Dynamic Linear Panel Regression Models with Interactive Fixed Effects *

Hyungsik Roger Moon‡ Martin Weidner‡

July 15, 2010

Abstract

This paper analyzes linear panel regression models with interactive fixed effects and predetermined regressors, e.g. lagged-dependent variables. We work out the first order asymptotic theory of the Gaussian quasi maximum likelihood estimator (QMLE) of the regression coefficients in the limit where both the cross sectional dimension and the number of time periods become large. We find that there are two sources of asymptotic bias of the QMLE: bias due to correlation or heteroscedasticity of the idiosyncratic error term, and bias due to predetermined (as opposed to strictly exogenous) regressors. For idiosyncratic errors that are independent across time and cross section we provide an estimator for the bias and a bias corrected QMLE. We also provide bias corrected versions of the three classical test statistics (Wald, LR and LM test) and show that their asymptotic distribution is a χ^2-distribution. Monte Carlo simulations show that the bias correction of the QMLE and of the test statistics also work well for finite sample sizes.

1 Introduction

This paper studies a panel regression model where 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 fixed effects and time-specific effects as special cases, but is significantly more flexible since it allows the factors f_t to affect each individual with a different loading λ_i.

In the absence of regressors, the model becomes an approximate factor model, as introduced by Chamberlain and Rothschild (1983) to study asset returns. Multifactor models in asset pricing are motivated by no-arbitrage arguments (Ross, 1976), and can be successful in explaining cross-sectional variations of stock returns, *e.g.* Fama and French (1993). Additional regressors in these

*This paper is based on an unpublished manuscript of the authors which was circulated under the title "Likelihood Expansion for Panel Regression Models with Factors". 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 Honor 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. Moon is grateful for the financial support from the NSF via grant SES 0920903 and the faculty development award from USC, and Weidner is grateful for travel funding from the Institute of Economic Policy and Research, USC.

‡Department of Economics, University of Southern California, 3620 South Vermont Avenue, Kaprielian Hall 300, Los Angeles, CA 90089-0253. Email: moonr@usc.edu, mweidner@usc.edu.

models are introduced to account for firm-specific characteristics, see *e.g* Daniel and Titman (1997).

In macroeconomics, factor models are used to account for international or national shocks that simultaneously affect multiple countries or multiple country specific variables. The diffusion index forecast model of Stock and Watson (2002) (see also Bai and Ng (2006)), and the factor augmented VAR of Bernanke, Boivin and Eliasz (2005) both describe the dynamics of the variables of interest by a combination of unobserved factors and observed covariates. The interactive fixed effect model examined in this paper can be viewed as a limited information version of these models, since no further assumptions on the dynamics of the covariates or of the factors are made. In this context it is crucial that we allow for predetermined regressors, like lagged dependent variables.

Lagged dependent variables are also important for microeconomic applications of the model. For example, Holtz-Eakin, Newey, and Rosen (1988) use the interactive fixed effect specification to study the relationship between wages and hours worked. There λ_i can describe the unobserved earnings abilities of individuals, while f_t can be interpreted as changes in local working conditions, or the macroeconomic state of the economy.

In the present paper we study the (Gaussian) quasi maximum likelihood estimator (QMLE), which jointly minimizes the fixed effect parameters, λ_i and f_t, and the sum of squared residuals of the linear panel model over the regression coefficients. Our analysis uses the alternative asymptotic where both the number of cross-sectional units N and the number of time periods T becomes large, which was shown to be a convenient tool to characterize the asymptotic bias due to incidental parameter problems, see *e.g.* Hahn and Kuersteiner (2002; 2004), Alvarez and Arellano (2003), Hahn and Newey (2004), and Hahn and Moon (2006).

The QMLE for linear panel regression models in the asymptotics $N, T \to \infty$ was previously analyzed by Bai (2009). Our paper extents Bai (2009)'s work in many respects. While Bai (2009) assumed regressors to be strictly exogenous, we allow for predetermined regressors, *e.g.* lagged-dependent variables, and find an additional incidental parameter bias that results from these regressors in the presence of interactive fixed effects. We consider models where both "low-rank regressors" (like time-invariant and common regressors) and "high-rank-regressors" (almost all other regressors that vary across individuals and over time) are present simultaneously, while Bai (2009) only considers them separately. In order to work out the asymptotic theory of the QMLE, we use the quadratic approximation of the profile likelihood function which is developed by Moon and Weidner (2010) using the perturbation theory of linear operators. This methodology differs from Bai (2009), who starts his analysis from the first order conditions of the QMLE. An advantage of starting the analysis from the likelihood function itself is that we can also work out the asymptotic distributions of the likelihood ratio (LR) test and the Lagrange multiplier (LM) test for testing hypotheses on the regression coefficients. Asymptotic analysis of these two tests is new in the literature. We provide modified versions of the three classical test statistics (Wald-, LR-, and LM-test). Modification is required due to the incidental parameter biases. We show that these modified tests have chi-square limiting distributions.

Both Bai (2009) and our analysis in this paper share the restriction that the number of factors is assumed to be known. For pure factors models, *i.e.* in the absence of regressors, there is a sizable literature on how to estimate or test for the number of factors, see *e.g.* Bai and Ng (2002), Onatski (2005) and Harding (2007). The supplementary material of Bai (2009) informally discussed how to estimate the number of factors in the presence of regressors. Moon and Weidner (2010) show that under certain regularity conditions the asymptotic distribution of the QMLE remains unchanged even if the number of factors is not known accurately and one fits on redundant factors. Unfortunately, relaxing their regularity conditions to include the conditions of the present paper turns out to be mathematically very challenging and goes beyond the scope of this work.

For estimation we considers the QMLE. In the literature, various other estimation techniques for interactive factor models are studied. Holtz-Eakin, Newey, and Rosen (1988) study a panel regression model with factors and lagged dependent variables, *i.e.* also allowing for predetermined regressors. In their asymptotics T is fixed, so that the factors f_t cause no incidental parameter bias. To solve the incidental parameter problem for λ_i they estimate a quasi-differenced version of the model using appropriated lagged variables as instruments. For small T their parameter estimates are easy to obtain and are unbiased. However, implementing their method for large T is difficult since one has to minimize a non-linear objective function (*e.g.* for GMM) over many parameters – since the f_t (or their quotients) are estimated jointly with the regression parameters. Thus, with respect to the size of T the Holtz-Eakin, Newey, and Rosen (1988) method is complementary to our our approach, since our asymptotic is accurate only for large T. The same is true for Ahn, Lee and Schmidt (2001), who study the QMLE and a GMM estimator in fixed T asymptotic. To achieve consistency in this asymptotic they have to assume that the regressors are iid distributed across individuals. Pesaran (2006) discusses common correlated effect estimators for multi-factor models.

Regarding hypothesis testing in panel regression model with factors, Holtz-Eakin, Newey, and Rosen (1988) show that the LR-test is asymptotically χ^2-distributed in their 2SLS estimation framework with fixed T. Bai and Ng (2004), Moon and Perron (2004), and Phillips and Sul (2003) discuss various unit-root tests and derive their limiting distribution for $N, T \to \infty$.

The paper is organized as follows. In Section 2 we introduce the interactive fixed effect model and the QMLE of the regression parameters, and we provide a set of assumptions that are sufficient to show consistency of the QMLE. The assumptions allow for a predetermined regressors, and for a general combination of "low-rank" and "high-rank regressors".

In Section 3 we first summarize the approximation results in Moon and Weidner (2010), and then apply these result to work out the asymptotic distribution of the QMLE in the limit where N and T go to infinity at the same rate. For this, we assume independent idiosyncratic error terms, but allow for heteroscedasticity across individuals and over time. We find two types of asymptotic biases in the QMLE: The first type of bias is due to the presence of predetermined regressors. This corresponds to the classical incidental parameter problem in the dynamic panel data literature with fixed effects (Nickell, 1981). The second type of bias, which was also derived in Bai (2009), is due to heteroscedasticity (and more generally also due to correlation) of the idiosyncratic errors. This second bias is thus a result of misspecification of the model, *i.e.* of minimizing the sum of squared residuals (iid likelihood function), as opposed to a likelihood function that accounts for the heteroscedasticity (and correlation) in the error terms. We also provide consistent estimators for each component of the asymptotic bias and a bias corrected QMLE.

In Section 4 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 distribution. In Section 5 we present Monte Carlo simulation results for an AR(1) model with interactive fixed effect. The simulations show that the QMLE for the AR(1) coefficient is biased, and that the tests based on it can have severe size distortions and power asymmetries, while the bias corrected QMLE and test statistics have better properties. In Section 6 we conclude. All proofs of theorems and some technical details are presented in the appendix.[1]

A few words on notation. 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 $\|A\|_F = \sqrt{\text{Tr}(AA')}$, 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 use $P_A = A(A'A)^{-1}A'$ and $M_A = \mathbb{I} - A(A'A)^{-1}A'$, where $\mathbb{I}$ is the $m \times m$ identity matrix, and $(A'A)^{-1}$ denotes some generalized inverse if A is not of full column rank. For square matrices B, C, we use $B > C$ (or $B \geq C$) to indicate that $B - C$ is positive (semi) definite. 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, *i.e.* $\nabla f(x)$ is the row vector of partial derivatives of f with respect to each component of x. We use "wpa1" for "with probability approaching one".

2 Model, QMLE, and Consistency

In this paper we study the following panel regression model with cross-sectional size N and T time periods

$$Y_{it} = \beta^{0\prime} X_{it} + \lambda_i^{0\prime} f_t^0 + e_{it}, \qquad i = 1 \dots N, \quad t = 1 \dots T, \tag{2.1}$$

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. Throughout this paper we assume that the true number of factors R is known.[2] In matrix notation the model can be written as

$$Y = \sum_{k=1}^{K} \beta_k^0 \, X_k \; + \; \lambda^0 f^{0\prime} + e \; , \tag{2.2}$$

where Y, X_k and e are $N \times T$ matrices, λ^0 is a $N \times R$ matrix, and f^0 is a $T \times R$ matrix. The (Gaussian) quasi likelihood function of the model is simply the sum of squared residuals, which in matrix notation can be expressed as

$$\mathcal{L}_{NT}(\beta, \lambda, f) = \frac{1}{NT} \, \text{Tr} \left[\left(Y - \sum_{k=1}^{K} \beta_k X_k - \lambda f' \right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k - \lambda f' \right) \right] \, . \tag{2.3}$$

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

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{B}} \; L_{NT}\left(\beta\right) \; , \tag{2.4}$$

[1] Some parts of the proofs and some further technical comments are available from the supplementary material, which can be downloaded at `http://www-rcf.usc.edu/~moonr`.

[2] 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 and Ng (2002), Onatski (2005) and Harding (2007)). Alternatively, one could rely on Moon and Weidner (2010) who consider a regression model with interactive fixed effects when only an upper bound on the number of factors is known. However, it is mathematically very challenging to extend their results for a more general set up as considered in this paper.

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

$$
\begin{aligned}
L_{NT}(\beta) &= \min_{\lambda, f} \; \mathcal{L}_{NT}(\beta, \lambda, f) \\
&= \min_{f} \; \frac{1}{NT} \operatorname{Tr}\left[\left(Y - \sum_{k=1}^{K} \beta_k X_k \right) M_f \left(Y - \sum_{k=1}^{K} \beta_k X_k \right)' \right] \\
&= \frac{1}{NT} \sum_{t=R+1}^{T} \mu_t \left[\left(Y - \sum_{k=1}^{K} \beta_k X_k \right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k \right) \right] . \qquad (2.5)
\end{aligned}
$$

Here, the first expression for $L_{NT}(\beta)$ is its definition as the the minimum value of $\mathcal{L}_{NT}(\beta, \lambda, f)$ over λ and f. We denote the minimizing incidental parameters by $\hat{\lambda}(\beta)$ and $\hat{f}(\beta)$. The minimum over $\mathcal{L}_{NT}(\beta, \lambda, f)$ is unique, but the minimizing incidental parameters are not uniquely determined, since $\mathcal{L}_{NT}(\beta, \lambda, f)$ is invariant under transformations $\lambda \to \lambda A$ and $f \to f A^{-1}$, where A is a non-singular $R \times R$ matrix.

The second expression for $L_{NT}(\beta)$ in equation (2.5) is obtained by concentrating out λ (analogously, one can concentrate out f to obtain a formulation where 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 $\left(Y - \sum_{k=1}^{K} \beta_k X_k\right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k\right)$. This leads to the third line that presents the profile quasi likelihood function as the sum over the $T - R$ smallest eigenvalues of this $T \times T$ matrix.

This last expression for $L_{NT}(\beta)$ is our starting point when expanding $L_{NT}(\beta)$ around β^0. It is also most convenient for numerical computations of the QMLE — at each step of the numerical optimization over β one only needs to calculate the eigenvalues of a $T \times T$ matrix, which is much faster than minimizing over the high dimensional parameters λ and f.[3]

The objective function $L_{NT}(\beta)$ is continuous in β, but is usually neither convex nor differentiable everywhere. Multiple local minima may exist, and one should use multiple starting values for the numerical optimization to guarantee that the true global minimum is found. Theorem B.1 in the appendix shows equivalence of the three expressions for $L_{NT}(\beta)$ given above.

To show consistency of the QMLE $\hat{\beta}$ 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.[4]

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

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

Assumption 1 guarantees that the matrices f^0 and λ^0 have full rank, *i.e.* that there are R distinct factors and factor loadings asymptotically, and that the norm of each factor and factor

[3]For numerical purposes one should use the last expression in (2.5) if T is smaller than N. If T is larger than N one should use the symmetry of the problem ($N \leftrightarrow T$, $\lambda \leftrightarrow f$, $Y \leftrightarrow Y'$, $X_k \leftrightarrow X_k'$) and calculate $L_{NT}(\beta)$ as the sum over the $N - R$ smallest eigenvalues of the $N \times N$ matrix $\left(Y - \sum_{k=1}^{K} \beta_k X_k\right) \left(Y - \sum_{k=1}^{K} \beta_k X_k\right)'$.

[4]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.

loading grows at a rate of $\sqrt{T}$ and $\sqrt{N}$, respectively. Assumption 2 demands that the regressors are weakly exogenous. Assumption 3 will be discussed in more detail in the next section. It is a regularity condition on the the error term e_{it}, and we give examples of error distributions that satisfy this condition in Appendix A. The final assumption needed for consistency is an assumption on the regressors X_k.

Assumption 4.

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

(ii) We assume that the K regressors can be decomposed into K_1 low-rank regressors X_l, $l = 1,\ldots,K_1$, and $K_2 = K - K_1$ high-rank regressors X_m, $m = K_1+1,\ldots,K$. The two types of regressors satisfy:

(a) Consider linear combinations $X_{\text{high},\alpha} = \sum_{m=K_1+1}^{K}\alpha_m X_m$ of the high-rank regressors X_m for K_2-vectors[5] α with $\|\alpha\| = 1$. We assume that there exists a constant $b > 0$ such that

$$\min_{\{\alpha\in\mathbb{R}^{K_2},\|\alpha\|=1\}}\sum_{i=2R+K_1+1}^{N}\mu_i\left(\frac{X_{\text{high},\alpha}X'_{\text{high},\alpha}}{NT}\right)\geq b \qquad \textit{wpa1.}$$

(b) For the low-rank regressors we assume $\text{rank}(X_l) = 1$, $l = 1,\ldots,K_1$, i.e. they can be written as $X_l = w_l v_l'$ for $N\times 1$ vectors w_l and $T\times 1$ 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 that there exists a constant $B > 0$ (independent of N, T) such that $N^{-1}\,\lambda^{0\prime}\,M_w\,\lambda^0 \;>\; B\,\mathbb{I}_R$ wpa1, and $T^{-1}\,f^{0\prime}\,M_v\,f^0 \;>\; B\,\mathbb{I}_R$ wpa1.

