Likelihood Expansion for Panel Regression Models with Factors *

Hyungsik Roger Moon‡ Martin Weidner‡

April 2009

Abstract

In this paper we provide a new methodology to analyze the (Gaussian) profile quasi likelihood function for panel regression models with interactive fixed effects, also called factor models. The number of factors is assumed to be known. Employing the perturbation theory of linear operators, we derive a power series expansion of the likelihood function in the regression parameters. Using this expansion we work out the first order asymptotic theory of the quasi maximum likelihood estimator (QMLE) 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 weak (as opposed to strict) exogeneity of the 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 discuss estimation in cases where the true parameter is on the boundary of the parameter set, and we 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 very successful in explaining cross-sectional variations of stock returns, *e.g.* Fama and French (1993). Additional regressors in these 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

*We thank Jinyong Hahn for very helpful discussions. We greatly appreciated 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, and from seminar participants at UCLA and USC. Moon is grateful for 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.

assumptions on the dynamics of the covariates or of the factors are made. In this context it is crucial that we allow for weakly exogenous regressors, like lagged dependent variables. Lagged dependent variables are also important for microeconomics 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 likelihood function of the interactive fixed effect model which is minimized over the parameters λ_i, f_t, and the regression coefficients. The profile quasi likelihood function of the model, in which λ_i and f_t are already concentrated out, becomes the sum of the $T - R$ smallest eigenvalues of the sample covariance matrix of the panel, where T is the cross-sectional size of the panel, and R is the number of factors.

The main contribution of the paper is to provide a general methodology to expand the profile quasi likelihood function as a power series expansion in the regression parameters. In particular, we derive the quadratic approximation which is necessary to establish the so-called first order asymptotic theory of the QMLE and to work out the limits of the classical test statistics (Wald, LR and LM test).

The conventional likelihood expansion is done mostly by a Taylor approximation in the regression coefficients. In our case this expansion is difficult to perform due to the implicit eigenvalue problem in the profile quasi likelihood function. The analytic properties of this objective function are not known in the literature so far. The approach we choose is to perform a joint expansion in the regression parameters and in the idiosyncratic error terms. Using the perturbation theory of linear operators we show that the profile quasi likelihood function is analytic in a neighborhood of the true parameter and we obtain a formula of the expansion coefficients for all orders.

Our likelihood expansion is valid with a general type of regressors, in particular we allow for weakly exogenous regressors and so called "low-rank" regressors, *e.g.* time-invariant and common regressors, or interacted dummy variables. We also allow for time-serial and cross-sectional correlation and heteroscedasticity of the idiosyncratic error terms. 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 quadratic likelihood expansion makes us understand the nature of the potential asymptotic bias in the QMLE caused by the incidental parameters, λ_i and f_t. This is possible because we know the approximate score in a closed form. What we find is that there are two main sources that may cause asymptotic bias. The first one is due to the presence of weakly exogenous regressors in either time or cross-sectional direction. The second one is due to heteroscedasticity or correlation of the idiosyncratic errors, again either in time or cross-sectional direction. These biases corresponds to the well-know incidental parameter problem in the panel data literature (Neyman and Scott, 1948).

As applications of the likelihood expansion we investigate three problems: (i) deriving the asymptotic distribution of the QMLE with weakly exogenous regressors using the alternative asymptotic $N, T \to \infty$, (ii) exploring the case where the true parameter is on the boundary of the parameter set, and (iii) studying the characteristics of the three classical test statistics for testing a general linear restriction on the regression parameters, again under the alternative asymptotic. The analysis of these three applications is new in the literature on panel regression models with interactive fixed effects.

To obtain the limiting distribution of the QMLE we need to derive the asymptotic properties of the approximated Hessian and of the approximated score, both known in explicit form from the profile quasi likelihood expansion. Under the assumption of independent error terms (but allowing for heteroscedasticity) we show that the score (and thus the QMLE) converges to a normal distribution, and we provide estimators for its asymptotic bias and covariance matrix, as well as for the probability limit of the approximated Hessian. These estimators do not require knowledge on whether the regressors are strictly or weakly exogenous. Using these estimators we construct a bias corrected QMLE. To prove consistency of the estimators it is convenient to use the expansions of the regression residuals and of the projectors of the estimated factors and factor loadings in the regression parameters. These expansions are a byproduct of the perturbation theory that is used to derive the likelihood expansion,

and they can be used whenever the factors and factor loadings are estimated by principal components even if the regression parameters are not estimated by maximum likelihood.

The analysis of the QMLE as described so far is performed under the assumption that the true parameter is an interior point of the parameter set. Combining our likelihood expansion with the results in Andrews (1999) we derive the asymptotic QMLE distribution for situations where the true parameter is on the boundary, given that the parameter set is locally approximated by a convex cone. Under these assumptions we also define a "bias corrected" QMLE and show that its distribution is the one that the QMLE would have for unbiased score function.

For testing a general linear hypothesis we consider the Wald, LR and LM tests. Knowing the limit of the QMLE, the analysis of the Wald test is straightforward. However, for the asymptotics of the LR and the LM tests, one needs to find the asymptotics of the likelihood and the score process. Again, using the likelihood expansion, we are able to approximate the LR and the LM tests to show that these tests are still asymptotically equivalent to the Wald test, but have a non-central χ^2-distribution due to the bias of the QMLE and of the score. Using our estimators for the asymptotic Hessian and score bias we provide bias corrected versions of the three test statistics and show that their limiting distribution is a χ^2-distribution.

Monte Carlo simulations are performed for an AR(1) model with interactive fixed effect (for one factor, $R = 1$). We find that the QMLE for the AR(1) coefficient is biased due to weak exogeneity of the regressor, and that this bias causes severe size distortions and power asymmetries when testing hypotheses on the coefficient. In accordance with our asymptotic results, we find that the QMLE bias becomes smaller as N and T become larger, but that the size distortions of the classical tests do not get smaller with increasing N and T (keeping the ratio N/T fixed). The bias correction formulas are found to capture about half of the QMLE bias at $T = 5$, but already about 90% of the QMLE bias at $T = 80$, *i.e.* the bias correction works better and better as T becomes larger. At finite sample, the bias corrected QMLE is still biased, but its bias vanishes faster than the one of the QMLE as N and T increase. Consequently, one finds the size distortions and power asymmetries of the bias corrected tests to be much lower than of the non-corrected tests, and in contrast to the non-corrected tests the size distortions of the bias corrected tests become smaller as N and T becomes larger (keeping the ratio N/T fixed).

For estimation, this paper 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.* they also allow for weakly exogenous regressors. In their asymptotics T is fixed, i.e. 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, 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 $N, T \to \infty$.

A paper that is closely related to our work is Bai (2009). He studies the QMLE for panel regression models with interactive fixed effects, but assuming strictly exogenous regressors, and using a different methodology to derive the asymptotic distribution. Bai starts from the first order condition of the quasi likelihood maximization problem to derive the first order asymptotic theory of the QMLE. He finds that under the alternative asymptotic and for strictly exogenous regressors the QMLE is biased due to correlation and heteroscedasticity of the error terms. He gives consistent estimators for these bias terms and for the QMLE covariance matrix, and thus provides a bias corrected estimator. He also

studies time-invariant and common regressors. Compared to our paper, Bai focuses on the properties of the QMLE, while we first study the characteristics of the likelihood function by using our expansion results from perturbation theory. This allows us to investigate situations where the true parameter is on the boundary, and to study the limiting distribution of the LR and LM test. As opposed to Bai, we allow for weakly exogenous regressors. This is important from an empirical viewpoint because the weakly exogenous regressors can have feedback from the dependent variable. Also, from a theoretical viewpoint it is important because the weakly exogenous regressors cause *additional* bias terms. Our treatment of "low-rank regressors" is more general than Bai's discussion since we allow not only for time-invariant and common regressors, but for all kinds of "low-rank regressors", *e.g.* also for interacted dummy variables that appear in "difference in difference" estimation and that are ruled out by Bai's assumptions.

Both our analysis and the one of Bai (2009) share the restriction that the number of factors has 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), and Onatski (2005). Bai (2009) informally discusses how to consistently estimate the number of factors in the presence of regressors. In this paper we do not address this issue.[1]

The paper is organized as follows. In the next section 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. In section 3 we present the expansion of the profile quasi likelihood function in the regression parameters, give a general discussion of the asymptotic bias of the QMLE, and also provide useful expansions of the regression residuals and of the principal component projectors in the regression parameters. In section 4 we apply the likelihood expansion to work out the asymptotic distribution of the QMLE. Under independent idiosyncratic error terms, but allowing for heteroscedasticity and weakly exogenous regressors, we present estimators for the different components of the asymptotic bias and thus provide a bias corrected QMLE. We also discuss the limiting distribution of the QMLE when the true parameter is on the boundary of the parameter set, and we work out the asymptotic distribution of the (bias corrected) classical test statistics. In section 5 the Monte Carlo simulation results for the AR(1) model are presented. Afterwards we conclude. All proofs of theorems and some technical details have been moved to the appendix. Some parts of the proofs and some further technical comments have been transferred to the supplementary material.[2]

A few words on notation. For a column vectors v its Euclidean norm is defined by $\|v\| = \sqrt{v'v}$. For the n-th largest eigenvalues (counting multiple eigenvalues multiple times) of a symmetric matrix B we write $\mu_n(B)$. For an $m \times n$ matrix A the Frobenius norm is $\|A\|_F = \sqrt{\mathrm{Tr}(AA')}$, and the operator 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. The Kronecker-delta symbol is defined by $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. We use "wpa1" for "with probability approaching one", and 1(.) for the indicator function.

