paper is derived under weaker assumptions than in Moon and Weidner (2015), where no low-rank regressors are considered. Corollary 4.2 is therefore stated under weaker assumptions here, but the proof is unchanged. In the supplementary material, we show the assumptions of Corollary 4.2 already guarantee W_{NT} does not become singular as $N, T \to \infty$.

For each $k = 1, \ldots, K$, we define the $N \times T$ matrices $\overline{X}_k$, $\widetilde{X}_k$, and $\mathfrak{X}_k$ as follows:

$$\overline{X}_k = \mathbb{E}\left(X_k \middle| \mathcal{C}\right), \qquad \widetilde{X}_k = X_k - \mathbb{E}\left(X_k \middle| \mathcal{C}\right), \qquad \mathfrak{X}_k = M_{\lambda^0}\, \overline{X}_k\, M_{f^0} + \widetilde{X}_k.$$

Note the difference between $\mathfrak{X}_k$ and $\mathcal{X}_k = M_{\lambda^0} X_k M_{f^0}$, which was defined in Assumption 6. In particular, conditional on $\mathcal{C}$, the elements $\mathfrak{X}_{k,it}$ of $\mathfrak{X}_k$ are contemporaneously uncorrelated with the error term e_{it}, although the same is not true for $\mathcal{X}_k$.

To present the proof of Theorem 4.3, it is convenient to first state two technical lemmas.

LEMMA B.1. *Under the assumptions of Theorem 4.3, we have*

$$(a) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left(P_{f^0}\, e'\, P_{\lambda^0}\, \widetilde{X}_k\right) = o_p(1),$$

$$(b) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left(P_{\lambda^0}\, e\, \widetilde{X}_k'\right) = o_p(1),$$

$$(c) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left\{P_{f^0}\left[e'\, \widetilde{X}_k - \mathbb{E}\left(e'\, \widetilde{X}_k \middle| \mathcal{C}\right)\right]\right\} = o_p(1),$$

$$(d) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left(e P_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0 \left(f^{0\prime} f^0\right)^{-1} \left(\lambda^{0\prime} \lambda^0\right)^{-1} \lambda^{0\prime}\right) = o_p(1),$$

$$(e) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left(e'\, P_{\lambda^0}\, e\, M_{f^0}\, X_k'\, \lambda^0 \left(\lambda^{0\prime} \lambda^0\right)^{-1} \left(f^{0\prime} f^0\right)^{-1} f^{0\prime}\right) = o_p(1),$$

$$(f) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0 \left(\lambda^{0\prime} \lambda^0\right)^{-1} \left(f^{0\prime} f^0\right)^{-1} f^{0\prime}\right) = o_p(1),$$

$$(g) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left\{\left[ee' - \mathbb{E}\left(ee' \middle| \mathcal{C}\right)\right] M_{\lambda^0}\, X_k\, f^0 \left(f^{0\prime} f^0\right)^{-1} \left(\lambda^{0\prime} \lambda^0\right)^{-1} \lambda^{0\prime}\right\} = o_p(1),$$

$$(h) \qquad \frac{1}{\sqrt{NT}} \operatorname{Tr}\left\{\left[e'e - \mathbb{E}\left(e'e \middle| \mathcal{C}\right)\right] M_{f^0}\, X_k'\, \lambda^0 \left(\lambda^{0\prime} \lambda^0\right)^{-1} \left(f^{0\prime} f^0\right)^{-1} f^{0\prime}\right\} = o_p(1),$$

$$(i) \qquad \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[e_{it}^2\, \mathfrak{X}_{it}\, \mathfrak{X}_{it}' - \mathbb{E}\left(e_{it}^2\, \mathfrak{X}_{it}\, \mathfrak{X}_{it}' \middle| \mathcal{C}\right)\right] = o_p(1),$$

$$(j) \qquad \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2 \left(\mathfrak{X}_{it}\, \mathfrak{X}_{it}' - \mathcal{X}_{it}\, \mathcal{X}_{it}'\right) = o_p(1).$$

LEMMA B.2. *Under the assumptions of Theorem 4.3, we have*

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\mathfrak{X}_{it} \underset{d}{\to} \mathcal{N}(0,\Omega).$$

The proofs of Lemma B.1 and Lemma B.2 are provided in the supplementary material. We briefly want to discuss why the asymptotic variance-covariance matrix in Lemma B.2 turns out to be Ω. Note that because $e_{it}\mathfrak{X}_{it}$ is mean zero and uncorrelated across both i and t, conditional on $\mathcal{C}$, we have

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\mathfrak{X}_{it}\,\middle|\,\mathcal{C}\right) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left(e_{it}^2\,\mathfrak{X}_{it}\,\mathfrak{X}_{it}'\,|\,\mathcal{C}\right)\\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2\,\mathfrak{X}_{it}\,\mathfrak{X}_{it}' + o_p(1)\\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2\,\mathcal{X}_{it}\,\mathcal{X}_{it}' + o_p(1)\\
&= \Omega + o_p(1), \qquad \textbf{(B.3)}
\end{aligned}$$

where we also used part (i) of Lemma B.1 for the second equality and part (j) of Lemma B.1 for the third equality, and the definition of Ω in Assumptions 6 in the last step.