The distinction between low-rank and high-rank regressors introduced in Assumption 4 is essential for showing consistency of the QMLE. The two most prominent examples of low-rank regressors are time-invariant regressors, which satisfy $X_{l,it} = X_{l,i\tau}$ for all i, t, τ, and common (or cross-sectionally invariant) regressors, which satisfy $X_{l,it} = X_{l,jt}$ for all i, j, t. To give another example of a low-rank regressor, let D_i and $\tilde{D}_t$ be dummy variables that indicate whether i is in a particular subset of $\{1,\ldots,N\}$ (group dummy), and whether t is in a particular subset of $\{1,\ldots,T\}$ (*e.g.* monthly dummy). The interacted dummy variable $X_{l,it} = D_i\tilde{D}_t$ then is a low-rank regressor, but is neither time-invariant nor common. In these examples, and probably for the vast majority of applications, the low-rank regressors all satisfy $\text{rank}(X_l) = 1$, as demanded in Assumption 4. However, none of our conclusions and proofs would be different if we allowed for low-rank regressors with rank larger than one as long as their rank remains constant as $N, T \to \infty$.[6]

The appearance of the factors and factor loadings in the assumption on the low-rank regressors is inevitable in order to guarantee consistency. For example, consider a low-rank regressor that is cross-sectionally independent and proportional to the r'th unobserved factor, *e.g.* $X_{l,it} = f_{tr}$. 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

[5]The components of the K_2-vector α are denoted by α_{K_1+1} to α_K.

[6]We would then have $X_l = w_l v_l'$, where w_l is a $N\times\text{rank}(X_l)$ matrix, and v_l is a $T\times\text{rank}(X_l)$. The definition of w and v would remain the same, but they would be $N\times R_X$ and $T\times R_X$ matrices, where $R_X = \sum_{l=1}^{K_1}\text{rank}(X_l)$ is the sum over the rank of all low-rank regressors. In addition, we would have to make a slight change in Assumption 4(ii)(a) on the high-rank regressors, namely replacing K_1 by R_X, *i.e.* we would have $\sum_{i=2R+R_x+1}^{N}$.

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. High-rank regressor are those whose distribution guarantees that they have high rank (usually full rank) asymptotically. 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 $\text{rank}(X_m) = \min(N, T)$, wpa1.[7] We can now state our consistency result for the QMLE.

Theorem 2.1. *Let Assumption 1, 2, 3, 4 be satisfied, and let the parameter set $\mathbb{B}$ be compact. In the limit $N, T \to \infty$ we then have*

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

The proof of the theorem and of all theorems below can be found in the appendix. We assume compactness of $\mathbb{B}$ to guarantee existence of the minimizing $\hat{\beta}$. We also use boundedness of $\mathbb{B}$ in the consistency proof, but only for those parameters β_l, $l = 1 \ldots K_1$, that correspond to low-rank regressors, *i.e.* if there are only high-rank regressors ($K_1 = 0$) the compactness assumption can be omitted, as long as existence of $\hat{\beta}$ is guaranteed (*e.g.* for $\mathbb{B} = \mathbb{R}^K$).

Bai (2009) also proves consistency of the QMLE 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, but for his consistency proof this assumption is not used. Regarding consistency, the real 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 discussed the case where all regressors are either time-invariant or common regressors (*i.e.* are low-rank). In 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.

3 Asymptotic Distribution and Bias Correction of the QMLE

Since we have already shown consistency of the QMLE $\hat{\beta}$, it is sufficient to study the local properties of the objective function $L_{NT}(\beta)$ around β^0 in order to derive the first order asymptotic theory of $\hat{\beta}$. A useful approximation of $L_{NT}(\beta)$ around β^0 was derived in Moon and Weidner (2010), and we briefly summarize the ideas and results of this approximation in the following subsection. Afterwards we apply these results to derive the asymptotic distribution of the QMLE. By providing consistent estimates of the asymptotic bias we then obtain a bias corrected QMLE.

3.1 Expansion of the Profile Quasi Likelihood

The last expression in equation (2.5) for the profile quasi likelihood is convenient because it does not involve any minimization over the parameters λ or f. On the other hand, this is not an

[7]To give a brief explanation of the assumption on high-rank regressors, let the $K_2 \times K_2$ matrix $\tilde{W}$ be defined by $\tilde{W}_{m_1 m_2} = (NT)^{-1}\text{Tr}(X_{m_1} X_{m_2}')$. If the sum over the eigenvalues in Assumption 4(ii)(a) would run over all eigenvalues $i = 1$ to N, it could be replaced by a trace, and the assumption would just be the conventional no-collinearity condition $\text{plim}_{N,T\to\infty} \tilde{W} > 0$. Assumption 4(ii)(a) is stricter than that since the first $2R + K_1$ eigenvalues are omitted from the sum. In particular, the matrix $X_m X_m'$ for each high-rank regressors needs to have more than $2R + K_1$ non-zero eigenvalues, *i.e.* high-rank regressors need to satisfy $\text{rank}(X_m) > 2R + K_1$, which explains their name.

expression that can be easily discussed by analytic means, because in general there is no explicit formula for the eigenvalues of a matrix. The conventional method that involves a Taylor series expansion in the regression parameters β *alone* seems infeasible here. Moon and Weidner (2010) showed how to overcome this problem by expanding the profile quasi likelihood *jointly* in β and $\|e\|$. The key idea is the following decomposition

$$Y - \sum_{k=1}^{K} \beta_k X_k = \underbrace{\lambda^0 f^{0\prime}}_{\substack{\text{leading} \\ \text{term}}} \underbrace{- \sum_{k=1}^{K} \left(\beta_k - \beta_k^0\right) X_k + e}_{\text{perturbation term}} . \tag{3.1}$$

If the perturbation term is zero, then the profile quasi likelihood $L_{NT}(\beta)$ is also zero, since 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 (1980)) one can work out an expansion of $L_{NT}(\beta)$ in the perturbation term, and one can show that this expansion is convergent as long as the spectral norm of the perturbation term is sufficiently small. For details we refer to Moon and Weidner (2010).

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

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

In the rest of the paper we only consider asymptotics where N and T grow at the same rate, *i.e.* we could equivalently write $o_p(T^{2/3})$ instead of $o_p(N^{2/3})$ in Assumption 3*. In Appendix A we provide examples of error distributions that satisfy Assumption 3*. In fact, for these examples, we have $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$. There is large literature that studies the asymptotic behavior of the spectral norm of random matrices, see *e.g.* Geman (1980), Silverstein (1989), Bai, Silverstein, Yin (1988), Yin, Bai, and Krishnaiah (1988), and Latala (2005). Loosely speaking, we expect the result $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$ 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 quasi likelihood function $L_{NT}(\beta)$ that was derived in Moon and Weidner (2010).

Theorem 3.1. *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 quasi likelihood 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

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

with $K \times K$ matrix W_{NT} defined by $W_{NT,k_1k_2} = (NT)^{-1}\,\mathrm{Tr}(M_{f^0}\,X'_{k_1}\,M_{\lambda^0}\,X_{k_2})$, 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

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

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 the particular order above is sufficient to work out the first order asymptotic theory of the QMLE, as the following corollary shows.

Corollary 3.2. *Let the assumptions of Theorem 2.1 and 3.1 hold, assume that β^0 is an interior point of the parameter set $\mathbb{B}$, and furthermore assume that $C_{NT} = \mathcal{O}_p(1)$. We then have $\sqrt{NT}\left(\hat{\beta} - \beta^0\right) = W_{NT}^{-1}C_{NT} + o_p(1) = \mathcal{O}_p(1)$.*

Combining consistency of the QMLE and the expansion of the profile quasi likelihood function in Theorem 3.1, one obtains $\sqrt{NT}\,W_{NT}\left(\hat{\beta} - \beta^0\right) = C_{NT} + o_p(1)$ (see *e.g.* Andrews (1999)). To obtain the corollary one needs in addition that W_{NT} does not become degenerate as $N,T \to \infty$, *i.e.* the smallest eigenvalue of W_{NT} should by bounded by a positive constant. Our assumptions already guarantee this, as is shown in the supplementary material.

3.2 Asymptotic Distribution of the QMLE

We now apply Corollary 3.2 to work out the asymptotic distribution of the QMLE $\hat{\beta}$. For this purpose we need more specific assumptions on λ^0, f^0, X_k, and e.

Assumption 5.

(i) In addition to Assumption 1 on λ^0 and f^0, we assume that $\|\lambda_i^0\|$ and $\|f_t^0\|$ are uniformly bounded across i,t and N,T.

(ii) The errors e_{it} are independent across i and t, they satisfy $\mathbb{E}e_{it} = 0$, and the eighth moment $\mathbb{E}e_{it}^8$ is bounded uniformly across i,t and N,T.

(iii) In addition to Assumption 4, we assume that the regressors X_k, $k = 1,\ldots,K$, can be decomposed as $X_k = X_k^{\rm str} + X_k^{\rm weak}$. The component $X_k^{\rm str}$ is strictly exogenous, i.e. $X_{k,it}^{\rm str}$ is independent of $e_{j\tau}$ for all i,j,t,τ. The component $X_k^{\rm weak}$ is predetermined and we assume

$$X_{k,it}^{\rm weak} = \sum_{\tau=1}^{t-1} c_{k,i\tau}\,e_{i,t-\tau}\;, \tag{3.2}$$

for some coefficients $c_{k,i\tau}$ that satisfy

$$|c_{k,i\tau}| \,<\, \alpha^{\tau}\;, \tag{3.3}$$

where $\alpha \in (0,1)$ *is a constant that is independent of* $\tau = 1\ldots, T-1$, $k = 1\ldots K$ *and* $i = 1\ldots N$. *We also assume that* $\mathbb{E}(X^{\rm str}_{k,it})^{8+\epsilon}$ *is bounded uniformly over* i, t *and* N, T, *for some* $\epsilon > 0$.

(iv) We consider a limit $N, T \to \infty$ *with* $N/T \to \kappa^2$, *where* $0 < \kappa < \infty$.

Assumption 5 is sufficient for Assumption 2 and for Assumption 3*. Thus, Assumption 5 guarantees consistency of the QMLE $\hat{\beta}$ and makes sure that the quadratic expansion of the profile quasi likelihood function is applicable.

Assumption 5(i) is convenient in order to calculate probability limits of expressions that involve λ^0 and f^0. One could weaken this assumption and only ask for existence and boundedness of appropriate higher moments of λ^0_i and f^0_t. Assumption 5(ii) requires cross-sectional and time-serial independence of e_{it}, but heteroscedasticity in both directions is still allowed. Assumption 5(iii) requires that the regressors X_k are additively separable into a strictly exogenous and a predetermined component and assumes that the predetermined component can be written as an MA(∞) process with innovation e_{it}.[8] An example where this is satisfied is if the interactive fixed effect model is one equation of a vector auto-regression for each cross-sectional unit, *e.g.* for the VAR(1) case we would have

$$\begin{pmatrix} Y_{it} \\ Z_{it} \end{pmatrix} = \mathcal{B} \begin{pmatrix} Y_{i,t-1} \\ Z_{i,t-1} \end{pmatrix} + \begin{pmatrix} \lambda^{0\prime}_i f^0_t \\ d_{it} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \Gamma & \mathbb{I} \end{pmatrix} \begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}, \qquad (3.4)$$

where Z_{it} is an $m \times 1$ vector of additional variables, $\mathcal{B}$ is an $(m+1) \times (m+1)$ matrix of parameters, the $m \times 1$ vectors d_{it} and u_{it} are independent of e_{it}, and Γ is an $m \times m$ covariance matrix. Here we already applied a Cholesky decomposition to the general form of the innovation of a VAR model in order to single our the shocks e_{it} that are genuine to Y_{it}. The first row in equation (3.4) is our interactive factor model with regressors $Y_{i,t-1}$ and $Z_{i,t-1}$, and due to the structure of the VAR process these regressors have a decomposition into strictly exogenous and predetermined regressors as demanded in Assumption 5(iii). The generalization of this example to VAR processes of higher order is straightforward.

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

Assumption 6. *Let* $\mathfrak{X}_k = M_{\lambda^0} \, X^{\rm str}_k \, M_{f^0} + X^{\rm weak}_k$. *For each* i *and* t, *define the* K*-vector* $\mathfrak{X}_{it} =$

[8] Actually, $X^{\rm weak}_k$ is only a truncated MA(∞) process, because it only depends on e_{it} for $t \geq 1$, but not on e_{it} for $t \leq 0$. However, one can define the decomposition $X_k = \tilde{X}^{\rm weak}_k + \tilde{X}^{\rm str}_k$ where $\tilde{X}^{\rm weak}_k = \sum_{\tau=1}^{\infty} c_{k,i\tau} \, e_{i,t-\tau}$ is a non-truncated MA(∞) process with innovation e_{it}, and $\tilde{X}^{\rm str}_k = X^{\rm str}_k - \sum_{\tau=t}^{\infty} c_{k,i\tau} \, e_{i,t-\tau}$ is still strictly exogenous.

$(\mathfrak{X}_{1,it}, \ldots, \mathfrak{X}_{K,it})'$. *We assume existence of the following probability limits*

$$
\begin{aligned}
W &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathfrak{X}_{it}\, \mathfrak{X}_{it}' \,, \\
\Omega &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[e_{it}^2\, \mathfrak{X}_{it}\, \mathfrak{X}_{it}' \right] \,, \\
B_{1,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{N} \operatorname{Tr}\left[P_{f^0} \mathbb{E}\left(e' X_k^{\text{weak}} \right) \right] \,, \\
B_{2,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{T} \operatorname{Tr}\left[\mathbb{E}\left(e e' \right)\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime} \right] \,, \\
B_{3,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{N} \operatorname{Tr}\left[\mathbb{E}\left(e' e \right)\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime} \right] \,.
\end{aligned}
\tag{3.5}
$$

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 3.3. *Let Assumptions 5 and 6 be satisfied, and let the true parameter β^0 be an interior point of the compact parameter set $\mathbb{B}$. Then we have*

$$
\sqrt{NT}\left(\hat{\beta} - \beta^0\right) \underset{d}{\to} \mathcal{N}\left(W^{-1} B,\; W^{-1}\, \Omega\, W^{-1} \right) \,, \tag{3.6}
$$

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

From Corollary 3.2 we already know that the limiting distribution of $\hat{\beta}$ is given by the limiting distribution of $W_{NT}^{-1} C_{NT}$. To prove Theorem 3.3 one first has to show that $W = \operatorname{plim}_{N,T\to\infty} W_{NT}$. We could have defined W this way, but the definition given in Assumption 6 is equivalent. Note, however, that the predetermined part of $\mathfrak{X}_k$ is not projected with M_{f^0} and M_{λ^0}. The intuition here is that since by assumption X_k^{weak} is uncorrelated with λ^0 and f^0 it does not matter asymptotically whether the corresponding subspaces (of fixed dimension) are projected out of X_k^{weak}. For the strictly exogenous part of the regressors this is different, because X_k^{str} can be correlated with λ^0 and f^0, and may have a significant part that is proportional to λ^0 and f^0 and that is projected out by M_{f^0} and M_{λ^0}. Assumption 4 guarantees that W is positive definite.