[1] The discussion in Bai (2009) starts with the assertion of $\sqrt{NT}$-consistency of the QMLE of the regression parameters even when only an upper bound on the number of factors is known. The proof of this claim is non-trivial and will be the key for future research on estimating the number of factors in the presence of regressors.

[2] Which is available at `http://www-rcf.usc.edu/~moonr`.

2 Model, QMLE and Consistency

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

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

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.[3]

Model (2.1) can be written in matrix notation 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 reads[4]

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

The estimator we consider in this paper 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

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

where $\mathbb{B} \subset \mathbb{R}^K$ is a compact parameter set that contains the true parameter, *i.e.* $\beta^0 \in \mathbb{B}$,[5] and the objective function is the profile quasi likelihood function

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

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. This minimum is unique, but the minimizing parameters $\hat{\lambda}$ and $\hat{f}$ 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 form the first one by concentrating out λ, *i.e.* by eliminating it from the objective function by use of its own first order condition. Analogously, one can concentrate out f to obtain a formulation where only the parameter λ remains.

[3] Bai and Ng (2002) and Onatski (2005) provide methods to estimate the number of factors in pure factors models but without a regressor.

[4] More accurately, after concentrating out the variance of e_{it}, the (Gaussian) quasi likelihood function of the model is a constant times $[\mathcal{L}_{NT}(\beta, \lambda, f)]^{-NT/2}$, *i.e.* maximizing the likelihood is equivalent to minimizing $\mathcal{L}_{NT}(\beta, \lambda, f)$. Note also that $\mathcal{L}_{NT}(\beta, \lambda, f)$ is just $1/NT$ times the sum of squared residuals $\hat{e}_{it} = Y_{it} - \beta' X_{it} - \lambda_i^{0\prime} f_t^0$, because $\sum_i \sum_t \hat{e}_{it}^2 = \operatorname{Tr}(\hat{e}'\hat{e})$. This trace notation will be used extensively throughout the paper.

[5] If there are multiple global minima in $\mathbb{B}$ we want $\hat{\beta}$ to be one of them.

It turns out that the optimal f is obtained by combining the R eigenvectors that correspond to the R largest eigenvalues of the $T \times T$ matrix $\left(Y - \sum_{k=1}^{K} \beta_k X_k\right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k\right)$. From this follows the third way to write the profile quasi likelihood function, namely 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, and it is also most convenient for numerical computations of the QMLE – at each step of the numerical optimization over β one 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.[6] 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 $N, T \to \infty$, *i.e.* more precisely we want $\min(N,T) \to \infty$, but we allow for $\max(N,T)$ to grow at a faster rate. In the following we present assumptions on X_k, e, λ and f that guarantee consistency.[7]

Assumption 1. *The probability limits of $\lambda^{0\prime}\lambda^0/N$ and $f^{0\prime}f^0/T$ are finite and have full rank, i.e. (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. *(i)* $\operatorname{plim}_{N,T\to\infty}\left[(NT)^{-1}\mathrm{Tr}(X_k\, e')\right] = 0$, *(ii)* $\operatorname{plim}_{N,T\to\infty}\left[(NT)^{-1}\mathrm{Tr}(\lambda^0\, f^{0\prime}e')\right] = 0$.

Assumption 3. *The operator norm of the error matrix e grows at a rate smaller than $\sqrt{NT}$, i.e.* $\operatorname{plim}_{N,T\to\infty}\left(\|e\|/\sqrt{NT}\right) = 0$.

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 $f^0_{\cdot r}$ and factor loading $\lambda^0_{\cdot r}$ grows at a rate of $\sqrt{T}$ and $\sqrt{N}$, respectively. Assumption 2 demands that the regressors are weakly exogenous and that the idiosyncratic errors are weakly independent from the factors and factor loadings. 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.

(a) We assume that the probability limit of the $K \times K$ matrix $(NT)^{-1}\sum_{i,t} X_{it}X'_{it}$ exists and is positive definite, i.e. $\operatorname{plim}_{N,T\to\infty}\left[(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}X'_{it}\right] > 0$.

(b) 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:

(i) Consider linear combinations $X_{\mathrm{high},\alpha} = \sum_{m=K_1+1}^{K} \alpha_m X_m$ of the high-rank regressors X_m for K_2-vectors[8] α with $\|\alpha\| = 1$. We assume that there exists a constant $b > 0$ such that

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

(ii) For the low-rank regressors we assume $\operatorname{rank}(X_l) = 1$, $l = 1,\ldots,K_1$, i.e. the 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 $B > 0$ (independent of N,T) such that $N^{-1}\lambda^{0\prime}M_v\lambda^0 > B\,\mathbb{I}_R$ wpa1, and $T^{-1}f^{0\prime}M_w f^0 > B\,\mathbb{I}_R$ wpa1.

[6] 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 $\left(Y - \sum_{k=1}^{K}\beta_k X_k\right)\left(Y - \sum_{k=1}^{K}\beta_k X_k\right)'$.

[7] In principle we should 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.

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

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 $D_i = 1(i \in \mathbb{A})$ and $\tilde{D}_t = 1(t \in \mathbb{B})$ be dummy variables that indicate whether individual i is in $\mathbb{A} \subset \{1, \ldots, N\}$ (group dummy), and whether t is in $\mathbb{B} \subset \{1, \ldots, T\}$ (*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 $\text{rank}(X_{l,it}) = 1$, 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 $N, T \to \infty$.[9]

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 well known from ordinary fixed effect models, where the coefficients of time-invariant regressors are not identified. Assumption 4(b)(ii) 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 where their distribution guarantees that they have high rank (usually full rank) asymptotically, for example $X_{m,it} = 1 + Z_{it}$, where $Z_{it} \sim iid\mathcal{N}(0, 1)$. However, a high-rank regressors may still have a significant "low-rank component", *e.g.* $X_{m,it} = 1 + Z_{it} + \lambda_{ir}^0 f_{tr}^0$, where Z_{it} as above and λ_{ir}^0 and f_{tr}^0 are the r'th factor loading and factor.[10]

We can now state our consistency result for the QMLE.

Theorem 2.1. *Let assumptions 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 different assumptions on the regressors. 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 gives a condition on the regressors (his assumption A) that rules out low-rank regressors, *i.e.* that works for the case of only high-rank regressors. This condition still involves λ^0, which we felt should better be avoided for the high-rank regressors since λ^0 is not observable (only for the low-rank regressors it is necessary that λ^0 and f^0 appear in assumption 4). In a separate section Bai (2009) gives a condition on the regressors (in his notation $D(F^0) > 0$) that is applicable in the case of only time-invariant and

[9] 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(b)(i) on the high-rank regressors, namely replacing K_1 by R_X, *i.e.* we would have $\sum_{i=2R+R_x+1}^{N}$.

[10] 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(b)(i) 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(b)(i) 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.

common regressors, *i.e.* that does not guarantee consistency for high-rank regressors and for more general low-rank regressors.[11] 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 Profile Quasi Likelihood Expansion

The last expression in equation (2.5) for the profile quasi likelihood function is on the one hand very convenient because it does not involve any minimization over the parameters λ or f. On the other hand, this does not seem like an expression that can be easily discussed by analytic means, because in general there is no explicit formula for the n-th largest eigenvalue of a matrix. This complicates the analysis of the asymptotic distribution of the QMLE using the conventional method that involves Taylor approximation, because it is not straightforward how to compute derivatives in order to expand $L_{NT}(\beta)$ around β^0.

The key idea of this paper is to use the perturbation theory of linear operators to perform the expansion of $L_{NT}(\beta)$ around β^0. More precisely, we expand simultaneously in β and in the operator norm of the error term e. Let the $K+1$ expansion parameters be defined by $\epsilon_0 = \|e\|/\sqrt{NT}$ and $\epsilon_k = \beta_k^0 - \beta_k$, $k = 1, \ldots, K$, and define the $N \times T$ matrix $X_0 = (\sqrt{NT}/\|e\|)e$. With these definitions we obtain

$$\frac{1}{\sqrt{NT}}\left(Y - \sum_{k=1}^{K} \beta_k X_k\right) = \frac{\lambda^0 f^{0\prime}}{\sqrt{NT}} + \sum_{\kappa=0}^{K} \epsilon_k \, \frac{X_\kappa}{\sqrt{NT}} \; , \tag{3.1}$$

and according to equation (2.5) the profile quasi likelihood function $L_{NT}(\beta)$ can be written as the sum over the $T-R$ smallest eigenvalues of this matrix multiplied with its transposed. We consider $\sum_{\kappa=0}^{K} \epsilon_k \, X_\kappa/\sqrt{NT}$ as a small perturbation of the unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$. The goal is to expand the profile quasi likelihood $L_{NT} = L_{NT}(\epsilon)$ in the perturbation parameters $\epsilon = (\epsilon_0, \ldots, \epsilon_K)$, *i.e.* in a neighborhood of $\epsilon = 0$ we want to write

$$L_{NT}(\epsilon) = \frac{1}{NT} \sum_{g=2}^{\infty} \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \ldots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \, \epsilon_{\kappa_2} \, \ldots \, \epsilon_{\kappa_g} \, L^{(g)}\left(\lambda^0, \, f^0, \, X_{\kappa_1}, \, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \; , \tag{3.2}$$

where $L^{(g)} = L^{(g)}\left(\lambda^0, \, f^0, \, X_{\kappa_1}, \, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)$ are the expansion coefficients.