Using those lemmas, we can now prove the theorem on the limiting distribution of $\widehat{\beta}$ in the main text.

Proof of Theorem 4.3. Assumption 5 implies $\|e\| = \mathcal{O}_p(N^{1/2})$ as N and T grow at the same rate, as discussed in Section S.2 of the supplementary material; that is, Assumption 3* is satisfied. We can therefore apply Corollary 4.2 to calculate the limiting distribution of $\widehat{\beta}$. Note that $M_{\lambda^0} X_k M_{f^0} = \mathfrak{X}_k - \widetilde{X}_k\,P_{f^0} - P_{\lambda^0}\,\widetilde{X}_k + P_{\lambda^0}\,\widetilde{X}_k\,P_{f^0}$. Using Lemmas B.1 and B.2 and Assumption 6, we find

$$\begin{aligned}
\frac{1}{\sqrt{NT}}C^{(1)}\left(\lambda^0,\,f^0,\,X_k,\,e\right) &= \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(e'\,M_{\lambda^0}\,X_k\,M_{f^0}\right)\\
&= \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\operatorname{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,\widetilde{X}_k\,|\,\mathcal{C}\right)\right]\\
&\quad - \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(e'\,P_{\lambda^0}\,\widetilde{X}_k\right) + \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(P_{f^0}\,e'\,P_{\lambda^0}\,\widetilde{X}_k\right)\\
&\quad - \frac{1}{\sqrt{NT}}\operatorname{Tr}\left\{P_{f^0}\left[e'\,\widetilde{X}_k - \mathbb{E}\left(e'\,\widetilde{X}_k\,|\,\mathcal{C}\right)\right]\right\}\\
&= \frac{1}{\sqrt{NT}}\operatorname{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\operatorname{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,X_k\,|\,\mathcal{C}\right)\right] + o_p(1).\\
&\underset{d}{\to} \mathcal{N}\left(-\kappa B_1,\,\Omega\right),
\end{aligned}$$

where we also used that $\mathbb{E}\left(e'\,\widetilde{X}_k\,|\,\mathcal{C}\right)=\mathbb{E}\left(e'\,X_k\,|\,\mathcal{C}\right)$. Using Lemma B.1, we also find

$$\begin{aligned}
&\frac{1}{\sqrt{NT}}C^{(2)}\left(\lambda^0,\,f^0,\,X_k,\,e\right)\\
&\quad=-\frac{1}{\sqrt{NT}}\Big[\mathrm{Tr}\left(eM_{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]\\
&\quad=\frac{1}{\sqrt{NT}}\mathrm{Tr}\left(eP_{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-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{\left[ee'-\mathbb{E}\left(ee'|\mathcal{C}\right)\right]M_{\lambda^0}\,X_k\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right\}\\
&\qquad-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left[\mathbb{E}\left(ee'|\mathcal{C}\right)M_{\lambda^0}\,X_k\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right]\\
&\qquad+\frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'P_{\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-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{\left[e'e-\mathbb{E}\left(e'e|\mathcal{C}\right)\right]M_{f^0}\,X_k'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\right\}\\
&\qquad-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left[\mathbb{E}\left(e'e|\mathcal{C}\right)M_{f^0}\,X_k'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\right]\\
&\qquad+\frac{1}{\sqrt{NT}}\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)\\
&\quad=-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left[\mathbb{E}\left(ee'|\mathcal{C}\right)M_{\lambda^0}\,X_k\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right]\\
&\qquad-\frac{1}{\sqrt{NT}}\mathrm{Tr}\left[\mathbb{E}\left(e'e|\mathcal{C}\right)M_{f^0}\,X_k'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\right]+o_p(1),\\
&\quad=-\kappa^{-1}B_2-\kappa B_3+o_p(1).
\end{aligned}$$

Combining these results, we obtain

$$\begin{aligned}
\sqrt{NT}\left(\widehat{\beta}-\beta^0\right)&=W_{NT}^{-1}\frac{1}{\sqrt{NT}}C_{NT}\\
&\underset{d}{\to}\mathcal{N}\left(-W^{-1}\left(\kappa B_1+\kappa^{-1}B_2+\kappa B_3\right),\;W^{-1}\,\Omega\,W^{-1}\right),
\end{aligned}$$

which is what we wanted to show. ■