The second step in proving the theorem is to show that the approximated score at the true parameter satisfies $C_{NT} \to_d \mathcal{N}(B, \Omega)$. The asymptotic variance Ω and the asymptotic bias B_1 originate exclusively from the $C^{(1)}$ term. The strictly exogenous part of the regressors only contributes to the asymptotic variance, but the predetermined part contributes to both, namely to the asymptotic variance via the term $\operatorname{Tr}(e' X_k^{\text{weak}})$ and to the bias B_1 via the term $\operatorname{Tr}(P_{f^0}\, e'\, X_k^{\text{weak}})$. 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 $\tau > t$), and does not appear if the regressor is strictly exogenous as in Bai (2009).

The three term in $C^{(2)}$ contribute no variance, *i.e.* they converge to constants in probability. One term in $C^{(2)}$ is vanishing. The other two terms contribute the asymptotic biases B_2 and B_3 that are due to cross-sectional and time-serial heteroscedasticity, as in Bai (2009). Note that the predetermined part of the regressors does not contribute to B_2 and B_3.

3.3 Bias Correction

In order to express our estimators for the asymptotic bias and the asymptotic variance of $\hat{\beta}$ we first have to introduce some notation.

Definition 3.4. *Let η_i and η_t be the N and T dimensional unit column vectors that have unity at position i and t, respectively, and zeros everywhere else. Let $\Gamma(.)$ 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. For an $N \times N$ matrix A and a $T \times T$ matrix B we define*

(i) the diagonal truncation $A^{\text{truncD}} = \sum_{i=1}^{N} \eta_i \eta_i' A \eta_i \eta_i'$, $B^{\text{truncD}} = \sum_{t=1}^{T} \eta_t \eta_t' B \eta_t \eta_t'$,

(ii) the right-sided and left-sided Kernel truncation $B^{\text{truncR}} = \sum_{t=1}^{T-1} \sum_{\tau=t+1}^{T} \Gamma\left(\frac{t-\tau}{M}\right) \eta_t \eta_t' B \eta_\tau \eta_\tau'$, $B^{\text{truncL}} = \sum_{t=2}^{T} \sum_{\tau=1}^{t-1} \Gamma\left(\frac{t-\tau}{M}\right) \eta_t \eta_t' B \eta_\tau \eta_\tau'$.

We now define our estimators for W, Ω, B_1, B_2, and B_3.

Definition 3.5. *Let $\hat{\mathfrak{X}}_k(\beta) = M_{\hat{\lambda}}(\beta) X_k M_{\hat{f}}(\beta)$. For each i and t, define the K-vector $\hat{\mathfrak{X}}_{it}(\beta) = (\hat{\mathfrak{X}}_{1,it}(\beta), \ldots, \hat{\mathfrak{X}}_{K,it}(\beta))'$. We define the $K \times K$ matrices $\hat{W}(\beta)$ and $\hat{\Omega}(\beta)$, and the K-vectors $\hat{B}_1(\beta)$, $\hat{B}_2(\beta)$, $\hat{B}_3(\beta)$, and $\hat{B}(\beta)$ as follows*

$$\begin{aligned}
\hat{W}(\beta) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\mathfrak{X}}_{it} \hat{\mathfrak{X}}_{it}' , \\
\hat{\Omega}(\beta) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2 \hat{\mathfrak{X}}_{it} \hat{\mathfrak{X}}_{it}' , \\
\hat{B}_{1,k}(\beta) &= \frac{1}{N} \operatorname{Tr}\left[P_{\hat{f}} \left(\hat{e}' X_k\right)^{\text{truncR}} \right] , \\
\hat{B}_{2,k}(\beta) &= \frac{1}{T} \operatorname{Tr}\left[\left(\hat{e}\, \hat{e}'\right)^{\text{truncD}} M_{\hat{\lambda}} X_k \hat{f} (\hat{f}'\hat{f})^{-1} (\hat{\lambda}'\hat{\lambda})^{-1} \hat{\lambda}' \right] , \\
\hat{B}_{3,k}(\beta) &= \frac{1}{N} \operatorname{Tr}\left[\left(\hat{e}'\, \hat{e}\right)^{\text{truncD}} M_{\hat{f}} X_k' \hat{\lambda} (\hat{\lambda}'\hat{\lambda})^{-1} (\hat{f}'\hat{f})^{-1} \hat{f}' \right] , \\
\hat{B}(\beta) &= -\sqrt{\frac{N}{T}} \hat{B}_{1,k}(\beta) - \sqrt{\frac{T}{N}} \hat{B}_{2,k}(\beta) - \sqrt{\frac{N}{T}} \hat{B}_{3,k}(\beta) ,
\end{aligned} \tag{3.7}$$

where we suppressed the β-dependence of $\mathfrak{X}$, $\hat{e}$, $\hat{f}$, and $\hat{\lambda}$ on the right hand side.[9]

The estimators above are dependent on β, since one needs an estimator for β in order to obtain the residuals $\hat{e}$ and the estimators for the factors and factor loadings.

Theorem 3.6. *Under Assumptions 5 and 6, for $M \to \infty$ and $M^5/T \to 0$, and for any $\sqrt{NT}$-consistent estimator $\hat{\beta} = \beta^0 + \mathcal{O}_p((NT)^{-1/2})$ we have $\hat{W}(\hat{\beta}) = W + o_p(1)$, $\hat{\Omega}(\hat{\beta}) = \Omega + o_p(1)$, $\hat{B}_1(\hat{\beta}) = B_1 + o_p(1)$, $\hat{B}_2(\hat{\beta}) = B_2 + o_p(1)$, and $\hat{B}_3(\hat{\beta}) = B_3 + o_p(1)$.*

[9]Here $\hat{f}(\beta)$ and $\hat{\lambda}(\beta)$ are the principal component estimators defined above, and we write $M_{\hat{\lambda}}(\beta)$ and $M_{\hat{f}}(\beta)$ instead of $M_{\hat{\lambda}(\beta)}$ and $M_{\hat{f}(\beta)}$. The corresponding residuals are given by $\hat{e}(\beta) = Y - \sum_{k=1}^{K} \beta_k X_k - \hat{\lambda}(\beta) \hat{f}'(\beta)$. If there is serial correlation in e_{it}, one can replace truncD in $\hat{B}_{3,k}(\beta)$ by truncK, which is a Kernel truncation on the first few center diagonals of a $T \times T$ matrix A, *i.e.* $A^{\text{truncK}} = \sum_{t=1}^{T} \sum_{\tau=1}^{T} \Gamma\left(\frac{t-\tau}{M}\right) \eta_t \eta_t' A \eta_\tau \eta_\tau'$.

Notice that the estimators $\hat{\Omega}(\hat{\beta})$, $\hat{B}_2(\hat{\beta})$, and $\hat{B}_3(\hat{\beta})$ are similar to White's standard error estimator under heteroskedasticity and the estimator $\hat{B}_1(\hat{\beta})$ is similar to the HAC estimator with a kernel. 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 $\hat{W}$, $\hat{\Omega}$, and $\hat{B}_i$, $i = 1, 2, 3$, it is not necessary to know whether the regressors are strictly exogenous or predetermined; 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 3.7. *Let Assumptions 5 and 6 hold, let β^0 be an interior point of the compact parameter set $\mathbb{B}$, and let $M \to \infty$ with $M^5/T \to 0$. Then, the bias corrected QMLE*

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

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

According to Theorem 3.6, a consistent estimator of the asymptotic variance of $\hat{\beta}^*$ is given by $\hat{W}^{-1}(\hat{\beta})\,\hat{\Omega}(\hat{\beta})\,\hat{W}^{-1}(\hat{\beta})$.

4 Testing Restrictions on β^0

In this section we discuss the three classical test statistics for testing general linear restrictions on β^0, *i.e.* 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 easily generalize the discussion to the testing of non-linear hypotheses. Throughout this subsection we assume that β^0 is an interior point of $\mathbb{B}$, *i.e.* there are no local restrictions 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

$$\tilde{\beta} = \underset{\beta\in\tilde{\mathbb{B}}}{\operatorname{argmin}}\ L_{NT}(\beta)\ , \tag{4.1}$$

where $\tilde{\mathbb{B}} = \{\beta \in \mathbb{B} |\, H\beta = h\}$ is the restricted parameter set. Analogous to Theorem 3.3 for the unrestricted estimator $\hat{\beta}$, we can use the expansion of the profile quasi likelihood function to derive the limiting distribution of the restricted estimator. Under Assumptions 5 and 6 we have

$$\sqrt{NT}(\tilde{\beta} - \beta^0)\ \underset{d}{\longrightarrow}\ \mathcal{N}\left(\mathfrak{W}^{-1}B,\ \mathfrak{W}^{-1}\,\Omega\,\mathfrak{W}^{-1}\right)\ , \tag{4.2}$$

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

Wald Test

Using the results above we find that under the null-hypothesis $\sqrt{NT}\left(H\hat{\beta} - h\right)$ 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 statistics $WD_{NT} = NT\left(H\hat{\beta}-h\right)'\left(H\hat{W}^{-1}\,\hat{\Omega}\,\hat{W}^{-1}H'\right)^{-1}\left(H\hat{\beta}-h\right)$ is not asymptotically χ^2_r distributed. Using our estimator for the bias, it is natural to define the bias corrected Wald test statistics as

$$WD^*_{NT} = \left[\sqrt{NT}\left(H\hat{\beta}-h\right) - H\hat{W}^{-1}\hat{B}\right]'\left(H\hat{W}^{-1}\,\hat{\Omega}\,\hat{W}^{-1}H'\right)^{-1}\left[\sqrt{NT}\left(H\hat{\beta}-h\right) - H\hat{W}^{-1}\hat{B}\right]. \tag{4.3}$$

Under the null hypothesis we find $WD^*_{NT} \to_d \chi^2_r$ if Assumptions 5 and 6 are satisfied. Here we used $\hat{B} = \hat{B}(\hat{\beta})$, $\hat{W} = \hat{W}(\hat{\beta})$, and $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$.

Likelihood Ratio Test

We start the discussion of the LR test by assuming that $\Omega = cW$ for some scalar constant $c > 0$, and that we have a consistent estimator $\hat{c}$ for c. This condition is satisfied in our interactive fixed effect model if Assumptions 5 and 6 hold, and if $\mathbb{E}e^2_{it} = c$, *i.e.* if the error is homoskedastic. A consistent estimator for c in this context is $\hat{c} = (NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T \hat{e}^2_{it}$, where $\hat{e} = \hat{e}(\hat{\beta})$. Since the likelihood function for the interactive fixed effect model is just the sum of squared residuals, we have $\hat{c} = L_{NT}(\hat{\beta})$. However, different estimators for c can be used.

The likelihood ratio test statistics is defined by

$$LR_{NT} = \hat{c}^{-1}\,NT\left[L_{NT}\left(\tilde{\beta}\right) - L_{NT}\left(\hat{\beta}\right)\right]\ . \tag{4.4}$$

Under Assumptions 5 and 6 we then have

$$LR_{NT} \underset{d}{\longrightarrow} c^{-1}C'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}C\ , \tag{4.5}$$

where $C \sim \mathcal{N}(B,\Omega)$, i.e. $C_{NT} \to_d C$. This is the same limiting distribution that one finds for the Wald test under $\Omega = cW$ (in fact, one can show $WD_{NT} = LR_{NT} + o_p(1)$). Therefore, we need to do a bias correction for the LR test in order to achieve a χ^2 limiting distribution. We define

$$LR^*_{NT} = \hat{c}^{-1}NT\left[\min_{\{\beta\in\mathbb{B}|\,H\beta=h\}} L_{NT}\left(\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}\right) - \min_{\beta\in\mathbb{B}} L_{NT}\left(\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}\right)\right], \tag{4.6}$$

where $\hat{B} = \hat{B}(\hat{\beta})$ and $\hat{W} = \hat{W}(\hat{\beta})$ do not depend on the parameter β in the minimization problem.[10] Asymptotically we have $\min_{\beta\in\mathbb{B}} L_{NT}\left(\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}\right) = L_{NT}(\hat{\beta})$, because $\beta \in \mathbb{B}$ does not impose local constraints, *i.e.* close to β^0 it does not matter for the value of the minimum whether one minimizes over β or over $\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}$. The correction to the LR test therefore originates from the first term in LR^*_{NT}. For the minimization over the restricted parameter set it matters whether the argument of L_{NT} is β or $\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}$,

[10] Alternatively, one could use $\hat{B}(\tilde{\beta})$ and $\hat{W}(\tilde{\beta})$ 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 which specification should be preferred is left for future research.

because generically we have $HW^{-1}B \neq 0$ (otherwise no correction would be necessary for the LR statistics). One can show that

$$LR^*_{NT} \underset{d}{\longrightarrow} c^{-1}(C-B)'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}(C-B)\,, \tag{4.7}$$

i.e. we obtain the same formula as for LR_{NT}, but the limit of the score C is replaced by the bias corrected term $C - B$. Under Assumptions 5 and 6, if H_0 is satisfied we therefore find $LR^*_{NT} \to_d \chi^2_r$. One can also show that $LR^*_{NT} = WD^*_{NT} + o_p(1)$ under H_0.

Lagrange Multiplier Test

The quasi-likelihood function was defined in equation (2.3). Its gradient with respect to β evaluated at the restricted estimates is denoted $\tilde{\nabla}\mathcal{L}_{NT}$, *i.e.*

$$\begin{aligned}\tilde{\nabla}\mathcal{L}_{NT} \equiv \nabla\mathcal{L}_{NT}(\tilde{\beta},\tilde{\lambda},\tilde{f}) &= \left(\left.\frac{\partial\mathcal{L}_{NT}(\beta,\tilde{\lambda},\tilde{f})}{\partial\beta_1}\right|_{\beta=\tilde{\beta}}, \ldots, \left.\frac{\partial\mathcal{L}_{NT}(\beta,\tilde{\lambda},\tilde{f})}{\partial\beta_K}\right|_{\beta=\tilde{\beta}} \right)' \\ &= -\frac{2}{NT}\Big(\mathrm{Tr}\left(X_1'\tilde{e}\right), \ldots, \mathrm{Tr}\left(X_K'\tilde{e}\right)\Big)'\,, \end{aligned} \tag{4.8}$$

where $\tilde{\lambda} = \hat{\lambda}(\tilde{\beta})$, $\tilde{f} = \hat{f}(\tilde{\beta})$, and $\tilde{e} = \hat{e}(\tilde{\beta})$. Under Assumptions 5 and 6, and if the null hypothesis $H_0:\ H\beta^0 = h$ is satisfied, one finds that[11]

$$\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} = \sqrt{NT}\,\nabla L_{NT}(\tilde{\beta}) + o_p(1). \tag{4.9}$$

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

Using the bound on the remainder $R_{NT}(\beta)$ given in Theorem 3.1, one cannot infer any properties of the score function, *i.e.* of the gradient $\nabla L_{NT}(\beta)$, because nothing is said about $\nabla R_{NT}(\beta)$. The following theorem gives the bound on $\nabla R_{NT}(\beta)$ that is sufficient to derive the limiting distribution of the Lagrange multiplier.