Note that the unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$ has rank R. Thus, the $T-R$ smallest eigenvalues of the unperturbed $T \times T$ matrix $f^0 \lambda^{0\prime} \lambda^0 f^{0\prime}/NT$ are all zero, and due to assumption 1 on λ^0 and f^0 we find that the R non-zero eigenvalues of this $T \times T$ matrix converge to positive constants as $N, T \to \infty$. In more technical terms this means that the "separating distance" of the zero-eigenvalue of the unperturbed $T \times T$ converges to a positive constant. Under this condition the perturbation theory of linear operators guarantees that the above expansion of L_{NT} in ϵ exists and is convergent as long as the operator norm of the perturbation matrix $\sum_{\kappa=0}^{K} \epsilon_k \, \frac{X_\kappa}{\sqrt{NT}}$ is smaller than the convergence radius $r_0(\lambda^0, f^0)$. For details, see Kato (1980) and appendix C. In the appendix the convergence radius $r_0(\lambda^0, f^0)$ is defined and it is shown that under assumption 1 it converges to a positive constant in probability as $N, T \to \infty$.

Thus, the above expansion of the profile quasi likelihood function is applicable whenever the operator norm of the perturbation matrix $\sum_{\kappa=0}^{K} \epsilon_k \, \frac{X_\kappa}{\sqrt{NT}}$ is smaller than $r_0(\lambda^0, f^0)$. Fortunately, when evaluated at a consistent estimator $\beta = \hat{\beta}$ this is the case asymptotically. Note that $\|X_\kappa/\sqrt{NT}\| = \mathcal{O}_p(1)$ for $\kappa = 0, \ldots, K$. For $\kappa = 0$ this is true by definition, and for $\kappa = k = 1, \ldots, K$ this is satisfied due to assumption 4, namely we have $\|X_k\| \leq \|X_k\|_F = \mathcal{O}_p(\sqrt{NT})$. In addition, assumption 3 guarantees that $\epsilon_0 \to_p 0$, and for $\beta = \hat{\beta}$ with $\hat{\beta} \to_p \beta^0$ we also have $\epsilon_k \to_p 0$ for $\kappa = k = 1, \ldots, K$. Thus, the operator norm of the perturbation converges to zero in probability if evaluated for a consistent

[11] In the supplementary material we give two examples that show that Bai's condition $D(F^0) > 0$ does not guarantee consistency in a more general case.

estimator of β. This shows how our assumption on the model play together to guarantee that the above likelihood expansion is valid asymptotically.[12]

Perturbation theory (*e.g.* Kato (1980)) also provides an explicit formula for the expansion coefficients $L^{(g)}$. For example, $L^{(1)}\left(\lambda^0,\, f^0,\, X_\kappa\right) = 0$, and $L^{(2)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}\right) = \mathrm{Tr}(M_{\lambda^0} X_{\kappa_1} M_{f^0} X'_{\kappa_2})$. The general formula is given in theorem C.2 in the appendix. Using this formula one can derive the following bound

$$\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2},\ldots,X_{\kappa_g}\right)\right| \;\leq\; a_{NT}\,(b_{NT})^g\,\frac{\|X_{\kappa_1}\|}{\sqrt{NT}}\,\frac{\|X_{\kappa_2}\|}{\sqrt{NT}}\,\cdots\,\frac{\|X_{\kappa_g}\|}{\sqrt{NT}}\,, \tag{3.3}$$

where a_{NT} and b_{NT} are functions of λ^0 and f^0 that converge to finite positive constants in probability, *i.e.* $a_{NT} \to_p a < \infty$ and $b_{NT} \to_p b < \infty$. This bound on the coefficients $L^{(g)}$ allows to work out a bound on the remainder term, when the likelihood expansion is truncated at a particular order.

3.1 Quadratic Approximation of the Likelihood Function

The assumptions on the model made so far are sufficient to expand $L_{NT}(\beta)$ in $(\beta-\beta^0)$ and $\|e\|/\sqrt{NT}$. But in order to cut the expansion in $\|e\|/\sqrt{NT}$ at a finite order and be able to give a useful bound on the remainder term, we need to strengthen assumption 3 slightly.

Assumption 3*. *We assume that there exists a deterministic ξ_{NT} and and a positive integer G_e such that $\|e\|/\sqrt{NT} = \mathcal{O}_p(\xi_{NT})$, for some series ξ_{NT} that satisfies $\sqrt{NT}\,(\xi_{NT})^{G_e} \to 0$ as $N,T \to \infty$.*

Note that the value of the constant G_e not only depends on the distributional assumptions for the error term e_{it}, but also on the particular convergence scheme of N and T. For all examples of error distributions given in appendix A we have $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$, *i.e.* $\xi_{NT} = \min(N,T)^{-\frac{1}{2}}$. There is a large literature that studies the asymptotic behavior of the operator norm of random matrices, see *e.g.* German (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). Assuming this is satisfied and considering the limit $N,T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we find assumption 3* to be satisfied with $G_e = 3$.

We can now present the quadratic approximation of the profile quasi likelihood function $L_{NT}(\beta)$.

Theorem 3.1. *Let assumptions 1, 3*, and 4(a) be satisfied with $G_e \geq 3$. Then, the profile quasi likelihood function satisfies $L_{NT}(\beta) = L_{q,NT}(\beta) + I_{NT} + (NT)^{-1} R_{NT}(\beta)$, where I_{NT} is independent of β, the remainder $R_{NT}(\beta)$ is such that for any sequence $\eta_{NT} \to 0$ we have*

$$\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}} \frac{|R_{NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} = o_p\left(1\right), \tag{3.4}$$

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

$$L_{q,NT}(\beta) \;=\; (\beta-\beta^0)'\,W_{NT}\,(\beta-\beta^0) \;-\; \frac{2}{\sqrt{NT}}\,(\beta-\beta^0)'\,C_{NT}\,, \tag{3.5}$$

with $K \times K$ matrix $W_{NT} = W_{NT}(\lambda^0, f^0, X)$ 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 $C_{NT} = C_{NT}(\lambda^0, f^0, e, X)$ given by $C_{NT,k} = \sum_{g=2}^{G_e} C^{(g)}\left(\lambda^0\,, f^0\,, X_k\,e\right)$. The general formula for the coefficients $C^{(g)}$ is $C^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = g(4NT)^{-1/2}\,L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e,\, e,\ldots,e\right)$,

[12] Note that all we need for this result is assumptions 1 and 3, $\|X_k\| = \mathcal{O}_p(\sqrt{NT})$, and consistency of $\hat{\beta}$. However, in order to achieve consistency of the QMLE we also have to impose assumptions 2 and 4.

with $L^{(g)}$ defined in theorem C.2 of the appendix. For $g = 2$ and $g = 3$ we have

$$C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = \frac{1}{\sqrt{NT}}\,\mathrm{Tr}(M_{f^0}\, e'\, M_{\lambda^0}\, X_k)\;,$$

$$\begin{aligned} C^{(3)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = -\,\frac{1}{\sqrt{NT}}\,\Big[& \mathrm{Tr}\left(e M_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) \\ & + \mathrm{Tr}\left(e' M_{\lambda^0}\, e\, M_{f^0}\, X_k'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\ & + \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right)\Big]\;. \end{aligned} \tag{3.6}$$

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.

Using the bound on the remainder $R_{NT}(\beta)$ given in equation (3.4), 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 a bound on $\nabla R_{NT}(\beta)$ that is useful to derive the limiting distribution of the Lagrange multiplier test in the application section below.