Theorem 4.1. *Under the assumptions of Theorem 3.1 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 $\eta_{NT} \to 0$

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

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

$$\begin{aligned}\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} &= \sqrt{NT}\,\nabla L_{q,NT}(\tilde{\beta}) + o_p(1) \\ &= 2\sqrt{NT}\,W_{NT}\,(\tilde{\beta}-\beta^0) - 2\,C_{NT} + o_p(1)\,.\end{aligned} \tag{4.10}$$

[11]The proof of the statement is given in the appendix as part of the proof of Theorem 4.2.

Using this result and the known limiting distribution of $\tilde{\beta}$ we now find

$$\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} \;\underset{d}{\longrightarrow}\; -2H'(HW^{-1}H')^{-1}HW^{-1}C \;. \tag{4.11}$$

The LM test statistics is given by[12]

$$LM_{NT} = \frac{NT}{4}\,(\tilde{\nabla}\mathcal{L}_{NT})'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}\tilde{\nabla}\mathcal{L}_{NT} \;, \tag{4.12}$$

where $\tilde{B} = \hat{B}(\tilde{\beta})$, $\tilde{W} = \hat{W}(\tilde{\beta})$ and $\tilde{\Omega} = \hat{\Omega}(\tilde{\beta})$. One can show that the LM test is asymptotically equivalent to the Wald test: $LM_{NT} = WD_{NT} + o_p(1)$, *i.e.* again bias correction is necessary. We define the bias corrected LM test statistics as

$$LM^*_{NT} = \frac{1}{4}\left(\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} + 2\tilde{B}\right)'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}\left(\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} + 2\tilde{B}\right) \;. \tag{4.13}$$

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

Theorem 4.2. *Let Assumptions 5 and 6 and the null hypothesis $H_0 :\; H\beta^0 = h$ be satisfied, and let $\hat{\beta}$ and $\tilde{\beta}$ be the unrestricted and restricted parameter estimates. Let the estimators $\hat{W}(\beta)$, $\hat{\Omega}(\beta)$, and $\hat{B}(\beta)$ be the ones given in definition 3.5. For the bias corrected Wald and LM test statistics introduced in equation* (4.3) *and* (4.13) *we then have*

$$WD^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r \;, \qquad\qquad LM^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r \;. \tag{4.14}$$

If in addition we assume $\mathbb{E}e^2_{it} = c$, i.e. the idiosyncratic errors are homoscedastic, and we use $\hat{c} = L_{NT}(\hat{\beta})$ as an estimator for c, then the LR test statistics defined in equation (4.6) *satisfies*

$$LR^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r \;. \tag{4.15}$$

5 Monte Carlo Simulations

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

$$Y_{it} \;=\; \rho\, Y_{i,t-1} \;+\; \lambda_i\, f_t \;+\; e_{it} \;. \tag{5.1}$$

We estimate the model as an interactive fixed effect model, *i.e.* no distributional assumption on λ_i and f_t are made in the estimation, but Assumption 5 is assumed to hold, in particular the e_{it} are assumed to be independent across i and t. The parameter of interest is ρ. The estimators we consider are the OLS estimator (which completely ignores the presence of the factors), the QMLE defined in equation (2.4),[13] and the bias corrected QMLE (BC-QMLE) defined in Theorem 3.7.

For the simulation we draw e_{it} independently distributed from $\mathcal{N}(0,1)$, the λ^0_i independently distributed from $\mathcal{N}(1,1)$, and we generate the factors from an AR(1) specification, *i.e.* $f^0_t = \rho_f\, f^0_{t-1} + u_t$, where $u_t \sim \text{iid}\mathcal{N}(0,(1-\rho^2_f)\sigma^2_f)$, and σ_f is the standard deviation of f_t.[14] In this

[12] Note also that $\sqrt{NT}HW^{-1}\nabla L_{NT}(\tilde{\beta}) \to_d -2HW^{-1}C$.

[13] 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.

[14] For all simulations we generate 1000 initial time periods for f^0_t and y_{it} that are not used for estimation. This guarantees that that the simulated data used for estimation is distributed according to the stationary distribution of the model. We also note that the distributional assumptions on f^0_t and λ^0_i made here do not satisfy Assumption 5(i), but nevertheless all theorems above are applicable since f^0_t and λ^0_i have arbitrary high uniformly bounded moments.

setup there is no correlation and heteroscedasticity in e_{it}, *i.e.* only the bias term B_1 of the QMLE is non-zero, but we ignore this information in the estimation, *i.e.* we correct for all three bias terms (B_1, B_2, and B_3, as introduced in Assumption 6) in the bias corrected QMLE.

Table 1 shows the simulation results for the bias, standard error and root mean square error of the three different estimators for the case $N = 100$, $\rho_f = 0.5$, $\sigma_f = 0.5$, and different values of ρ and T. As expected, the OLS estimator is biased due to the factor structure and its bias does not vanish (it actually increases) as T increases. The QMLE is also biased, but as predicted by the theory its bias vanishes as T increases. The bias corrected QMLE performs even better than the non-corrected QMLE, in particular its bias vanishes even faster. Since we only correct for the first order bias of the QMLE, we could not expect the bias corrected QMLE to be unbiased. However, as T gets larger more and more of the QMLE bias is corrected for: at $T = 5$ the bias correction only corrects for about half for the QMLE bias, while at $T = 80$ it already corrects for about 90% of it.

In our setup we have $\|\lambda f'\| \approx \sqrt{2NT}\sigma_f$ and $\|e\| \approx \sqrt{N} + \sqrt{T}$.[15] Assumption 1 and 3 imply that asymptotically $\|\lambda f'\| \gg \|e\|$. We can therefore only be sure that the asymptotic results for the QMLE distribution are a good approximation of the finite sample properties if $\|\lambda f'\| \gtrsim \|e\|$, *i.e.* if $\sqrt{2NT}\sigma_f \gtrsim \sqrt{N} + \sqrt{T}$. In table 2 we present simulation results for $N = 100$, $T = 20$, $\rho = 0.6$ and different values of ρ_f and σ_f. In the case $\sigma_f = 0$ we have $0 = \|\lambda f'\| \ll \|e\|$, 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 QMLE and the bias corrected QMLE perform worse than OLS at finite sample (though we suspect that all three estimators are asymptotically equivalent), but the bias corrected QMLE has a lower bias and a lower variance than the non-corrected QMLE. The case $\sigma_f = 0.2$ corresponds to $\|\lambda f'\| \approx \|e\|$, and one finds that the bias and the variance of the OLS estimator and of the QMLE are of comparable size. However, the bias corrected QMLE already has much smaller bias and a bit smaller variance in this case. Finally, in the case $\sigma_f = 0.5$ we have $\|\lambda f'\| > \|e\|$, and we expect our asymptotic result 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 QMLE, while the bias corrected QMLE performs even better in terms of bias and variance.

An import issue is the choice of bandwidth M for the bias correction. Table 3 gives the fraction of the QMLE 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 optimal bandwidth depends on ρ: it is approximately $M = 2$ for $\rho = 0$, $M = 4$ for $\rho = 0.3$ and $\rho = 0.6$, and $M = 6$ for $\rho = 0.9$. Choosing the bandwidth too large or too small results in a smaller fraction of the bias to be corrected, *i.e.* in a larger bias of the bias corrected QMLE. 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 of optimizing the bandwidth.

In table 4 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 distortions of these test 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 $T = 80$. In contrast, the size distortions for the bias corrected Wald, LR, and LM test are much smaller, and tend to

[15] To be precise, we have $\|\lambda f'\|/(\sqrt{2NT}\sigma_f) \to_p 1$, and $\|e\|/(\sqrt{N} + \sqrt{T}) \to_p 1$.

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

In table 5 and 6 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. Since both the QMLE and the bias corrected QMLE 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. In 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, while the size corrected power for the right sided alternative is about the same.

6 Conclusions

This paper studies the (Gaussian) QMLE for dynamic linear panel regression models with interactive fixed effects. We provide conditions under which the QMLE is consistent and that allow for predetermined regressors, and for a general combination of "low-rank" and "high-rank" regressors. We show how a quadratic approximation of the profile likelihood function $L_{NT}(\beta)$ can be used to derive the first order asymptotic theory of the QMLE of β under the alternative asymptotic $N, T \to \infty$. We work out the asymptotic distribution of the QMLE and find that it can be asymptotically biased (i) due to weak exogeneity of the regressors and (ii) due and heteroscedasticity (and correlation) of the idiosyncratic errors e_{it}. Consistent estimators for the asymptotic covariance matrix and for the asymptotic bias of the QMLE are provided, and thus a bias corrected QMLE is given. We derive the asymptotic distribution of the Wald, LR and LM test statistics for testing a general linear hypothesis on β. The test statistics are not chi-square due to the asymptotic bias of the score and of the QMLE. We provide bias corrected test statistics and show that their asymptotic distribution is chi-square. The findings of the Monte Carlo experiments show that our asymptotic results on the distribution of the (bias corrected) QMLE and of the (bias corrected) test statistics provide a good approximation of their finite sample properties. Although the bias corrected QMLE has a non-zero bias at finite sample, this bias is much smaller than the one of the QMLE. Analogously, the size distortions and power asymmetries of the bias corrected Wald, LR and LM test are much smaller than for the non-bias corrected versions.

Appendix

A Examples of Error Distributions

Under each of the following distributional assumptions on the errors e_{it}, $i = 1,\ldots,N$, $t = 1,\ldots,T$, we have $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$. The proofs are given in the supplementary material.

(i) The e_{it} are independent across i and t, they satisfy $\mathbb{E}e_{it} = 0$, and $\mathbb{E}e_{it}^4$ is bounded uniformly over i,t and N,T.

(ii) The e_{it} follow different MA(∞) process for each i, namely

$$e_{it} = \sum_{\tau=0}^{\infty} \psi_{i\tau}\, u_{i,t-\tau} \; , \qquad \text{for } i = 1\ldots N,\; t = 1\ldots T \; , \tag{A.1}$$

where the u_{it}, $i = 1\ldots N$, $t = -\infty\ldots T$ are independent random variables with $\mathbb{E}u_{it} = 0$ and $\mathbb{E}u_{it}^4$ uniformly bounded across i,t and N,T. The coefficients $\psi_{i\tau}$ satisfy

$$\sum_{\tau=0}^{\infty} \tau \max_{i=1\ldots N} \psi_{i\tau}^2 \, < \, B \; , \qquad \sum_{\tau=0}^{\infty} \max_{i=1\ldots N} |\psi_{i\tau}| \, < \, B \; , \tag{A.2}$$

for a finite constant B which is independent of N and T.

(iii) The error matrix e is generated as $e = \sigma^{1/2}\, u\, \Sigma^{1/2}$, where u is an $N \times T$ matrix with independently distributed entries u_{it} and $\mathbb{E}u_{it} = 0$, $\mathbb{E}u_{it}^2 = 1$, and $\mathbb{E}u_{it}^4$ is bounded uniformly across i,t and N,T. Here σ is the $N \times N$ cross-sectional covariance matrix, and Σ is $T \times T$ time-serial covariance matrix, and they satisfy

$$\max_{j=1\ldots N} \sum_{i=1}^{N} |\sigma_{ij}| \, < \, B \; , \qquad \max_{\tau=1\ldots T} \sum_{t=1}^{T} |\Sigma_{t\tau}| \, < \, B \; , \tag{A.3}$$

for some finite constant B which is independent of N and T. In this example we have $\mathbb{E}e_{it}e_{j\tau} = \sigma_{ij}\Sigma_{t\tau}$.

B Proof of Consistency (Theorem 2.1)

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

Theorem B.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 a $N \times R$, f be a $T \times R$ matrix, $\tilde{\lambda}$ be a $N \times R_1$ matrix, and $\tilde{f}$ be a $T \times R_2$ matrix. Then the following six expressions (that are functions of Z only) are equivalent:*

$$\begin{aligned}
&\min_{f,\lambda} \operatorname{Tr}\left[\left(Z - \lambda f'\right)\left(Z' - f\lambda'\right)\right] = \min_{f} \operatorname{Tr}(Z\, M_f\, Z') = \min_{\lambda} \operatorname{Tr}(Z'\, M_\lambda\, Z) \\
&= \min_{\tilde{\lambda},\tilde{f}} \operatorname{Tr}(M_{\tilde{\lambda}}\, Z\, M_{\tilde{f}}\, Z') = \sum_{i=R+1}^{T} \mu_i(Z'Z) = \sum_{i=R+1}^{N} \mu_i(ZZ')
\end{aligned}$$

In the above minimization problems we do not have to restrict the matrices λ, f, $\tilde\lambda$ and $\tilde f$ to be of full rank. If for example λ is not of full rank we can still define $(\lambda'\lambda)^{-1}$ as the generalized inverse (*e.g.* via singular value decomposition). The projector M_λ is therefore still defined in this case and satisfied $M_\lambda \lambda = 0$ and $\text{rank}(M_\lambda) = N - \text{rank}(\lambda)$. If $\text{rank}(Z) \geq R$ then the optimal λ, f, $\tilde\lambda$ and $\tilde f$ always have full rank.

Theorem B.1 shows the equivalence of the three different versions of the profile quasi likelihood function in equation (2.5). It goes beyond this by also considering minimization of $\text{Tr}(M_{\tilde\lambda}\, Z\, M_{\tilde f}\, Z')$ over $\tilde\lambda$ and $\tilde f$, 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 B.2. *Under the assumptions of Theorem 2.1 we have*

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

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

Proof. By Assumption 2 we know that the first equation in Lemma B.2 is satisfied when replacing M_f by the identity matrix. So we are left to show $\max_f \left| \frac{1}{NT} \text{Tr}(\Xi\, e') \right| = o_p(1)$, 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 $\text{rank}(\Xi) \leq R$. We therefore find[16]

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

∎

Proof of Theorem 2.1. For the second version of the profile quasi likelihood function in equation (2.5) we write $L_{NT}(\beta) = \min_f\ S_{NT}(\beta, f)$, where

$$S_{NT}(\beta, f) = \frac{1}{NT}\ \text{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^K (\beta_k^0 - \beta_k) X_k + e \right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^K (\beta_k^0 - \beta_k) X_k + e \right)' \right], \tag{B.2}$$

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

$$\begin{aligned} S_{NT}(\beta, f) &= S_{NT}(\beta^0, f^0) + \tilde S_{NT}(\beta, f) \\ &\qquad + \frac{2}{NT}\ \text{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^K (\beta_k^0 - \beta_k) X_k \right)\, M_f\, e' \right] + \frac{1}{NT}\ \text{Tr}\left(e\, (P_{f^0} - P_f)\, e' \right) \\ &= S_{NT}(\beta^0, f^0) + \tilde S_{NT}(\beta, f) + o_p(\|\beta - \beta^0\|) + o_p(1)\ , \end{aligned} \tag{B.3}$$

where we defined

$$\tilde S_{NT}(\beta, f) = \frac{1}{NT}\ \text{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^K (\beta_k^0 - \beta_k) X_k \right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^K (\beta_k^0 - \beta_k) X_k \right)' \right]. \tag{B.4}$$

[16] Here we use $|\text{Tr}\,(C)| \leq \|C\|\, \text{rank}\,(C)$, which holds for all square matrices C, see the supplementary material.

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, since we allow for more general low-rank regressors, and since we allow for high-rank and low-rank regressors simultaneously. We split $\tilde{S}_{NT}(\beta, f) = \tilde{S}^{(1)}_{NT}(\beta, f) + \tilde{S}^{(2)}_{NT}(\beta, f)$, where

$$\begin{aligned}
\tilde{S}^{(1)}_{NT}(\beta, f) &= \frac{1}{NT} \operatorname{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta^0_k - \beta_k) X_k\right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta^0_k - \beta_k) X_k\right)' M_{(\lambda_0, w)}\right] \\
&= \frac{1}{NT} \operatorname{Tr}\left[\left(\sum_{m=K_1+1}^{K} (\beta^0_m - \beta_m) X_m\right) M_f \left(\sum_{m=K_1+1}^{K} (\beta^0_m - \beta_m) X_m\right)' M_{(\lambda_0, w)}\right], \\
\tilde{S}^{(2)}_{NT}(\beta, f) &= \frac{1}{NT} \operatorname{Tr}\left[\left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta^0_k - \beta_k) X_k\right) M_f \left(\lambda^0 f^{0\prime} + \sum_{k=1}^{K} (\beta^0_k - \beta_k) X_k\right)' P_{(\lambda_0, w)}\right],
\end{aligned} \tag{B.5}$$

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 $\tilde{S}^{(1)}_{NT}(\beta, f)$ we can apply Theorem B.1 with $\tilde{f} = f$ and $\tilde{\lambda} = (\lambda^0, w)$ (the R in the theorem is now $2R + K_1$) to find

$$\begin{aligned}
\tilde{S}^{(1)}_{NT}(\beta, f) &\geq \frac{1}{NT} \sum_{i=2R+K_1+1}^{N} \mu_i \left[\left(\sum_{m=K_1+1}^{K} (\beta^0_m - \beta_m) X_m\right)\left(\sum_{m=K_1+1}^{K} (\beta^0_m - \beta_m) X_m\right)'\right] \\
&\geq b \left\|\beta^{\text{high}} - \beta^{\text{high}}_0\right\|^2, \qquad \text{wpa1},
\end{aligned} \tag{B.6}$$

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 that $\tilde{S}^{(2)}_{NT}(\beta, f)$ can be written as the trace of a positive definite matrix, and therefore $\tilde{S}^{(2)}_{NT}(\beta, f) \geq 0$. Note also that we can choose $\beta = \beta^0$ and $f = f^0$ in the minimization problem over $S_{NT}(\beta, f)$, *i.e.* the optimal $\beta = \hat{\beta}$ and $f = \hat{f}$ must satisfy $S_{NT}(\hat{\beta}, \hat{f}) \leq S_{NT}(\beta^0, f^0)$. Using this, equation (B.3), $\tilde{S}^{(2)}_{NT}(\beta, f) \geq 0$, and the bound in (B.6) we find

$$0 \geq b \left\|\hat{\beta}^{\text{high}} - \beta^{\text{high}}_0\right\|^2 + o_p\left(\left\|\hat{\beta}^{\text{high}} - \beta^{\text{high}}_0\right\|\right) + o_p\left(\left\|\hat{\beta}^{\text{low}} - \beta^{\text{low}}_0\right\|\right) + o_p(1) . \tag{B.7}$$

Since we assume that $\hat{\beta}^{\text{low}}$ is bounded, the last equation implies that $\left\|\hat{\beta}^{\text{high}} - \beta^{\text{high}}_0\right\| = o_p(1)$, *i.e.* $\hat{\beta}^{\text{high}}$ is consistent. What is left to show is that $\hat{\beta}^{\text{low}}$ is consistent, too. In the supplementary material we show that Assumption 4(ii)(b) guarantees that there exist finite positive constants a_0, a_1, a_2, a_3 and a_4 such that

$$\begin{aligned}
\tilde{S}^{(2)}_{NT}(\beta, f) \geq &\frac{a_0 \left\|\beta^{\text{low}} - \beta^{\text{low}}_0\right\|^2}{\left\|\beta^{\text{low}} - \beta^{\text{low}}_0\right\|^2 + a_1 \left\|\beta^{\text{low}} - \beta^{\text{low}}_0\right\| + a_2} \\
&- a_3 \left\|\beta^{\text{high}} - \beta^{\text{high}}_0\right\| - a_4 \left\|\beta^{\text{high}} - \beta^{\text{high}}_0\right\| \left\|\beta^{\text{low}} - \beta^{\text{low}}_0\right\| , \qquad \text{wpa1.}
\end{aligned} \tag{B.8}$$

Using consistency of $\hat{\beta}^{\text{high}}$ and again boundedness of β^{low} this implies that there exists $a > 0$ such that $\tilde{S}^{(2)}_{NT}(\hat{\beta}, f) \geq a \left\| \hat{\beta}^{\text{low}} - \beta^{\text{low}}_0 \right\|^2 + o_p(1)$. With the same argument as for $\hat{\beta}^{\text{high}}$ we therefore find $\left\| \hat{\beta}^{\text{low}} - \beta^{\text{low}}_0 \right\| = o_p(1)$, *i.e.* $\hat{\beta}^{\text{low}}$ is consistent. This is what we wanted to show. ∎

C Proof of Limiting Distribution of the QMLE (Theorem 3.3)

Theorem 3.1 and Corollary 3.2 are from Moon and Weidner (2010), and the proof can be found there. Note that Assumption 4(i) implies $\|X_k\| = \mathcal{O}_p(\sqrt{NT})$, which is assumed in Moon and Weidner (2010). They also assume $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)}) = \mathcal{O}_p(\sqrt{N})$, while we assume $\|e\| = o_p(\|N^{2/3}\|)$. It is, however, straightforward to verify that their proof for Theorem 3.1 is also valid under this weaker assumption. They also use different consistency assumptions that are demanded in Corollary 3.2, which is not important for the proof of the corollary, since only consistency of the QMLE $\hat{\beta}$, not the detailed assumptions for it, enters into the proof. In the supplementary material we show that the assumptions of Corollary 3.2 already guarantee that W_{NT} does not become singular as $N, T \to \infty$.

To present the proof of Theorem 3.3 it is convenient to first establish some technical lemmas.

Lemma C.1. *Under Assumption 5 we have*

$$\begin{aligned}
\|X_k^{\text{weak}}\| &= \mathcal{O}_p(\sqrt{N}) , && k = 1, \ldots, K , \\
\|P_{\lambda^0} e P_{f^0}\| &= \mathcal{O}_p(1) , && \\
\|P_{\lambda^0} e X_k^{\text{str}\,\prime}\| &= \mathcal{O}_p(\sqrt{NT}) , && k = 1, \ldots, K , \\
\|P_{f^0} e' X_k^{\text{str}}\| &= \mathcal{O}_p(\sqrt{NT}) , && k = 1, \ldots, K .
\end{aligned}$$

Lemma C.2. *Under Assumption 5 we have*

$$(a) \qquad \frac{1}{NT} \mathrm{Tr}(M_{f^0}\, X_{k_1}^{\text{weak}\prime}\, P_{\lambda^0}\, X_{k_2}^{\text{weak}}) = o_p(1) \; ,$$

$$(b) \qquad \frac{1}{NT} \mathrm{Tr}(P_{f^0}\, X_{k_1}^{\text{weak}\prime}\, X_{k_2}^{\text{weak}}) = o_p(1) \; ,$$

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

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

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

$$(f) \qquad \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(P_{f^0}\, e'\, P_{\lambda^0}\, X_k^{\text{weak}}\right) = o_p(1) \; ,$$

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

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

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

$$(j) \qquad \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e'\, P_{\lambda^0}\, X_k^{\text{weak}}\right) = o_p(1) \; ,$$

$$(k) \qquad \frac{1}{\sqrt{NT}} \mathrm{Tr}\left\{\left[ee' - \mathbb{E}\left(ee'\right)\right]\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right\} = o_p(1) \; ,$$

$$(l) \qquad \frac{1}{\sqrt{NT}} \mathrm{Tr}\left\{\left[e'e - \mathbb{E}\left(e'e\right)\right]\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right\} = o_p(1) \; ,$$

$$(m) \qquad \frac{1}{\sqrt{NT}} \mathrm{Tr}\left\{P_{f^0}\, \left[e'\, X_k^{\text{weak}} - \mathbb{E}\left(e'\, X_k^{\text{weak}}\right)\right]\right\} = o_p(1) \; .$$

Lemma C.3. *Under Assumptions 5 and 6 we have*

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

The proofs of the three preceding lemmas are given in the supplementary material. The proof of Lemma C.3 makes use of the cross-sectional independence of (e_{it}, X_{it}^{weak}), and applies Theorem 2 in Phillips and Moon (1999). We now prove the theorem on the limiting distribution of $\hat{\beta}$ in the main text.

Proof of Theorem 3.3. We have $\|e\| = \mathcal{O}_p(N^{1/2})$, *i.e.* Assumption 3* is satisfied. We can therefore apply Corollary 3.2 to calculate the limiting distribution of $\hat{\beta}$. Using Lemma C.2 and

Assumption 6 we find for the matrix in the numerator

$$
\begin{aligned}
W_{NT,k_1k_2} &= \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X'_{k_1}\, M_{\lambda^0}\, X_{k_2}) \\
&= \frac{1}{NT}\mathrm{Tr}(\mathfrak{X}'_{k_1}\mathfrak{X}_{k_2}) - \frac{1}{NT}\mathrm{Tr}(P_{f^0}\, X^{\mathrm{weak}\prime}_{k_1}\, X^{\mathrm{weak}}_{k_2}) - \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X^{\mathrm{weak}\prime}_{k_1}\, P_{\lambda^0}\, X^{\mathrm{weak}}_{k_2}) \\
&= \frac{1}{NT}\mathrm{Tr}(\mathfrak{X}'_{k_1}\mathfrak{X}_{k_2}) + o_p(1)\ . \\
&= W + o_p(1)\ . \qquad \text{(C.1)}
\end{aligned}
$$

Using Lemmas C.2 and C.3 and Assumption 6 we find for the denominator terms

$$
\begin{aligned}
\frac{1}{\sqrt{NT}}C^{(1)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) =& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(M_{f^0}\, e'\, M_{\lambda^0}\, X_k\right) \\
=& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\, X^{\mathrm{weak}}_k\right)\right] \\
&- \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, P_{\lambda^0}\, X^{\mathrm{weak}}_k\right) + \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(P_{f^0}\, e'\, P_{\lambda^0}\, X^{\mathrm{weak}}_k\right) \\
&- \frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{P_{f^0}\,\left[e'\, X^{\mathrm{weak}}_k - \mathbb{E}\left(e'\, X^{\mathrm{weak}}_k\right)\right]\right\} \\
=& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\, X^{\mathrm{weak}}_k\right)\right] + o_p(1)\ . \\
\underset{d}{\rightarrow}& \, \mathcal{N}\left(-\kappa B_1,\, \Omega\right)\ , \qquad \text{(C.2)}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{\sqrt{NT}} C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = &- \frac{1}{\sqrt{NT}} \Big[\mathrm{Tr}\left(e M_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) \\
&\quad + \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) \\
&\quad + \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] \\
= &- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e\, M_{f^0}\, e'\, M_{\lambda^0}\, X_k^{\text{weak}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) \\
&+ \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e P_{f^0}\, e'\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left\{\left[ee' - \mathbb{E}\left(ee'\right)\right]\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right\} \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left[\mathbb{E}\left(ee'\right)\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right] \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e'\, M_{\lambda^0}\, e\, M_{f^0}\, X_k^{\text{weak}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
&+ \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e' P_{\lambda^0}\, e\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left\{\left[e'e - \mathbb{E}\left(e'e\right)\right]\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right\} \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left[\mathbb{E}\left(e'e\right)\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right] \\
&+ \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k^{\text{weak}}\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
&+ \frac{1}{\sqrt{NT}} \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k^{\text{str}}\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
= &- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left[\mathbb{E}\left(ee'\right)\, M_{\lambda^0}\, X_k^{\text{str}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right] \\
&- \frac{1}{\sqrt{NT}} \mathrm{Tr}\left[\mathbb{E}\left(e'e\right)\, M_{f^0}\, X_k^{\text{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right] + o_p(1)\,, \\
= &- \kappa^{-1} B_2 - \kappa B_3 + o_p(1)\,, \qquad \text{(C.3)}
\end{aligned}
$$

Combining these results we obtain

$$
\begin{aligned}
\sqrt{NT}\left(\hat\beta - \beta^0\right) &= W_{NT}^{-1} \left(\frac{1}{\sqrt{NT}} C^{(1)} + \frac{1}{\sqrt{NT}} C^{(1)}\right)\,, \\
&\underset{d}{\rightarrow} \mathcal{N}\left(-\,W^{-1}\,\left(\kappa B_1 + \kappa^{-1} B_2 + \kappa B_3\right),\; W^{-1}\,\Omega\, W^{-1}\right)\,, \qquad \text{(C.4)}
\end{aligned}
$$

which is what we wanted to show. ∎

D Expansions of Projectors and Residuals

The incidental parameter estimators $\hat f$ and $\hat\lambda$ as well as the residuals $\hat e$ enter into the asymptotic bias and variance estimators for the QMLE $\hat\beta$. To describe the properties of $\hat f$, $\hat\lambda$ and $\hat e$, it is convenient to have asymptotic expansions of the projectors $M_{\hat{\lambda}}(\beta)$ and $M_{\hat{f}}(\beta)$ that correspond to the minimizing parameters $\hat{\lambda}(\beta)$ and $\hat{f}(\beta)$ in equation (2.5). Note that the minimizing $\hat{\lambda}(\beta)$ and $\hat{f}(\beta)$ can be defined for all values of β, not only for the optimal value $\beta = \hat{\beta}$. The corresponding residuals are $\hat{e}(\beta) = Y - \sum_{k=1}^{K} \beta_k X_k - \hat{\lambda}(\beta) \hat{f}'(\beta)$.