Theorem 3.2. *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 $\eta_{NT} \to 0$

$$\sup_{\{\beta:\|\beta-\beta^0\|\le\eta_{NT}\}} \frac{\|\nabla R_{NT}(\beta)\|}{\sqrt{NT}\left(1 + \sqrt{NT}\,\left\|\beta - \beta^0\right\|\right)} = o_p\,(1)\,. \tag{3.7}$$

3.2 Expansions of Projectors and Residuals

It is convenient to also have the asymptotic β-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 minimizing value $\beta = \hat\beta$. The corresponding residuals are defined by

$$\hat e(\beta) = Y \;-\; \sum_{k=1}^{K} \beta_k\, X_k \;-\; \hat\lambda(\beta)\, \hat f'(\beta)\;. \tag{3.8}$$

Theorem 3.3. *Under assumptions 1, 3, and 4(a) we have the following expansions*

$$\begin{aligned} M_{\hat\lambda}(\beta) &= M_{\lambda^0} + M^{(1)}_{\hat\lambda,e} + M^{(2)}_{\hat\lambda,e} - \sum_{k=1}^{K} \left(\beta_k - \beta_k^0\right) M^{(1)}_{\hat\lambda,k} + M^{(\mathrm{rem})}_{\hat\lambda}(\beta)\;, \\ M_{\hat f}(\beta) &= M_{f^0} + M^{(1)}_{\hat f,e} + M^{(2)}_{\hat f,e} - \sum_{k=1}^{K} \left(\beta_k - \beta_k^0\right) M^{(1)}_{\hat f,k} + M^{(\mathrm{rem})}_{\hat f}(\beta)\;, \\ \hat e(\beta) &= M_{\lambda^0}\, e\, M_{f^0} + \hat e^{(1)}_e - \sum_{k=1}^{K} \left(\beta_k - \beta_k^0\right) \hat e^{(1)}_k + \hat e^{(\mathrm{rem})}(\beta)\;, \end{aligned} \tag{3.9}$$

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

$$\begin{aligned}
\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}} \frac{\left\|M^{(\mathrm{rem})}_{\hat{\lambda}}(\beta)\right\|}{\|\beta-\beta^0\|^2+(NT)^{-1/2}\|e\|\,\|\beta-\beta^0\|\ +(NT)^{-3/2}\|e\|^3} &= \mathcal{O}_p\left(1\right)\,,\\
\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}} \frac{\left\|M^{(\mathrm{rem})}_{\hat{f}}(\beta)\right\|}{\|\beta-\beta^0\|^2+(NT)^{-1/2}\|e\|\,\|\beta-\beta^0\|\ +(NT)^{-3/2}\|e\|^3} &= \mathcal{O}_p\left(1\right)\,,\\
\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}} \frac{\left\|\hat{e}^{(\mathrm{rem})}(\beta)\right\|}{(NT)^{1/2}\|\beta-\beta^0\|^2+\|e\|\,\|\beta-\beta^0\|+(NT)^{-1}\|e\|^3} &= \mathcal{O}_p\left(1\right)\,, \qquad (3.10)
\end{aligned}$$

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

$$\begin{aligned}
M^{(1)}_{\hat{\lambda},e} &= -M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\ -\ \lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\,,\\
M^{(1)}_{\hat{\lambda},k} &= -M_{\lambda^0}\,X_k\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\ -\ \lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,X_k'\,M_{\lambda^0}\,,\\
M^{(2)}_{\hat{\lambda},e} &= M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\\
&\quad+\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\\
&\quad-M_{\lambda^0}\,e\,M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\\
&\quad-\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\,e'\,M_{\lambda^0}\\
&\quad-M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\\
&\quad+\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\,, \qquad (3.11)
\end{aligned}$$

analogously

$$\begin{aligned}
M^{(1)}_{\hat{f},e} &= -M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}f^{0\prime}\ -\ f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\,,\\
M^{(1)}_{\hat{f},k} &= -M_{f^0}\,X_k'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}f^{0\prime}\ -\ f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,\epsilon_k\,M_{f^0}\,,\\
M^{(2)}_{\hat{f},e} &= M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}f^{0\prime}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}f^{0\prime}\\
&\quad+f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\\
&\quad-M_{f^0}\,e'\,M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\\
&\quad-f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\,e\,M_{f^0}\\
&\quad-M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\\
&\quad+f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}f^{0\prime}\,, \qquad (3.12)
\end{aligned}$$

and finally

$$\begin{aligned}
\hat{e}^{(1)}_k &= M_{\lambda^0}\,X_k\,M_{f^0}\,,\\
\hat{e}^{(1)}_e &= -M_{\lambda^0}\,e\,M_{f^0}\,e'\,\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\\
&\quad-\lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime}f^0)^{-1}\,f^{0\prime}\,e'\,M_{\lambda^0}\,e\,M_{f^0}\\
&\quad-M_{\lambda^0}\,e\,f^0\,(f^{0\prime}f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\,e\,M_{f^0}\,. \qquad (3.13)
\end{aligned}$$

In theorem C.2 of the appendix we give the general expansion of $M_{\hat{\lambda}}(\beta)$ up to arbitrary orders in β and e. The general expansion of $M_{\hat{f}}(\beta)$ can be obtained from the one for $M_{\hat{\lambda}}(\beta)$ by applying symmetry ($N \leftrightarrow T$, $\lambda \leftrightarrow f$, $Y \leftrightarrow Y'$, $X_k \leftrightarrow X_k'$), and the general expansion for $\hat{e}(\beta)$ can be obtained via $\hat{e}(\beta) = M_{\hat{\lambda}}(\beta)\left[Y\ -\ \sum_{k=1}^K\,\beta_k\,X_k\right]$, with Y given in equation (2.2). For most purposed the expansions up to the finite orders given above should be sufficient.

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.

These expansions are very useful. In the present paper we make use of them in the proof of theorem 4.4 below in order to derive the properties of the variance and bias estimates of the QMLE, *i.e.* of objects that contain $M_{\hat{\lambda}}(\beta)$, $M_{\hat{f}}(\beta)$, and $\hat{e}$. More generally, one can use these expansions in situations where $\hat{\lambda}$ and $\hat{f}$ are still defined as principal components estimators, but where a different estimator for β (not the QMLE) is used. For those alternative estimators the likelihood expansion in theorem 3.1 is irrelevant, but the expansions in theorem 3.3 are still applicable as long as principal components are used to estimate factors and factor loadings.

3.3 Remarks

$\sqrt{NT}$-consistency of the QMLE

The following corollary is the key for working out the asymptotic distribution of the QMLE.

Corollary 3.4. *Under the assumptions of the theorems 2.1 and 3.1, and assuming that β^0 is an interior point of the parameter set $\mathbb{B}$ we have $\sqrt{NT}\left(\hat{\beta}_{k_1} - \beta^0_{k_1}\right) = W_{NT}^{-1} C_{NT} + o_p(1)$.*

Having consistency of the QMLE and the expansion of the profile quasi likelihood function in theorem 3.1, in particular the bound on the remainder term given there, one finds $\sqrt{NT} W_{NT}\left(\hat{\beta}_{k_1} - \beta^0_{k_1}\right) = C_{NT} + o_p(1)$, see *e.g.* Andrews (1999). Alternatively, one can solve the first order conditions and use the bound on $\nabla R_{NT}(\beta)$ in theorem 3.2 to obtain the same result. To obtain the above corollary one needs in addition that W_{NT} does not degenerate as $N, T \to \infty$, *i.e.* the smallest eigenvalue of W_{NT} should by bounded by a positive constant. Our assumptions made so far already guarantee this, as is shown in the supplementary material. The corollary shows that the QMLE $\hat{\beta}$ is $\sqrt{NT}$-consistent if $C_{NT} = \mathcal{O}_p(1)$.

Asymptotic Bias of the QMLE

Corollary 3.4 can be used to derive the limiting distribution of the QMLE $\hat{\beta}$ under different distributional assumptions on λ^0, f^0, e, and X_k, and for different asymptotics $T, N \to \infty$. The restriction on e and X_k made to derive the corollary still allow for very general cross-sectional and time-serial correlation of the errors, and for very general weakly exogenous regressors. In order to actually compute the limiting distribution of $\hat{\beta}$ more specific assumptions on λ^0, f^0, e, and X_k have to me made, depending on the particular application in mind. A concrete example of these more specific assumptions is given in the application section below.

It is natural to assume that the approximated Hessian W_{NT} converges to a constant matrix in probability as $N, T \to \infty$, see also Bai (2009). Thus, according to corollary 3.4 the asymptotic distribution of $\hat{\beta}$ is up to a matrix multiplication given by the asymptotic distribution of the approximated score C_{NT}. Asymptotic bias of $\hat{\beta}$ therefore corresponds to asymptotic bias of C_{NT}, and we now give an informal discussion of the different bias terms that can occur.

According to theorem 3.1 the approximated score is proportional to the sum over the terms $C^{(g)}\left(\lambda^0, f^0, X_k, e\right)$ from $g = 2$ to G_e. In the following we restrict attention to the terms $g = 2$ and $g = 3$, and discuss under what conditions these terms contribute an asymptotic bias to the QMLE. As discussed previously, for $\|e\| = \mathcal{O}_p(\max(N, T))$ and $N/T \to \kappa^2$, $0 < \kappa < \infty$, asymptotically we have $G_e = 3$, *i.e.* under these conditions higher order score terms do not contribute to the limiting distribution of $\hat{\beta}$. In the following, for expositional simplicity, λ^0 and f^0 are treated as non-stochastic.