Theorem D.1. *Under Assumption 1, 3, and 4(i) we have the following expansions*

$$
\begin{aligned}
M_{\hat{\lambda}}(\beta) &= M_{\lambda^0} + M^{(1)}_{\hat{\lambda},e} + M^{(2)}_{\hat{\lambda},e} - \sum_{k=1}^{K} \left(\beta_k - \beta^0_k\right) M^{(1)}_{\hat{\lambda},k} + M^{(\mathrm{rem})}_{\hat{\lambda}}(\beta) \ , \\
M_{\hat{f}}(\beta) &= M_{f^0} + M^{(1)}_{\hat{f},e} + M^{(2)}_{\hat{f},e} - \sum_{k=1}^{K} \left(\beta_k - \beta^0_k\right) M^{(1)}_{\hat{f},k} + M^{(\mathrm{rem})}_{\hat{f}}(\beta) \ , \\
\hat{e}(\beta) &= M_{\lambda^0}\, e\, M_{f^0} + \hat{e}^{(1)}_e - \sum_{k=1}^{K} \left(\beta_k - \beta^0_k\right) \hat{e}^{(1)}_k + \hat{e}^{(\mathrm{rem})}(\beta) \ ,
\end{aligned} \tag{D.1}
$$

where the spectral norms of the remainders satisfy for any series $\eta_{NT} \to 0$

$$
\begin{aligned}
\sup_{\{\beta : \|\beta - \beta^0\| \le \eta_{NT}\}} \frac{\left\| M^{(\mathrm{rem})}_{\hat{\lambda}}(\beta) \right\|}{\|\beta - \beta^0\|^2 + (NT)^{-1/2} \|e\| \, \|\beta - \beta^0\| \ + (NT)^{-3/2} \|e\|^3} &= \mathcal{O}_p(1) \ , \\
\sup_{\{\beta : \|\beta - \beta^0\| \le \eta_{NT}\}} \frac{\left\| M^{(\mathrm{rem})}_{\hat{f}}(\beta) \right\|}{\|\beta - \beta^0\|^2 + (NT)^{-1/2} \|e\| \, \|\beta - \beta^0\| \ + (NT)^{-3/2} \|e\|^3} &= \mathcal{O}_p(1) \ , \\
\sup_{\{\beta : \|\beta - \beta^0\| \le \eta_{NT}\}} \frac{\left\| \hat{e}^{(\mathrm{rem})}(\beta) \right\|}{(NT)^{1/2} \|\beta - \beta^0\|^2 + \|e\| \, \|\beta - \beta^0\| + (NT)^{-1} \|e\|^3} &= \mathcal{O}_p(1) \ ,
\end{aligned} \tag{D.2}
$$

and we have $\mathrm{rank}(\hat{e}^{(\mathrm{rem})}(\beta)) \le 7R$, *and the expansion coefficients are given by*

$$
\begin{aligned}
M^{(1)}_{\hat{\lambda},e} &= - M_{\lambda^0}\, e\, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \lambda^{0\prime} \ - \ \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, e' \, M_{\lambda^0} \ , \\
M^{(1)}_{\hat{\lambda},k} &= - M_{\lambda^0}\, X_k \, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \lambda^{0\prime} \ - \ \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, X_k' \, M_{\lambda^0} \ , \\
M^{(2)}_{\hat{\lambda},e} &= M_{\lambda^0}\, e\, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \lambda^{0\prime} \, e\, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \lambda^{0\prime} \\
&\quad + \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, e' \, \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, e' \, M_{\lambda^0} \\
&\quad - M_{\lambda^0}\, e\, M_{f^0}\, e' \, \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \, \lambda^{0\prime} \\
&\quad - \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \, \lambda^{0\prime} \, e\, M_{f^0}\, e' \, M_{\lambda^0} \\
&\quad - M_{\lambda^0}\, e\, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, e' \, M_{\lambda^0} \\
&\quad + \lambda^0 \, (\lambda^{0\prime} \lambda^0)^{-1} \, (f^{0\prime} f^0)^{-1} \, f^{0\prime} \, e' \, M_{\lambda^0}\, e\, f^0 \, (f^{0\prime} f^0)^{-1} \, (\lambda^{0\prime} \lambda^0)^{-1} \lambda^{0\prime} \, ,
\end{aligned} \tag{D.3}
$$

analogously

$$
\begin{aligned}
M^{(1)}_{\hat f,e} &= -\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1} f^{0\prime} \,-\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, M_{f^0}\ , \\
M^{(1)}_{\hat f,k} &= -\, M_{f^0}\, X_k'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1} f^{0\prime} \,-\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, X_k\, M_{f^0}\ , \\
M^{(2)}_{\hat f,e} &= M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1} f^{0\prime}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1} f^{0\prime} \\
&\quad + f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, M_{f^0} \\
&\quad - M_{f^0}\, e'\, M_{\lambda^0}\, e\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime} \\
&\quad - f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime}\, e'\, M_{\lambda^0}\, e\, M_{f^0} \\
&\quad - M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, M_{f^0} \\
&\quad + f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1} f^{0\prime}\ ,
\end{aligned}
\tag{D.4}
$$

and finally

$$
\begin{aligned}
\hat e^{(1)}_k &= M_{\lambda^0}\, X_k\, M_{f^0}\ , \\
\hat e^{(1)}_e &= -M_{\lambda^0}\, e\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime} \\
&\quad - \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime}\, e'\, M_{\lambda^0}\, e\, M_{f^0} \\
&\quad - M_{\lambda^0}\, e\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\, e\, M_{f^0}\ .
\end{aligned}
\tag{D.5}
$$

Proof. The general expansion of $M_{\hat\lambda}(\beta)$ is given Moon and Weidner (2010), and in the theorem we just make this expansion explicit up to a particular order. The result for $M_{\hat f}(\beta)$ is just obtained by symmetry ($N \leftrightarrow T$, $\lambda \leftrightarrow f$, $e \leftrightarrow e'$, $X_k \leftrightarrow X_k'$). For the residuals $\hat e$ we have

$$
\hat e = M_{\hat\lambda}\left(Y - \sum_{k=1} \hat\beta_k\, X_k\right) = M_{\hat\lambda}\left[e - \sum_{k=1}^{K}\left(\hat\beta_k - \beta_k^0\right) X_k + \lambda^0 f^{0\prime}\right] , \tag{D.6}
$$

and plugging in the expansion of $M_{\hat\lambda}$ gives the expansion of $\hat e$. We have $\hat e(\beta) = A_0 + \lambda^0 f^{0\prime} - \hat\lambda(\beta)\hat f'(\beta)$, where $A_0 = e - \sum_k(\beta_k - \beta_k^0)X_k$. Therefore $\hat e^{(\mathrm{rem})}(\beta) = A_1 + A_2 + A_3$ with $A_1 = A_0 - M_{\lambda^0}\, A_0\, M_{f^0}$, $A_2 = \lambda^0 f^{0\prime} - \hat\lambda(\beta)\hat f'(\beta)$, and $A_3 = -\hat e^{(1)}_e$. We find $\mathrm{rank}(A_1) \le 2R$, $\mathrm{rank}(A_2) \le 2R$, $\mathrm{rank}(A_3) \le 3R$, and thus $\mathrm{rank}(\hat e^{(\mathrm{rem})}(\beta)) \le 7R$, as stated in the theorem. ∎

Having expansions for $M_{\hat\lambda}(\beta)$ and $M_{\hat f}(\beta)$ we also have expansions for $P_{\hat\lambda}(\beta) = \mathbb{I}_N - M_{\hat\lambda}(\beta)$ and $P_{\hat f}(\beta) = \mathbb{I}_T - M_{\hat f}(\beta)$. The reason why we give expansions of the projectors and not expansions of $\hat\lambda(\beta)$ and $\hat f(\beta)$ directly is that for the latter we would need to specify a normalization, while the projectors are independent of any normalization choice. An expansion for $\hat\lambda(\beta)$ can for example be defined by $\hat\lambda(\beta) = P_{\hat\lambda}(\beta)\lambda^0$, in which case the normalization of $\hat\lambda(\beta)$ is implicitly defined by the normalization of λ^0.

E Consistency Proof for Bias and Variance Estimators (Theorem 3.6)

Corollary E.1. *Under Assumptions 5 and 6 we have* $\sqrt{NT}\left(\hat\beta - \beta^0\right) = \mathcal{O}_p(1)$.

This corollary directly follows from Theorem 3.3.

Corollary E.2. *Under Assumption 5 we have*

$$\left\|P_{\hat\lambda} - P_{\lambda^0}\right\| = \left\|M_{\hat\lambda} - M_{\lambda^0}\right\| = \mathcal{O}_p(N^{-1/2}) ,$$
$$\left\|P_{\hat f} - P_{f^0}\right\| = \left\|M_{\hat f} - M_{f^0}\right\| = \mathcal{O}_p(T^{-1/2}) .$$

Proof. Using $\|e\| = \mathcal{O}_p(N^{1/2})$ and $\|X_k\| = \mathcal{O}_p(N)$ we find that the expansion terms in Theorem D.1 satisfy

$$\left\|M^{(1)}_{\hat\lambda,e}\right\| = \mathcal{O}_p(N^{-1/2}) , \qquad \left\|M^{(2)}_{\hat\lambda,e}\right\| = \mathcal{O}_p(N^{-1}) , \qquad \left\|M^{(1)}_{\hat\lambda,k}\right\| = \mathcal{O}_p(1) . \tag{E.1}$$

Together with corollary E.1 the result for $\left\|M_{\hat\lambda} - M_{\lambda^0}\right\|$ immediately follows. In addition we have $P_{\hat\lambda} - P_{\lambda^0} = -M_{\hat\lambda} + M_{\lambda^0}$. The proof for $M_{\hat f}$ and $P_{\hat f}$ is analogous. ∎

Lemma E.3. *Under Assumption 5 we have*

$$A_1 \equiv \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T e_{it}^2 \left(\mathfrak{X}_{it}\mathfrak{X}_{it}' - \hat{\mathfrak{X}}_{it}\hat{\mathfrak{X}}_{it}'\right) = o_p(1) , \quad A_2 \equiv \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(e_{it}^2 - \hat e_{it}^2\right)\hat{\mathfrak{X}}_{it}\hat{\mathfrak{X}}_{it}' = o_p(1) .$$

Lemma E.4. *Let $\hat f$ and f^0 be normalized as $\hat f'\hat f/T = \mathbb{I}_R$ and $f^{0\prime}f^0/T = \mathbb{I}_R$. Then, under the assumptions of Theorem 3.6, there exists an $R \times R$ matrices $H = H_{NT}$ such that*[17]

$$\left\|\hat f - f^0\, H\right\| = O_p\left(1\right) , \qquad \left\|\hat\lambda - \lambda^0\,\left(H'\right)^{-1}\right\| = O_p\left(1\right) .$$

Furthermore

$$\left\|\hat\lambda\,(\hat\lambda'\hat\lambda)^{-1}\,(\hat f'\hat f)^{-1}\,\hat f' - \lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\right\| = O_p\left(N^{-3/2}\right) .$$

Lemma E.5. *Under Assumption 5 we have*

$$\text{(i)} \qquad N^{-1}\left\|\mathbb{E}(e'M_{\lambda^0}X_k) - \left(\hat e'\,X_k\right)^{\text{truncR}}\right\| = o_p(1) ,$$
$$\text{(ii)} \qquad N^{-1}\left\|\mathbb{E}(e'M_{\lambda^0}e) - \left(\hat e'\,\hat e\right)^{\text{truncD}}\right\| = o_p(1) ,$$
$$\text{(iii)} \qquad T^{-1}\left\|\mathbb{E}(eM_{f^0}e') - \left(\hat e\,\hat e'\right)^{\text{truncD}}\right\| = o_p(1) .$$

Lemma E.6. *Under the Assumption 5 we have*

$$\text{(i)} \qquad N^{-1}\left\|\left(\hat e'\,X_k\right)^{\text{truncR}}\right\| = \mathcal{O}_p(MT^{1/8}) ,$$
$$\text{(ii)} \qquad N^{-1}\left\|\left(\hat e'\,\hat e\right)^{\text{truncD}}\right\| = \mathcal{O}_p(1) ,$$
$$\text{(iii)} \qquad T^{-1}\left\|\left(\hat e\,\hat e'\right)^{\text{truncD}}\right\| = \mathcal{O}_p(1) .$$

The proof of the above lemmas is given in the supplementary material. Using these lemmas we can now prove Theorem 3.6.

[17] We consider a limit $N, T \to \infty$ and for different N, T different H-matrices can be chosen, but we write H instead of H_{NT} to keep notation simple.

Proof of Theorem 3.6, Part I: show $\hat{W} = W + o_p(1)$.
According to Assumption 6 we have $W_{k_1k_2} = W_{NT,k_1k_2} + o_p(1)$, where

$$
\begin{aligned}
W_{NT,k_1k_2} &= (NT)^{-1}\mathrm{Tr}(\mathfrak{X}_{k_1}\mathfrak{X}'_{k_2}) \\
&= (NT)^{-1}\mathrm{Tr}(M_{\lambda^0}\, X^{\mathrm{str}}_{k_1}\, M_{f^0}\, X^{\mathrm{str}\,\prime}_{k_2}) + (NT)^{-1}\mathrm{Tr}(M_{\lambda^0}\, X^{\mathrm{str}}_{k_1}\, M_{f^0}\, X^{\mathrm{weak}\,\prime}_{k_2}) \\
&\quad + (NT)^{-1}\mathrm{Tr}(M_{\lambda^0}\, X^{\mathrm{weak}}_{k_1}\, M_{f^0}\, X^{\mathrm{str}\,\prime}_{k_2}) + (NT)^{-1}\mathrm{Tr}(X^{\mathrm{weak}}_{k_1} X^{\mathrm{weak}\,\prime}_{k_2}) \; .
\end{aligned} \tag{E.2}
$$

In order to prove $\hat{W} = W + o_p(1)$, it is therefore sufficient to show $\hat{W}_{k_1k_2} = W_{NT,k_1k_2} + o_p(1)$, where $\hat{W}_{k_1k_2} = (NT)^{-1}\mathrm{Tr}(\hat{\mathfrak{X}}_{k_1}\hat{\mathfrak{X}}'_{k_2}) = (NT)^{-1}\mathrm{Tr}(M_{\hat{\lambda}}\, X_{k_1}\, M_{\hat{f}}\, X'_{k_2})$. Using $|\mathrm{Tr}\,(C)| \leq \|C\|\,\mathrm{rank}\,(C)$ corollary E.2, and the result $\|X^{\mathrm{weak}}_{k_1}\| = \mathcal{O}_p(N^{-1})$ from Lemma C.1, we find

$$
\begin{aligned}
&\left|\hat{W}_{k_1k_2} - W_{NT,k_1k_2}\right| \\
&\quad = \Big|(NT)^{-1}\mathrm{Tr}\left[\left(M_{\hat{\lambda}} - M_{\lambda^0}\right)\, X_{k_1}\, M_{\hat{f}}\, X'_{k_2}\right] + (NT)^{-1}\mathrm{Tr}\left[M_{\lambda^0}\, X_{k_1}\, \left(M_{\hat{f}} - M_{f^0}\right)\, X'_{k_2}\right] \\
&\qquad - (NT)^{-1}\mathrm{Tr}(M_{\lambda^0} X^{\mathrm{weak}}_{k_1} P_{f^0} X^{\mathrm{weak}\,\prime}_{k_2}) - (NT)^{-1}\mathrm{Tr}(P_{\lambda^0} X^{\mathrm{weak}}_{k_1} X^{\mathrm{weak}\,\prime}_{k_2})\Big| \\
&\quad \leq \frac{2R}{NT}\left\|M_{\hat{\lambda}} - M_{\lambda^0}\right\| \|X_{k_1}\|\|X_{k_2}\| \frac{2R}{NT}\left\|M_{\hat{f}} - M_{f^0}\right\| \|X_{k_1}\|\|X_{k_2}\| \\
&\qquad + \frac{R}{NT}\|X^{\mathrm{weak}}_{k_1}\|\|X^{\mathrm{weak}}_{k_2}\| + \frac{R}{NT}\|X^{\mathrm{weak}}_{k_1}\|\|X^{\mathrm{weak}}_{k_2}\| \\
&\quad = \frac{2R}{NT}\mathcal{O}_p(N^{-1})\mathcal{O}_p(NT) + \frac{2R}{NT}\mathcal{O}_p(T^{-1})\mathcal{O}_p(NT) + \frac{2R}{NT}\mathcal{O}_p(N) \\
&\quad = o_p(1) \; .
\end{aligned} \tag{E.3}
$$

This is what we wanted to show. ∎

Proof of Theorem 3.6, Part II: show $\hat{\Omega} = \Omega + o_p(1)$.
Let $\Omega_{NT} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 \mathfrak{X}_{it}\mathfrak{X}'_{it}$. First, we want to show that $\Omega = \Omega_{NT} + o_p(1)$. By definition of Ω we have $\Omega = \mathbb{E}\,(\Omega_{NT}) + o(1)$. Thus, once we show for all $k_1, k_2 = 1, \ldots, K$ that $\mathrm{Var}\,(\Omega_{NT,k_1k_2}) = o(1)$ we are done.