We start with the discussion of the $C^{(2)}$ term. If the regressors X_k are strictly exogenous we have $\mathbb{E}\left[C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right)\right] = 0$, *i.e.* no asymptotic bias originates from $C^{(2)}$ in this case. However, if the regressors are weakly exogenous we have[13]

$$\begin{aligned}\mathbb{E}\left[C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right)\right] &= -\sqrt{\frac{N}{T}}\,\mathrm{Tr}\left[P_{f^0}\mathbb{E}\left(\frac{1}{N}e'\,X_k\right)\right] - \sqrt{\frac{T}{N}}\,\mathrm{Tr}\left[P_{\lambda^0}\,\mathbb{E}\left(\frac{1}{T}e\,X_k'\right)\right] + o(1) \\ &= -\sqrt{\frac{N}{T}}\sum_{t=1}^{T}\sum_{\tau=1}^{T} P_{f^0,t\tau}\,\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(e_{it}\,X_{k,i\tau}\right) \\ &\quad - \sqrt{\frac{T}{N}}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{\lambda^0,ij}\,\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(e_{it}\,X_{k,jt}\right) + o(1)\,. \end{aligned} \tag{3.14}$$

The first bias term we find here is non-zero if $\mathbb{E}\left(e_{it}\,X_{k,i\tau}\right) \neq 0$ for $t > \tau$, *i.e.* if the past innovation e_{it}, or equivalently the past Y_{it}, influences the future regressors $X_{k,i\tau}$. This bias term would also be present if the factors f^0 would be observed. In the special case of only one factor which is observed to be $f_i^0 = 1$ the QMLE becomes just the within-group estimator for the fixed effect model. For this special case the above bias term was first derived by Nickell (1981). We have given the generalization of this bias for more general f^0, and we have shown that the same bias term is present if the factors are unobserved and estimated jointly with the regression coefficients.

The second bias term we find here is non-zero if $\mathbb{E}\left(e_{it}\,X_{k,jt}\right) \neq 0$. This bias term would be relevant in applications in which one allows for spatial correlation, for example, when the dependent variable Y_{it} for unit i appears as a regressors in the equation for Y_{jt} of unit $j \neq i$.[14]

For the discussion of the $C^{(3)}$ terms, we assume for simplicity that the regressors X_k are strictly exogenous and non-stochastic. We then have[15]

$$\begin{aligned}\mathbb{E}\left[C^{(3)}\left(\lambda^0,\, f^0,\, X_k,\, e\right)\right] &= -\sqrt{\frac{T}{N}}\mathrm{Tr}\left[\lambda^{0\prime}\mathbb{E}\left(\frac{1}{T}\,e\,e'\right)\,M_{\lambda^0}\,X_k\,f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\right] \\ &\quad - \sqrt{\frac{N}{T}}\mathrm{Tr}\left[f^{0\prime}\mathbb{E}\left(\frac{1}{N}\,e'\,e\right)\,M_{f^0}\,X_k'\,\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}\right] + o(1)\,. \end{aligned} \tag{3.15}$$

These are the two bias terms that were already found by Bai (2009). For error terms e_{it} that are cross-sectionally independent and homoscedastic we have $\mathbb{E}\left(T^{-1}\,e\,e'\right) = \mathbb{I}_N$, and the first bias term in equation (3.15) then is zero since $\lambda^{0\prime}M_{\lambda^0} = 0$. However, under cross-sectional correlation or heteroscedasticity of e_{it} this bias term is non-zero. Analogously, for errors e_{it} that are time-serial independent and homoscedastic we have $\mathbb{E}\left(N^{-1}\,e'\,e\right) = \mathbb{I}_T$, *i.e.* the second bias term in equation (3.15) is zero. This term contributes asymptotic bias to the QMLE only under time-serial correlation or heteroscedasticity.

Thus, if e_{it} is iid across i and t we expect no asymptotic bias from the $C^{(3)}$ terms (this is true even if regressors are not strictly exogenous), but there may still be asymptotic bias from the $C^{(2)}$ term due to weak exogeneity.

---

[13] Here we assumed that $\mathbb{E}\left[(NT)^{-1/2}\mathrm{Tr}(P_{f^0}\,e'\,P_{\lambda^0}\,X_k)\right] = o(1)$, which can be shown to be true under additional assumptions on e and X_k, and for N and T growing at the same rate, see section 4.1.

[14] For this to be consistent with weak exogeneity we need a partial ordering on the cross-sectional labels so that Y_{it} only appears in the equation for Y_{jt} if $i > j$ according to this ordering.

[15] Here we assume that $\mathrm{Tr}\left(eP_{f^0}\,e'\,M_{\lambda^0}\,X_k\,f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right) = o_p(1)$, $\mathrm{Tr}\left(e'P_{\lambda^0}\,e\,M_{f^0}\,X_k'\,\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}f^{0\prime}\right) = o_p(1)$, and $\mathrm{Tr}\left(e'M_{\lambda^0}\,X_k\,M_{f^0}\,e'\,\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}f^{0\prime}\right) = o_p(1)$. In the application section below we give an example of low-level assumptions on e and X_k under which this is true. In general, the above equations are satisfied as soon as one can show that $\|P_{\lambda^0}eP_{\lambda^0}\| = \mathcal{O}_p(1)$, and $\|P_{\lambda^0}eX_k'\| = \mathcal{O}_p(\sqrt{NT})$.

4 Applications of the Likelihood Expansion

4.1 Asymptotic Distribution and Bias Correction of the QMLE

In this subsection we apply corollary 3.4 to work out the asymptotic distribution of the QMLE $\hat{\beta}$, and to correct for the asymptotic bias. For this purpose we need more specific assumptions on λ^0, f^0, X_k and e. These additional specifications can be made differently, depending on the particular empirical application one has in mind.

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 $X_k = X_k^{\text{str}} + X_k^{\text{weak}}$. The component X_k^{str} is strictly exogenous, i.e. $X_{k,it}^{\text{str}}$ is independent of $e_{j\tau}$ for all i,j,t,τ. The component X_k^{weak} is weakly exogenous and we assume

$$X_{k,it}^{\text{weak}} = \sum_{\tau=1}^{t-1} c_{k,i\tau}\, e_{i,t-\tau}\ , \tag{4.1}$$

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

$$|c_{k,i\tau}| \ < \ \alpha^{\tau}\ , \tag{4.2}$$

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 $\mathbb{E}(X_{k,it}^{\text{str}})^{8+\epsilon}$ 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(i) is needed 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 some higher moments of λ_i^0 and f_t^0, but the assumptions as it is now is very convenient from a theoretical perspective, *e.g.* it guarantees that $P_{f^0,t\tau}$ is of order $1/T$ uniformly across t,τ and T.

Assumption 5(ii) requires cross-sectional and time-serial independence of e_{it}, but heteroscedasticity in both directions is still allowed, *i.e.* we still expect an asymptotic bias of the QMLE due to the $C^{(3)}$ term. In the supplementary material we show that assumption 5(ii) guarantees that $\|e\| = \mathcal{O}_p(\max(N,T))$, *i.e.* for the asymptotics $N,T\to\infty$ that is specified in assumption 5(iv) we find assumption 3^* to be satisfied with $G_e=3$. Assumption 2 is also satisfied as a consequence of assumption 5, *i.e.* assumption 5 guarantees that our quadratic expansion of the profile quasi likelihood function is applicable.

Assumption 5(iii) requires that the regressors X_k are additively separable into a strictly and a weakly exogenous component and assumes that the weakly exogenous component can be written as an MA(∞) process with innovation e_{it}.[16] 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

$$\begin{pmatrix} Y_{it} \\ Z_{it} \end{pmatrix} = B \begin{pmatrix} Y_{i,t-1} \\ Z_{i,t-1} \end{pmatrix} + \begin{pmatrix} \lambda_i^{0\prime} f_t^0 \\ d_{it} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \Gamma & \mathbb{I} \end{pmatrix} \begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}, \tag{4.3}$$

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

where Z_{it} is an $r\times 1$ vector of additional variables, B is an $(r+1)\times(r+1)$ matrix of parameters, the $r\times 1$ vectors d_{it} and u_{it} are independent of e_{it}, and Γ is an $r\times r$ 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}.[17] The first row in equation (4.3) 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 and weakly exogenous regressors as demanded in assumption 5(iii). The generalization of this example to VAR processes of higher order is straightforward.