Using cross-sectional independence of e_{it} and X^{weak}_{it}, we find that conditional on X^{str}_{k} (or alternatively, treating $X_k^{\rm str}$ as non-stochastic) we have

$$\begin{aligned}
\mathrm{Var}\left(\Omega_{NT,k_1k_2}\right) &= \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,\tau=1}^{T}\bigg[\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\, e_{j\tau}^2\mathfrak{X}_{k_1,j\tau}\mathfrak{X}_{k_2,j\tau}\right) \\
&\qquad - \mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\mathbb{E}\left(e_{j\tau}^2\mathfrak{X}_{k_1,j\tau}\mathfrak{X}_{k_2,j\tau}\right)\bigg] \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\bigg[\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\, e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) \\
&\qquad - \mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\mathbb{E}\left(e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right)\bigg] \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\left\{\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\, e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) - \left[\sum_{t=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\right]^2\right\} \\
&\leq \frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\, e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) \\
&\leq \frac{1}{N}\sqrt{\frac{1}{N\,T^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^4 e_{i\tau}^4\right)\frac{1}{N\,T^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(\mathfrak{X}_{k_1,it}^2\mathfrak{X}_{k_2,it}^2\,\mathfrak{X}_{k_1,i\tau}^2\mathfrak{X}_{k_2,i\tau}^2\right)} \\
&= \frac{1}{N}\,\mathcal{O}(1) = o(1)\,, \qquad\qquad (\mathrm{E.4})
\end{aligned}$$

where we used that both e and X_k have uniformly bounded 8'th moments. Since the conditional variance of Ω_{NT,k_1k_2} is $o(1)$, the same is true for the unconditional variance by the law of iterated expectations, so we have shown $\Omega = \Omega_{NT} + o_p(1)$.

We have $\Omega_{NT} - \hat{\Omega} = A_1 + A_2$, where A_1 and A_2 are defined in Lemma E.3 and the lemmas states that both A_1 and A_2 are $o_p(1)$. Therefore we have $\Omega_{NT} = \hat{\Omega} + o_p(1)$, and thus also $\hat{\Omega} = \Omega + o_p(1)$, which is what we wanted to show. ∎

Proof of Theorem 3.6, Part III: show $\hat{B}_1 = B_1 + o_p(1)$.
Let $B_{1,k,NT} = N^{-1}\,\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,X_k^{\rm weak}\right)\right]$, and $\tilde{B}_{1,k,NT} = N^{-1}\,\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,M_{\lambda^0}\,X_k^{\rm weak}\right)\right]$. According to Assumption 6 we have $B_{1,k} = B_{1,k,NT} + o_p(1)$. Applying part (f) of Lemma C.2 we obtain $B_{1,k,NT} = \tilde{B}_{1,k,NT} + o_p(1)$. So what is left to show is that $\tilde{B}_{1,k,NT} = \hat{B}_{1,k} + o_p(1)$. Using $|\mathrm{Tr}\,(C)| \leq \|C\|\,\mathrm{rank}\,(C)$ we find

$$\begin{aligned}
\left|\tilde{B}_{1,k,NT} - \hat{B}_1\right| &= \left|\mathbb{E}\left[\frac{1}{N}\mathrm{Tr}(P_{f^0}\,e'\,M_{\lambda^0}\,X_k)\right] - \frac{1}{N}\mathrm{Tr}\left[P_{\hat{f}}\left(\hat{e}'X_k\right)^{\rm truncR}\right]\right| \\
&\leq \left|\frac{1}{N}\mathrm{Tr}\left[\left(P_{f^0} - P_{\hat{f}}\right)\left(\hat{e}'X_k\right)^{\rm truncR}\right]\right| \\
&\quad + \left|\frac{1}{N}\mathrm{Tr}\left\{P_{f^0}\left[\mathbb{E}\left(e'\,M_{\lambda^0}\,X_k\right) - \left(\hat{e}'X_k\right)^{\rm truncR}\right]\right\}\right| \\
&\leq \frac{2R}{N}\left\|P_{f^0} - P_{\hat{f}}\right\|\left\|\left(\hat{e}'X_k\right)^{\rm truncR}\right\| \\
&\quad + \frac{R}{N}\left\|P_{f^0}\right\|\left\|\mathbb{E}\left(e'\,M_{\lambda^0}\,X_k\right) - \left(\hat{e}'X_k\right)^{\rm truncR}\right\|\,. \qquad\qquad (\mathrm{E.5})
\end{aligned}$$

We have $\left\|P_{f^0}\right\| = 1$. We now apply Lemmas E.5, E.2 and E.6 to find

$$\left|\tilde{B}_{1,k,NT} - \hat{B}_1\right| = N^{-1}\left(\mathcal{O}_p(N^{-1/2})\mathcal{O}_p(MNT^{1/8}) + o_p(N)\right) = o_p(1) \ . \tag{E.6}$$

This is what we wanted to show. ▮

Proof of Theorem 3.6, final part: show $\hat{B}_i = B_i + o_p(1)$, $i = 2, 3$.
Define

$$\begin{aligned}
B_{2,k,NT} &= \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(ee'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right] \ , \\
\tilde{B}_{2,k,NT} &= \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(eM_{f^0}e'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right] \ .
\end{aligned} \tag{E.7}$$

According to Assumption 6 we have $B_{2,k} = B_{2,k,NT} + o_p(1)$. Applying part (g) of Lemma C.2 we obtain $B_{2,k,NT} = \tilde{B}_{2,k,NT} + o_p(1)$. What is left to show is that $\tilde{B}_{2,k,NT} = \hat{B}_{2,k} + o_p(1)$.

We can decompose $\hat{B}_2 = \hat{B}_2^{\mathrm{str}} + \hat{B}_2^{\mathrm{weak}}$, according to the decomposition of the regressors into predetermined and strictly exogenous part. As a consequence of Lemma C.1, *i.e.* $\|X_k^{\mathrm{weak}}\| = \mathcal{O}_p(N^{1/2})$, of part (ii) of Lemma E.6, and of Lemma E.4, we find that the predetermined part of the regressors does not contribute to $\hat{B}_2$ asymptotically, namely

$$\begin{aligned}
\hat{B}_{2,k}^{\mathrm{weak}} &= \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{weak}}\, \hat{f}\,(\hat{f}'\hat{f})^{-1}\,(\hat{\lambda}'\hat{\lambda})^{-1}\,\hat{\lambda}'\right] \\
&\leq \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|X_k^{\mathrm{weak}}\right\|\left\|\hat{f}\,(\hat{f}'\hat{f})^{-1}\,(\hat{\lambda}'\hat{\lambda})^{-1}\,\hat{\lambda}'\right\| \\
&= \frac{R}{T}\,\mathcal{O}_p(T)\,\mathcal{O}_p(N^{1/2})\,\mathcal{O}_p((NT)^{-1/2}) = o_p(1) \ .
\end{aligned} \tag{E.8}$$

We are left to consider the contribution from the strictly exogenous part of the regressor in $\hat{B}_2$. We have

$$\begin{aligned}
\tilde{B}_{2,k} - \hat{B}_{2,k}^{\mathrm{str}} =& \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(eM_{f^0}e'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right] \\
& - \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{str}}\, \hat{f}\,(\hat{f}'\hat{f})^{-1}\,(\hat{\lambda}'\hat{\lambda})^{-1}\,\hat{\lambda}'\right] \\
=& \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{str}}\,\left(f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime} - \hat{f}\,(\hat{f}'\hat{f})^{-1}\,(\hat{\lambda}'\hat{\lambda})^{-1}\,\hat{\lambda}'\right)\right] \\
& + \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\,\left(M_{\lambda^0} - M_{\hat{\lambda}}\right)\, X_k^{\mathrm{str}}\, f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right] \\
& + \frac{1}{T}\mathrm{Tr}\left\{\left[\mathbb{E}\left(eM_{f^0}e'\right) - \left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right]\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right\} \ .
\end{aligned} \tag{E.9}$$

Using $|\mathrm{Tr}\,(C)| \leq \|C\|\,\mathrm{rank}\,(C)$ (which is true for every square matrix C, see the supplementary material) we find

$$\begin{aligned}
\left|\tilde{B}_{2,k} - \hat{B}_{2,k}^{\mathrm{str}}\right| \leq& \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|X_k^{\mathrm{str}}\right\|\left\|f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime} - \hat{f}\,(\hat{f}'\hat{f})^{-1}\,(\hat{\lambda}'\hat{\lambda})^{-1}\,\hat{\lambda}'\right\| \\
& + \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|M_{\lambda^0} - M_{\hat{\lambda}}\right\|\left\|X_k^{\mathrm{str}}\right\|\left\|f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right\| \\
& + \frac{R}{T}\left\|\mathbb{E}\left(eM_{f^0}e'\right) - \left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\,\left\|X_k^{\mathrm{str}}\right\|\,\left\|f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right\| \ .
\end{aligned} \tag{E.10}$$

Here we used $\|M_{f^0}\| = \left\|M_{\hat{f}}\right\| = 1$. Using $\left\|X_k^{\rm str}\right\| = \mathcal{O}_p(\sqrt{NT})$, and applying Lemmas E.2, E.4, E.5, and E.6, we now find

$$\begin{aligned}\left|\tilde{B}_{2,k} - \hat{B}^{\rm str}_{2,k}\right| = T^{-1}\Big[&\mathcal{O}_p(T)\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p(N^{-3/2})\\ &+ \mathcal{O}_p(T)\,\mathcal{O}_p(N^{-1/2})\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p((NT)^{-1/2})\\ &+ o_p(T)\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p((NT)^{-1/2})\Big] = o_p(1)\,. \qquad \text{(E.11)}\end{aligned}$$

This is what we wanted to show. The proof of $\hat{B}_3 = B_3 + o_p(1)$ is analogous. ∎

F Proofs for Section 4 (Testing)

Proof of Theorem 4.1. Using the expansion for $L_{NT}(\beta)$ in Lemma A.1 of Moon and Weidner (2010) we find for the derivative (the sign convention $\epsilon_k = \beta_k^0 - \beta_k$ results in the minus sign below)

$$\begin{aligned}\frac{\partial L_{NT}}{\partial \beta_k} &= -\frac{1}{NT}\sum_{g=2}^{\infty} g \sum_{\kappa_1=0}^{K}\sum_{\kappa_2=0}^{K}\ldots\sum_{\kappa_{g-1}=0}^{K} \epsilon_{\kappa_1}\,\epsilon_{\kappa_2}\,\ldots\,\epsilon_{\kappa_{g-1}}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_{\kappa_1},\ldots,X_{\kappa_{g-1}}\right)\\ &= \left[2W_{NT}(\beta-\beta^0)\right]_k - \frac{2}{\sqrt{NT}}C_{NT,k} + \frac{1}{NT}\nabla R_{1,NT,k} + \frac{1}{NT}\nabla R_{2,NT,k}\,, \qquad \text{(F.1)}\end{aligned}$$

where

$$\begin{aligned}W_{NT,k_1k_2} &= \frac{1}{NT}\,L^{(2)}\left(\lambda^0,\,f^0,\,X_{k_1},\,X_{k_2}\right)\,,\\ C_{NT,k} &= \frac{1}{2\sqrt{NT}}\sum_{g=2}^{G_e} g\,(\epsilon_0)^{g-1}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_0,\ldots,X_0\right)\\ &= \sum_{g=2}^{G_e}\frac{g}{2\sqrt{NT}}\;L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,e,\ldots,e\right)\,, \qquad \text{(F.2)}\end{aligned}$$

and

$$\begin{aligned}\nabla R_{1,NT,k} &= -\sum_{g=G_e+1}^{\infty} g\,(\epsilon_0)^{g-1}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_0,\ldots,X_0\right)\,,\\ &= -\sum_{g=G_e+1}^{\infty} g\;L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,e,\ldots,e\right)\,,\\ \nabla R_{2,NT,k} &= -\sum_{g=3}^{\infty} g\sum_{r=1}^{g-1}\binom{g-1}{r}\sum_{k_1=1}^{K}\ldots\sum_{k_r=1}^{K}\epsilon_{k_1}\,\ldots\,\epsilon_{k_r}\,(\epsilon_0)^{g-r-1}\\ &\qquad\qquad L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_{k_1},\ldots,X_{k_r},X_0,\ldots,X_0\right)\,.\\ &= -\sum_{g=3}^{\infty} g\sum_{r=1}^{g-1}\binom{g-1}{r}\sum_{k_1=1}^{K}\ldots\sum_{k_r=1}^{K}(\beta^0_{k_1}-\beta_{k_1})\,\ldots\,(\beta^0_{k_r}-\beta_{k_r})\\ &\qquad\qquad L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_{k_1},\ldots,X_{k_r},e,\ldots,e\right)\,. \qquad \text{(F.3)}\end{aligned}$$

The above expressions for W_{NT} and C_{NT} are equivalent to their definitions given in theorem 3.1. Using the bound on $L^{(g)}$ we find[18]

$$
\begin{aligned}
|\nabla R_{1,NT,k}| &\leq c_0\, NT\, \frac{\|X_k\|}{\sqrt{NT}} \sum_{g=G_e+1}^{\infty} g^2 \left(\frac{c_1\|e\|}{\sqrt{NT}}\right)^{g-1} \\
&\leq 2\, c_0\, (1+G_e)^2\, NT\, \frac{\|X_k\|}{\sqrt{NT}} \left(\frac{c_1\|e\|}{\sqrt{NT}}\right)^{G_e} \left[1-\left(\frac{c_1\|e\|}{\sqrt{NT}}\right)\right]^{-3} = o_p(\sqrt{NT})\,, \\
|\nabla R_{2,NT,k}| &\leq c_0\, NT\, \frac{\|X_k\|}{\sqrt{NT}} \sum_{g=3}^{\infty} g^2 \sum_{r=1}^{g-1} \binom{g-1}{r} c_1^{g-1} \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right) \\
&\qquad\qquad \times \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right)^{g-2} \\
&\leq c_0\, NT\, \frac{\|X_k\|}{\sqrt{NT}} \sum_{g=3}^{\infty} g^3\, (4c_1)^{g-1} \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right) \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right)^{g-2} \\
&\leq c_2\, NT\, \frac{\|X_k\|}{\sqrt{NT}} \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right) \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right)\,, \qquad \text{(F.4)}
\end{aligned}
$$

where $c_0 = 8Rd_{\max}(\lambda^0, f^0)/2$ and $c_1 = 16 d_{\max}(\lambda^0, f^0)/d^2_{\min}(\lambda^0, f^0)$ both converge to a constants as $N, T \to \infty$, and the very last inequality is only true if $4c_1 \left(\sum_{\tilde{k}=1}^{K} |\beta_{\tilde{k}} - \beta_{\tilde{k}}^0| \frac{\|X_{\tilde{k}}\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right) < 1$, and $c_2 > 0$ is an appropriate positive constant. To show $\nabla R_{1,NT,k} = o_p(NT)$ we used Assumption 3*. From the above inequalities we find for $\eta_{NT} \to \infty$

$$
\begin{aligned}
\sup_{\{\beta: \|\beta-\beta^0\| \leq \eta_{NT}\}} \frac{\|\nabla R_{1,NT}(\beta)\|}{\sqrt{NT}} &= o_p(1)\,, \\
\sup_{\{\beta: \|\beta-\beta^0\| \leq \eta_{NT}\}} \frac{\|\nabla R_{2,NT}(\beta)\|}{NT\, \|\beta-\beta^0\|} &= o_p(1)\,. \qquad \text{(F.5)}
\end{aligned}
$$

Thus $R_{NT}(\beta) = R_{1,NT}(\beta) + R_{2,NT}(\beta)$ satisfies the bound in the theorem. ∎

Proof of Theorem 4.2. Using Theorem 3.3 it is straightforward to show that WD^*_{NT} has limiting distribution χ^2_r.