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

Assumption 6. *Let $\mathfrak{X}_k = M_{\lambda^0} X_k^{\text{str}} M_{f^0} + X_k^{\text{weak}}$ and for each i,t define the K-vector $\mathfrak{X}_{it} = (\mathfrak{X}_{1,it},\ldots,\mathfrak{X}_{K,it})'$. The $K\times K$ matrices W and Ω, and the K-vectors B_1, B_2 and B_3, are defined below, and we assume that they exist:*

$$
\begin{aligned}
W &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathfrak{X}_{it}\,\mathfrak{X}_{it}'\,,\\
\Omega &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}\left[e_{it}^2\,\mathfrak{X}_{it}\,\mathfrak{X}_{it}'\right]\,,\\
B_{1,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{N}\operatorname{Tr}\left[P_{f^0}\mathbb{E}\left(e' X_k^{\text{weak}}\right)\right]\,,\\
B_{2,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{T}\operatorname{Tr}\left[\mathbb{E}\left(ee'\right)\,M_{\lambda^0}\,X_k^{\text{str}}\,f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right]\,,\\
B_{3,k} &= \operatorname*{plim}_{N,T\to\infty} \frac{1}{N}\operatorname{Tr}\left[\mathbb{E}\left(e'e\right)\,M_{f^0}\,X_k^{\text{str}\,\prime}\,\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}f^{0\prime}\right]\,.
\end{aligned}
\tag{4.4}
$$

Theorem 4.1. *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*

$$
\sqrt{NT}\left(\hat\beta-\beta^0\right)\underset{d}{\to}\mathcal{N}\left(W^{-1}B,\;W^{-1}\,\Omega\,W^{-1}\right)\,, \tag{4.5}
$$

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

From corollary 3.4 we already know that the limiting distribution of $\hat\beta$ is given by the limiting distribution of $W_{NT}^{-1}C_{NT}$. To proof theorem 4.1 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, although the equivalence is non-trivial since in $\mathfrak{X}_k$ the weakly exogenous part 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 whether the corresponding subspaces (of fixed dimension) are projected out of X_k^{weak} (whose dimension grows to infinity). 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}. For later applications the definition of W given in assumption 6 may be easier to evaluate (*e.g.* in a lagged dependent variable model we have $X_k^{\text{str}}=0$.) Note that 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 $C^{(2)}$ term. The strictly exogenous part of the regressors only contributes to the asymptotic variance, but the weakly exogenous 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 $\tau>t$). In section

[17] To guarantee independence (not merely uncorrelatetness) of e_{it} and u_{it} one has to assume normally distributed errors in this example.

3.3 we also discussed a bias due to correlation of errors and regressors in the cross-sectional dimension, but here we assume cross-sectional independence, *i.e.* this second type of bias is not present.

The three $C^{(3)}$ terms contribute no variance, *i.e.* they converge to constants in probability. One $C^{(3)}$ term is vanishing, an the other two contribute the asymptotic biases B_2 and B_3 that are due to cross-sectional and time-serial heteroscedasticity. Note that the weakly exogenous part of the regressors does not contribute to B_2 and B_3.

In order to express our estimators for asymptotic bias and asymptotic variance we first have to introduce some notation.

Definition 4.2. *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 4.3. *Let $\hat{\mathfrak{X}}_k(\beta) = M_{\hat{\lambda}}(\beta) \, X_k \, M_{\hat{f}}(\beta)$, and for each i,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*

$$
\begin{aligned}
\hat{W}(\beta) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\mathfrak{X}}_{it} \, \hat{\mathfrak{X}}_{it}' \, , \\
\hat{\Omega}(\beta) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2 \, \hat{\mathfrak{X}}_{it} \, \hat{\mathfrak{X}}_{it}' \, , \\
\hat{B}_{1,k}(\beta) &= \frac{1}{N} \operatorname{Tr}\left[P_{\hat{f}} \, (\hat{e}' X_k)^{\text{truncR}} \right] \, , \\
\hat{B}_{2,k}(\beta) &= \frac{1}{T} \operatorname{Tr}\left[(\hat{e} \, \hat{e}')^{\text{truncD}} \, M_{\hat{\lambda}} \, X_k \, \hat{f} \, (\hat{f}' \hat{f})^{-1} \, (\hat{\lambda}' \hat{\lambda})^{-1} \, \hat{\lambda}' \right] \, , \\
\hat{B}_{3,k}(\beta) &= \frac{1}{N} \operatorname{Tr}\left[(\hat{e}' \, \hat{e})^{\text{truncD}} \, M_{\hat{f}} \, X_k' \, \hat{\lambda} \, (\hat{\lambda}' \hat{\lambda})^{-1} \, (\hat{f}' \hat{f})^{-1} \, \hat{f}' \right] \, , \\
\hat{B}(\beta) &= -\sqrt{\frac{N}{T}} \, \hat{B}_{1,k}(\beta) - \sqrt{\frac{T}{N}} \, \hat{B}_{2,k}(\beta) - \sqrt{\frac{N}{T}} \, \hat{B}_{3,k}(\beta) \, ,
\end{aligned} \tag{4.6}
$$

where we suppressed the β-dependence of $\mathfrak{X}$, $\hat{e}$, $\hat{f}$, and $\hat{\lambda}$ on the right hand side.[18]

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 4.4. *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 $\hat{B}_3(\hat{\beta}) = B_3 + o_p(1)$.*

Note that 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 weakly exogenous, 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.

[18] Here $\hat{f}(\beta)$ and $\hat{\lambda}(\beta)$ are the principal component estimators defined above, and $\hat{e}(\beta)$ are the corresponding residuals defined in equation (3.8).

Corollary 4.5. *Under assumptions 5 and 6, for β^0 being an interior point of the compact parameter set $\mathbb{B}$, and for $M \to \infty$ and $M^5/T \to 0$ we find that 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 4.4, a consistent estimator of the asymptotic variance of $\hat{\beta}^*$ is given by $\hat{W}^{-1}(\hat{\beta})\,\hat{\Omega}(\hat{\beta})\,\hat{W}^{-1}(\hat{\beta})$.

4.2 Asymptotic Distribution when the True Parameter is on the Boundary

In corollary 4.5 we gave a bias corrected estimator $\hat{\beta}^*$ and its limiting distribution under the assumption that β^0 is an interior point of the parameter set $\mathbb{B}$, *i.e.* when there are no local parameter restriction on β. In the present subsection we discuss situations where β^0 is on the boundary of $\mathbb{B}$, *i.e.* when local parameter restrictions are present. In this case, one can use the result of Andrews (1999) to obtain the limiting distribution of the QMLE, once the quadratic expansion of the profile quasi likelihood function is obtained and the limiting distribution of the approximated score and Hessian are derived, and it is not difficult to apply Andrews' method also to derive the limiting distribution of an appropriately defined "bias corrected" QMLE. The following assumption will be used in this subsection and in the next one.

Assumption 7.

(i) We have a scalar objective function $L_{NT}(\beta)$ that is used to estimate the parameter $\beta \in \mathbb{B} \subset \mathbb{R}^K$, whose true value $\beta^0 \in \mathbb{B}$. We assume that the objective function has an asymptotic quadratic expansion of the form $L_{NT}(\beta) = L_{q,NT}(\beta) + I_{NT} + \frac{1}{NT} R_{NT}(\beta)$, where I_{NT} is independent of β, the remainder $R_{NT}(\beta)$ satisfies the condition in equation (3.4), and $L_{q,NT}(\beta) = (\beta - \beta^0)' W_{NT} (\beta - \beta^0) - 2 (NT)^{-1/2} (\beta - \beta^0)' C_{NT}$ is a second order polynomial.

(ii) We consider a limit $N, T \to \infty$, which may satisfy additional restrictions (e.g. $N/T \to$ const.). For this asymptotics, we assume that there exist positive definite $K \times K$ matrices Ω and W and a K-vector B such that the approximated Hessian W_{NT} and the approximated score C_{NT} satisfy $W_{NT} \to_p W$, and $C_{NT} \to_d C$, where $C \sim \mathcal{N}(B, \Omega)$.

(iii) We assume that the estimator $\hat{\beta}$ that minimizes $L_{NT}(\beta)$ subject to $\beta \in \mathbb{B}$ is consistent for β^0.

(iv) We have estimators $\hat{W}(\beta)$, $\hat{\Omega}(\beta)$ and $\hat{B}(\beta)$ that are consistent for W, Ω and B when evaluated for any $\sqrt{NT}$-consistent estimator of β^0.

Assumption 7 can be satisfied in the interactive fixed effect model for different estimators of W, Ω and B, and under different assumptions on λ^0, f^0, X_k and e. In the last subsection we presented a concrete example for which the assumption holds, namely for the estimators in definition 4.3, and under the assumptions of corollary 4.5, but for assumption 7 to be satisfied it is not necessary that β^0 is an interior point of $\mathbb{B}$.

In this section we want to discuss the limiting distribution of the QMLE for cases where β^0 is on the boundary of the parameter set $\mathbb{B}$. More specifically, we consider the case where $\mathbb{B} - \beta^0$ is locally approximated by a convex cone $\Lambda \subset \mathbb{R}^K$.[19]

[19] We refer to Andrews (1999) for the definition of "locally approximated". A special case is when $\mathbb{B} - \beta^0$ is locally equal to a cone $\Lambda \subset \mathbb{R}^K$, *i.e.* if there exists $\epsilon > 0$ such that $B(0,\epsilon) \cap (\mathbb{B} - \beta^0) = B(0,\epsilon) \cap \Lambda$, where $B(0,\epsilon)$ is the ball with radius ϵ around the origin. Remember that $\Lambda \subset \mathbb{R}^K$ is a cone iff $az \in \Lambda$ for every $a > 0$ and $z \in \Lambda$, *i.e.* it is invariant under rescalings with positive scaling factor that are centered at the origin. Whenever $\beta^0 \in \mathbb{B}$ and $\mathbb{B}$ is defined by equality and inequality constraints on linear combinations of β we find that $\mathbb{B} - \beta^0$ is locally equal to a convex cone. Under non-linear equality and inequality constraints one usually finds $\mathbb{B} - \beta^0$ is locally approximated by a convex cone $\Lambda \subset \mathbb{R}^K$.

When β^0 is on the boundary of the parameter set it is not guaranteed that the bias corrected estimator $\hat{\beta}^*$ defined in corollary 4.5 satisfies $\hat{\beta}^* \in \mathbb{B}$ asymptotically. We therefore define an alternative "bias corrected" estimator by

$$\hat{\beta}^{**} = \underset{\beta\in\mathbb{B}}{\operatorname{argmin}}\, L_{NT}^{**}(\beta)\,, \qquad L_{NT}^{**}(\beta) = L_{NT}\left[\beta + (NT)^{-1/2}\,\hat{W}^{-1}(\hat{\beta})\,\hat{B}(\hat{\beta})\right]\,, \tag{4.7}$$

where $\hat{\beta}$ is the QMLE that minimizes $L_{NT}(\beta)$ subject to $\beta \in \mathbb{B}$, *i.e.* $\hat{\beta}^{**}$ is defined by a two-step minimization procedure. The estimator $\hat{\beta}^{**}$ is bias corrected in the sense that its limiting distribution is the one that the QMLE $\hat{\beta}$ would have if the asymptotic bias of the score would be vanishing, *i.e.* if $B = 0$. However, $\hat{\beta}^{**}$ usually has an asymptotic bias since its limiting distribution is a projection (or truncation) of a multivariate normal distribution, as described in the theorem below.

In order to describe the limiting distributions of $\hat{\beta}$ and $\hat{\beta}^{**}$ it is convenient to introduce the function $l_q(\phi) = \phi' W \phi - 2\phi' C$ for $\phi \in \mathbb{R}^K$. For all $\phi \in \mathbb{R}^K$ we find that under assumption 7 we have $NT\left[L_{NT}(\beta_{NT}) - L_{NT}(\beta^0)\right] \to_d l_q(\phi)$ for $\beta_{NT} = \beta^0 + (NT)^{-1/2}\phi$. Thus, $l_q(\phi)$ is the limit of the appropriately rescaled profile quasi likelihood function when holding $\phi = \sqrt{NT}(\beta - \beta^0)$ fixed.

Theorem 4.6. *Let assumption 7 be satisfied and let $\mathbb{B} - \beta^0$ be locally approximated by a closed convex cone $\Lambda \subset \mathbb{R}^K$. Define the random variables $\Phi = \operatorname{argmin}_{\phi\in\Lambda} l_q(\phi)$, and $\Phi^{**} = \operatorname{argmin}_{\phi\in\Lambda} l_q(\phi + W^{-1}B)$. Then*

$$\begin{aligned}
\sqrt{NT}\left(\hat{\beta} - \beta^0\right) &\underset{d}{\to} \Phi\,,\\
\sqrt{NT}\left(\hat{\beta}^{**} - \beta^0\right) &\underset{d}{\to} \Phi^{**}\,,\\
NT\left[L_{NT}\left(\hat{\beta}\right) - L_{NT}\left(\beta^0\right)\right] &\underset{d}{\to} l_q(\Phi)\,,\\
NT\left[L_{NT}^{**}\left(\hat{\beta}^{**}\right) - L_{NT}^{**}\left(\beta^0\right)\right] &\underset{d}{\to} l_q(\Phi^{**} + W^{-1}B) - l_q(W^{-1}B)\,.
\end{aligned}$$

Theorem 4.6 is a special case of theorem 3 in Andrews (1999). Although Andrews does not explicitly consider bias correction, it is easy to check that both objective functions $L_{NT}(\beta)$ and $L_{NT}^{**}(\beta)$ satisfy the assumptions necessary to apply Andrews' theorem for the limiting distributions.

By writing the limiting distribution of the approximated score as $C = B + \Omega^{1/2}\mathcal{Z}_K$, where $\mathcal{Z}_K$ is a K-dimensional standard normal distribution, we can give slightly more explicit expressions for Φ and Φ^{**}, namely

$$\begin{aligned}
\Phi &= \underset{\phi\in\Lambda}{\operatorname{argmin}}\left[\phi - W^{-1}(B + \Omega^{1/2}\mathcal{Z}_K)\right]' W \left[\phi - W^{-1}(B + \Omega^{1/2}\mathcal{Z}_K)\right]\,,\\
\Phi^{**} &= \underset{\phi\in\Lambda}{\operatorname{argmin}}\left[\phi - W^{-1}\Omega^{1/2}\mathcal{Z}_K\right]' W \left[\phi - W^{-1}\Omega^{1/2}\mathcal{Z}_K\right]\,.
\end{aligned} \tag{4.8}$$

Thus, the asymptotic distribution of $\sqrt{NT}(\hat{\beta} - \beta^0)$ is given by the orthogonal projection (relative to the metric W) of $W^{-1}(B + \Omega^{1/2}\mathcal{Z}_K) \sim \mathcal{N}(W^{-1}B,\, W^{-1}\Omega W^{-1})$ onto the cone Λ. For interior points of Λ the distribution of $\sqrt{NT}(\hat{\beta} - \beta^0)$ is the same as for $\mathcal{N}(W^{-1}B,\, W^{-1}\Omega W^{-1})$, but for a point on the boundary of Λ the distribution is given by an integral over those points that are projected on this point. The distribution for $\sqrt{NT}(\hat{\beta}^{**} - \beta^0)$ is given by almost the same formula, but without bias B. In the one-dimensional case ($K = 1$) the only non-trivial closed cones are $\Lambda = [0, \infty)$ and $\Lambda = (-\infty, 0]$, *i.e.* the distributions of $\sqrt{NT}(\hat{\beta} - \beta^0)$ and $\sqrt{NT}(\hat{\beta}^{**} - \beta^0)$ are truncated normal distributions.

4.3 Testing restriction on β^0

For our interactive fixed effect model, we now want to discuss the three classical test statistics for testing a general linear restriction 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 a $r \times K$ matrix of rank $r \leq K$, and h is a $r \times 1$ vector. 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.

For ease of exposition we restrict the presentation to testing a linear hypothesis, but using the tools provided above one can generalize the discussion to the testing of non-linear hypotheses. Using the expansion $L_{NT}(\beta)$ one could also discuss testing when the true parameter is on the boundary, as shown in Andrews (2001).

The unrestricted and restricted estimators are defined by

$$\hat{\beta} = \underset{\beta \in \mathbb{B}}{\operatorname{argmin}}\ L_{NT}(\beta)\ , \qquad \tilde{\beta} = \underset{\beta \in \tilde{\mathbb{B}}}{\operatorname{argmin}}\ L_{NT}(\beta)\ , \tag{4.9}$$

where $\tilde{\mathbb{B}} = \{\beta \in \mathbb{B} | \, H\beta = h\}$ is the restricted parameter set.

Under assumption 7 the limiting distribution of the restricted and unrestricted estimator is given by

$$\sqrt{NT}\left(\hat{\beta} - \beta^0\right) \ \underset{d}{\longrightarrow}\ \mathcal{N}\left(W^{-1}B,\ W^{-1}\,\Omega\,W^{-1}\right)\ , \quad \sqrt{NT}(\tilde{\beta} - \beta^0) \ \underset{d}{\longrightarrow}\ \mathcal{N}\left(\mathfrak{W}^{-1}B,\ \mathfrak{W}^{-1}\,\Omega\,\mathfrak{W}^{-1}\right)\ , \tag{4.10}$$

where $\mathfrak{W}^{-1} = W^{-1} - W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}$.[20]

For the unrestricted estimator this result was given in theorem 4.1 for a specific set of assumptions on λ^0, f^0, X_k and e, but it holds whenever assumption 7 is satisfied.[21] For the restricted estimator we note that $\tilde{\mathbb{B}} - \beta^0$ is locally equal to the r-dimensional subspace $\Lambda = \{\phi \in \mathbb{R}^K | \, H\phi = 0\}$, which is a special case of a convex cone, *i.e.* we can apply theorem 4.6 to obtain the limiting distribution of $\tilde{\beta}$.[22]

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

$$WD^*_{NT} = \left[\sqrt{NT}\left(H\hat{\beta} - h\right) - H\hat{W}^{-1}\hat{B}\right]'\left(H\hat{W}^{-1}\,\hat{\Omega}\,\hat{W}^{-1}H'\right)^{-1}\left[\sqrt{NT}\left(H\hat{\beta} - h\right) - H\hat{W}^{-1}\hat{B}\right]\ , \tag{4.11}$$

and under the null hypothesis we find $WD^*_{NT} \rightarrow_d \chi^2_r$ if assumption 7 is 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

For the discussion of the LR test we have to assume 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 there is no heteroscedasticity. 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 $\hat{e} = \hat{e}(\hat{\beta})$, and since the likelihood function for the interactive fixed effect model is just the sum of squared residuals we have $\hat{c} = L_{NT}\left(\hat{\beta}\right)$. However, different estimators for c can be used.

---

[20] The $K \times K$ covariance matrix in the limiting distribution of $\tilde{\beta}$ is not full rank, but satisfies $\operatorname{rank}(\mathfrak{W}^{-1}\,\Omega\,\mathfrak{W}^{-1}) = K - r$, because $H\mathfrak{W}^{-1} = 0$. The asymptotic distribution of $\sqrt{NT}(\tilde{\beta} - \beta^0)$ is therefore $K - r$ dimensional.