For the LR test we have to show that the estimator $\hat{c} = (NT)^{-1} \mathrm{Tr}(\hat{e}(\hat{\beta})\, \hat{e}'(\hat{\beta}))$ is consistent for $c = \mathbb{E}e^2_{it}$. As already noted in the main text we have $\hat{c} = L_{NT}\left(\hat{\beta}\right)$, and using our likelihood expansion and $\sqrt{NT}$-consistency of $\hat{\beta}$ we immediately obtain

$$
\hat{c} = \frac{1}{NT} \mathrm{Tr}(M_{\lambda^0} e M_{f^0} e') + o_p(1)\,. \qquad \text{(F.6)}
$$

Alternatively, one could use the expansion of $\hat{e}$ in Theorem D.1 to show this. From the above

[18] Here we use $\binom{n}{k} \leq 4^n$.

result we find

$$\left|\hat{c} - \frac{1}{NT}\text{Tr}(ee')\right| = \frac{1}{NT}\left|\text{Tr}(P_{\lambda^0} e M_{f^0} e') + \text{Tr}(e P_{f^0} e')\right| + o_p(1)$$
$$\leq \frac{2R}{NT}\|e\|^2 + o_p(1) = o_p(1) . \tag{F.7}$$

By the weak law of large numbers we thus have

$$\hat{c} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 + o_p(1) = c + o_p(1) , \tag{F.8}$$

i.e. $\hat{c}$ is indeed consistent for c. Having this one immediately obtains the result for the limiting distribution of LR^*_{NT}.

For the LM test we first want to show that equation (4.9) holds. Using the expansion of $\hat{e}$ in Theorem D.1 one obtains

$$\begin{aligned}
\sqrt{NT}(\tilde{\nabla}\mathcal{L}_{NT})_k &= -\frac{2}{\sqrt{NT}}\,\text{Tr}\left(X_k'\tilde{e}\right) \\
&= \left[2\sqrt{NT}\,W_{NT}\,\left(\tilde{\beta} - \beta^0\right)\right]_k + \frac{2}{NT}C^{(1)}(\lambda^0, f^0, X_k, e) + \frac{2}{NT}C^{(2)}(\lambda^0, f^0, X_k, e) \\
&\qquad - \frac{2}{\sqrt{NT}}\,\text{Tr}\left(X_k'\tilde{e}^{(\text{rem})}\right) \\
&= \left[2\sqrt{NT}\,W_{NT}\,\left(\tilde{\beta} - \beta^0\right) + \frac{2}{NT}C_{NT}\right]_k + o_p(1) \\
&= \sqrt{NT}\left[\nabla L_{NT}(\tilde{\beta})\right]_k + o_p(1) ,
\end{aligned} \tag{F.9}$$

which is what we wanted to show. Here we used that $\left|\text{Tr}\left(X_k'\tilde{e}^{(\text{rem})}\right)\right| \leq 7R\|X_k\|\|\tilde{e}^{(\text{rem})}\| = \mathcal{O}_p(N^{3/2})$. Note that $\|X_k\| = \mathcal{O}_p(N)$, and Theorem D.1 and $\sqrt{NT}$-consistency of $\tilde{\beta}$ imply $\|\tilde{e}^{(\text{rem})}\| = \mathcal{O}_p(\sqrt{N})$. We also used the expression for $\nabla L_{NT}(\beta)$ given in Theorem 4.1, and the bound on $\nabla R_{NT}(\beta)$ given there.

We now use equation (4.11) and $\tilde{W} = W + o_p(1)$, $\tilde{\Omega} = \Omega + o_p(1)$, and $\tilde{B} = B + o_p(1)$ to obtain

$$LM^*_{NT} \underset{d}{\longrightarrow} (C - B)'W^{-1}H'(HW^{-1}\Omega W^{-1}H')^{-1}HW^{-1}(C - B) . \tag{F.10}$$

Under H_0 we thus find $LM^*_{NT} \to_d \chi^2_r$. ∎

References

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

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

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.

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

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

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

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

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

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

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

Daniel, K. and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *The Journal of Finance*, 52(1):1–33.

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

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

Hahn, J. and Kuersteiner, G. (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. and Kuersteiner, G. (2004). Bias reduction for dynamic nonlinear panel models with fixed effects. Manuscript.

Hahn, J. and Moon, H. R. (2006). Reducing bias of MLE in a dynamic panel model. *Econometric Theory*, 22(03):499–512.

Hahn, J. and Newey, W. (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. Manuscript.

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

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.

Moon, H. and Weidner, M. (2010). Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects. Manuscript.

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

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

Onatski, A. (2005). Determining the number of factors from empirical distribution of eigenvalues. Discussion Papers 0405-19, Columbia University, Department of Economics.

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

Phillips, P. C. B. and Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111.

Phillips, P. C. B. and Sul, D. (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. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179.

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

Tables with Simulation Results

		$\rho = 0.3$			$\rho = 0.6$		
		OLS	QMLE	BC-QMLE	OLS	QMLE	BC-QMLE
$T = 5$, $M = 2$	bias	0.1257	-0.1470	-0.0640	0.0807	-0.2080	-0.1169
	std	0.1462	0.1435	0.0907	0.1195	0.1789	0.1253
	rmse	0.1929	0.2054	0.1110	0.1442	0.2743	0.1713
$T = 10$, $M = 3$	bias	0.1337	-0.0549	-0.0175	0.0918	-0.0596	-0.0236
	std	0.1125	0.0577	0.0404	0.0896	0.0679	0.0458
	rmse	0.1748	0.0796	0.0441	0.1283	0.0903	0.0515
$T = 20$, $M = 4$	bias	0.1443	-0.0261	-0.0057	0.1015	-0.0253	-0.0070
	std	0.0875	0.0278	0.0236	0.0691	0.0280	0.0216
	rmse	0.1688	0.0381	0.0242	0.1228	0.0378	0.0227
$T = 40$, $M = 5$	bias	0.1511	-0.0129	-0.0018	0.1083	-0.0114	-0.0017
	std	0.0653	0.0167	0.0158	0.0514	0.0154	0.0138
	rmse	0.1646	0.0211	0.0159	0.1199	0.0192	0.0139
$T = 80$, $M = 6$	bias	0.1552	-0.0066	-0.0006	0.1125	-0.0057	-0.0006
	std	0.0487	0.0112	0.0110	0.0382	0.0096	0.0092
	rmse	0.1627	0.0130	0.0110	0.1188	0.0112	0.0093

Table 1: Simulation results for the AR(1) model described in the main text with $N = 100$, $\rho_f = 0.5$, and $\sigma_f = 0.5$. The OLS estimator, QMLE, and bias corrected QMLE (BC-QMLE) were computed for 10,000 samples. The table lists the mean bias, the standard deviation (std), and the square root of the mean square error (rmse) for the three estimators.

		$\rho_f = 0.3$			$\rho_f = 0.7$		
		OLS	QMLE	BC-QMLE	OLS	QMLE	BC-QMLE
$\sigma_f = 0$	bias	-0.0007	-0.0108	-0.0059	-0.0007	-0.0108	-0.0059
	std	0.0180	0.0367	0.0256	0.0180	0.0367	0.0256
	rmse	0.0180	0.0383	0.0263	0.0180	0.0383	0.0263
$\sigma_f = 0.2$	bias	0.0156	-0.0131	-0.0037	0.0475	-0.0344	-0.0098
	std	0.0253	0.0294	0.0223	0.0381	0.0352	0.0249
	rmse	0.0297	0.0322	0.0226	0.0609	0.0492	0.0267
$\sigma_f = 0.5$	bias	0.0568	-0.0142	-0.0042	0.1487	-0.0404	-0.0120
	std	0.0622	0.0258	0.0208	0.0767	0.0297	0.0229
	rmse	0.0843	0.0295	0.0212	0.1673	0.0502	0.0259

Table 2: Simulation results for the AR(1) model with $N = 100$, $T = 20$, $M = 4$, and $\rho = 0.6$. The three different estimators were computed for 10,000 samples, and the mean bias, standard deviation (std), and root mean square error (rmse) are reported.

	$M = 2$	$M = 4$	$M = 6$	$M = 9$	$M = 12$	$M = 15$
$\rho = 0$	0.875	0.779	0.710	0.625	0.559	0.512
$\rho = 0.3$	0.754	0.777	0.710	0.622	0.555	0.507
$\rho = 0.6$	0.593	0.731	0.679	0.595	0.529	0.484
$\rho = 0.9$	0.295	0.480	0.513	0.492	0.446	0.405

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 and of the bandwidth M, we give the fraction of the QMLE bias that is accounted for by the bias correction, *i.e.* the fraction $\sqrt{NT}\,\mathbb{E}(\hat{\beta} - \beta)/\mathbb{E}(\hat{W}^{-1}\hat{B})$, computed over 10,000 samples.

			size			size	
		WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$	$N = 100,\ T = 20,\ M = 4$	0.219	0.210	0.195	0.069	0.063	0.059
	$N = 400,\ T = 80,\ M = 6$	0.205	0.203	0.199	0.054	0.053	0.053
	$N = 400,\ T = 20,\ M = 4$	0.560	0.549	0.533	0.095	0.090	0.083
	$N = 1600,\ T = 80,\ M = 6$	0.591	0.588	0.584	0.056	0.055	0.055
$\rho^0 = 0.6$	$N = 100,\ T = 20,\ M = 4$	0.321	0.303	0.273	0.092	0.080	0.073
	$N = 400,\ T = 80,\ M = 6$	0.261	0.257	0.250	0.052	0.049	0.052
	$N = 400,\ T = 20,\ M = 4$	0.609	0.595	0.572	0.175	0.161	0.141
	$N = 1600,\ T = 80,\ M = 6$	0.668	0.663	0.658	0.063	0.060	0.062

Table 4: 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 size of the different tests is reported, based on 7,500 simulation runs.

			power			power		
			WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$	$N = 100$, $T = 20$, $M = 4$	H_a^{left}	0.094	0.087	0.076	0.131	0.122	0.123
		H_a^{right}	0.523	0.510	0.486	0.233	0.221	0.206
	$N = 400$, $T = 80$, $M = 6$	H_a^{left}	0.062	0.061	0.059	0.150	0.149	0.150
		H_a^{right}	0.547	0.544	0.538	0.196	0.193	0.193
	$N = 400$, $T = 20$, $M = 4$	H_a^{left}	0.301	0.292	0.280	0.103	0.098	0.100
		H_a^{right}	0.796	0.789	0.776	0.304	0.293	0.276
	$N = 1600$, $T = 80$, $M = 6$	H_a^{left}	0.244	0.242	0.239	0.135	0.133	0.135
		H_a^{right}	0.870	0.867	0.865	0.216	0.213	0.213
$\rho^0 = 0.6$	$N = 100$, $T = 20$, $M = 4$	H_a^{left}	0.189	0.169	0.144	0.175	0.156	0.164
		H_a^{right}	0.633	0.617	0.581	0.341	0.315	0.298
	$N = 400$, $T = 80$, $M = 6$	H_a^{left}	0.078	0.076	0.072	0.175	0.199	0.205
		H_a^{right}	0.681	0.676	0.671	0.341	0.265	0.271
	$N = 400$, $T = 20$, $M = 4$	H_a^{left}	0.436	0.422	0.395	0.175	0.155	0.153
		H_a^{right}	0.798	0.792	0.778	0.341	0.431	0.409
	$N = 1600$, $T = 80$, $M = 6$	H_a^{left}	0.318	0.313	0.307	0.205	0.167	0.172
		H_a^{right}	0.914	0.911	0.909	0.272	0.314	0.319

Table 5: As table 4, but we report the power for testing the alternatives $H_a^{\text{left}} : \rho = \rho^0 - (NT)^{-1/2}$ and $H_a^{\text{right}} : \rho = \rho^0 + (NT)^{-1/2}$.

			size corrected power			size corrected power		
			WD	LR	LM	WD^*	LR^*	LM^*
$\rho^0 = 0$	$N = 100$, $T = 20$, $M = 4$	H_a^{left}	0.013	0.012	0.011	0.106	0.105	0.112
		H_a^{right}	0.216	0.218	0.211	0.195	0.196	0.193
	$N = 400$, $T = 80$, $M = 6$	H_a^{left}	0.008	0.008	0.008	0.145	0.144	0.145
		H_a^{right}	0.251	0.251	0.250	0.188	0.187	0.188
	$N = 400$, $T = 20$, $M = 4$	H_a^{left}	0.006	0.006	0.006	0.056	0.054	0.063
		H_a^{right}	0.177	0.173	0.172	0.203	0.203	0.199
	$N = 1600$, $T = 80$, $M = 6$	H_a^{left}	0.006	0.005	0.006	0.125	0.126	0.129
		H_a^{right}	0.237	0.235	0.236	0.204	0.205	0.204
$\rho^0 = 0.6$	$N = 100$, $T = 20$, $M = 4$	H_a^{left}	0.010	0.011	0.012	0.109	0.110	0.126
		H_a^{right}	0.200	0.206	0.200	0.237	0.239	0.239
	$N = 400$, $T = 80$, $M = 6$	H_a^{left}	0.005	0.005	0.005	0.200	0.200	0.200
		H_a^{right}	0.297	0.296	0.296	0.266	0.266	0.265
	$N = 400$, $T = 20$, $M = 4$	H_a^{left}	0.014	0.015	0.015	0.034	0.038	0.050
		H_a^{right}	0.124	0.123	0.121	0.191	0.191	0.202
	$N = 1600$, $T = 80$, $M = 6$	H_a^{left}	0.004	0.005	0.005	0.149	0.150	0.149
		H_a^{right}	0.223	0.223	0.224	0.288	0.288	0.288

Table 6: As table 4 and 5, but we report the size corrected power.