[21] The unrestricted case is the special case of theorem 4.6, namely $\Lambda = \mathbb{R}^K$, *i.e.* $\Phi = W^{-1}C$.

[22] One finds $\sqrt{NT}(\tilde{\beta} - \beta^0) \rightarrow_d \tilde{\Phi}$, with $\tilde{\Phi} = \operatorname{argmin}_\Lambda l_q(\phi) = M_{W,H'}W^{-1}C = \mathfrak{W}^{-1}C$, where $M_{W,H'} = \mathbb{I}_K - W^{-1}H'(HW^{-1}H')^{-1}H$ is the orthogonal projector onto the subspace Λ with respect to the metric W. One can easily check that the projector $M_{W,H'}$ as given here has all the required properties, namely $HM_{W,H'} = 0$ (thus, $(M_{W,H'}\phi) \in \Lambda$ for all $\phi \in \mathbb{R}^K$), $(M_{W,H'})^2 = M_{W,H'}$ (idempotence), $\operatorname{Tr}(M_{W,H'}) = K - r$ (projector on $K - r$ dimensional subspace), and $M'_{W,H'}W(\mathbb{I}_K - M_{W,H'}) = 0$ (orthogonality wrt to W). Note that $M_{W,H'} = M_{H'}$ if $W = \mathbb{I}_K$.

The likelihood ratio test statistics is defined by

$$LR_{NT} = \hat{c}^{-1}\, NT \left[L_{NT}\left(\tilde{\beta}\right) - L_{NT}\left(\hat{\beta}\right)\right] . \tag{4.12}$$

Applying theorem 4.6 we find that under assumption 7 we have

$$\begin{aligned} LR_{NT} \;\underset{d}{\longrightarrow}\; & c^{-1}\left[l(\tilde{\Phi}) - l(\Phi)\right] = c^{-1}\left[l(\mathfrak{W}^{-1}C) - l(W^{-1}C)\right] \\ & = c^{-1}C'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}C\,. \end{aligned} \tag{4.13}$$

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

It is natural to base the bias corrected LR test on the objective function L^{**}_{NT} used above to define the biased corrected estimator β^{**}. Thus, we define

$$LR^{*}_{NT} = \hat{c}^{-1} NT \left[\min_{\{\beta\in\mathbb{B}|\, H\beta = h\}} L_{NT}\left(\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}\right) - \min_{\beta\in\mathbb{B}} L_{NT}\left(\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}\right)\right] , \tag{4.14}$$

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.[23] 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 $\beta + (NT)^{-1/2}\hat{W}^{-1}\hat{B}$, because generically we have $HW^{-1}B \neq 0$ (otherwise no correction would be necessary for the LR statistics).

Using theorem 4.6 one finds

$$\begin{aligned} LR^{*}_{NT} \;\underset{d}{\longrightarrow}\; & c^{-1}\left[\min_{\{\phi\in\mathbb{R}^K|\, H\phi = 0\}} l\left(\phi + W^{-1}B\right) - l\left(\Phi\right)\right] \\ & = c^{-1}\left[l\left(\mathfrak{W}^{-1}(C-B) + W^{-1}B\right) - l\left(W^{-1}C\right)\right] \\ & = c^{-1}(C-B)'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}(C-B)\,, \end{aligned} \tag{4.15}$$

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 assumption 7 and 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.

$$\begin{aligned} \tilde{\nabla}\mathcal{L}_{NT} \;\equiv\; \nabla\mathcal{L}_{NT}(\tilde{\beta},\tilde{\lambda},\tilde{f}) &= \left(\left.\frac{\partial\mathcal{L}_{NT}(\beta,\tilde{\lambda},\tilde{f})}{\partial\beta_1}\right|_{\beta=\tilde{\beta}}, \ldots, \left.\frac{\partial\mathcal{L}_{NT}(\beta,\tilde{\lambda},\tilde{f})}{\partial\beta_K}\right|_{\beta=\tilde{\beta}}\right)' \\ &= -\frac{2}{NT}\Big(\mathrm{Tr}\left(X_1'\tilde{e}\right), \ldots, \mathrm{Tr}\left(X_K'\tilde{e}\right)\Big)' , \end{aligned} \tag{4.16}$$

[23] 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.

where $\tilde{e} = \hat{e}(\tilde{\beta})$ (for the definition of $\hat{e}(\beta)$ see equation (3.8)). Under assumptions 5 and 6, and if the null hypothesis $H_0 : \; H\beta^0 = h$ is satisfied, one finds that[24]

$$\begin{aligned}\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} &= \sqrt{NT}\,\nabla L_{NT}(\tilde{\beta}) + o_p(1) \\ &= \sqrt{NT}\,\nabla L_{q,NT}(\tilde{\beta}) + o_p(1) \\ &= 2\sqrt{NT}\,W_{NT}\,(\tilde{\beta} - \beta^0) - 2\,C_{NT} + o_p(1)\;. \end{aligned} \tag{4.17}$$

Due to the first line of the last 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. That one can also replace $\nabla L_{NT}(\tilde{\beta})$ by its approximation $\nabla L_{q,NT}(\tilde{\beta})$ is a consequence of theorem 3.2 and the fact that $\tilde{\beta}$ is $\sqrt{NT}$-consistent under the Null. Using this result and the known limiting distribution of $\tilde{\beta}$ we find

$$\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} \;\underset{d}{\longrightarrow}\; -2H'(HW^{-1}H')^{-1}HW^{-1}C\;. \tag{4.18}$$

The LM test statistics is given by[25]

$$LM_{NT} = \frac{NT}{4}\,(\tilde{\nabla}\mathcal{L}_{NT})'\tilde{W}^{-1}\tilde{H}'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}HW^{-1}\tilde{\nabla}\mathcal{L}_{NT}\;, \tag{4.19}$$

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

$$LM^*_{NT} = \frac{1}{4}\,\left(\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} + 2\tilde{B}\right)'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}\left(\sqrt{NT}\,\tilde{\nabla}\mathcal{L}_{NT} + 2\tilde{B}\right)\;. \tag{4.20}$$

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

Theorem 4.7. *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 4.3. For the bias corrected Wald and LM test statistics introduced in equation* (4.11) *and* (4.20) *we then have*

$$WD^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r\;, \qquad\qquad LM^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r\;. \tag{4.21}$$

If in addition we assume $\mathbb{E}e^2_{it} = c$, i.e. the idiosyncratic errors are homoscedastic, and we use $\hat{c} = L_{NT}\left(\hat{\beta}\right)$ as an estimator for c, then the LR test statistics defined in equation (4.14) *satisfies*

$$LR^*_{NT} \;\underset{d}{\longrightarrow}\; \chi^2_r\;. \tag{4.22}$$

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),[26] and the bias corrected QMLE (BC-QMLE) defined in theorem 4.5.

[24] The proof of the statement is given in the appendix as part of the proof of theorem 4.7.

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

[26] 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.

For the simulation we draw e_{it} independently distributed from $\mathcal{N}(0,1)$, the λ_i^0 independently distributed from $\mathcal{N}(1,1)$, and we generate the factors from an AR(1) specification, *i.e.* $f_t^0 = \rho_f f_{t-1}^0 + u_t$, where $u_t \sim \text{iid}\mathcal{N}(0,(1-\rho_f^2)\sigma_f^2)$, and σ_f is the standard deviation of f_t.[27] In this 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 $\|e\| \approx \sqrt{N} + \sqrt{T}$.[28] 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 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

[27] For all simulations we generate 1000 initial time periods for f_t^0 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_t^0 and λ_i^0 made here do not satisfy assumption 5(i), but nevertheless all theorems above are applicable since f_t^0 and λ_i^0 have arbitrary high uniformly bounded moments.

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

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

For the interactive fixed effect model (2.1) we provide a methodology that uses the perturbation theory of linear operators to expand the profile quasi likelihood function $L_{NT}(\beta)$ around the true regression parameter β^0. In particular, we work out the quadratic expansion of $L_{NT}(\beta)$ and show how it can be used to derive the first order asymptotic theory of the QMLE of β under the alternative asymptotic $N, T \to \infty$. It is found that the QMLE can be asymptotically biased (i) due to weak exogeneity of the regressors and (ii) due to correlation and heteroscedasticity of the idiosyncratic errors e_{it}. We also provide expansions of the projectors $M_{\hat{f}}$ and $M_{\hat{\lambda}}$, and of the residuals $\hat{e}$ in the the regression parameters that are very useful when working with these estimators, *e.g.* when proving consistency of the asymptotic bias and variance estimators of $\hat{\beta}$.

As applications of our general methodology, we work out the limiting distribution of the QMLE $\hat{\beta}$ under the assumption of independent error terms 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 also discuss the asymptotic distribution of the QMLE when the true parameter is on the boundary of the parameter set. Finally, we derive the asymptotic distribution of the Wald, LR and LM test statistics, which are not χ^2 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 χ^2.

The results 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.

The most important extension of the present paper will be to study the case where only an upper bound on the number of factors is know. The goal then is (i) to derive the limiting distribution of the QMLE when the number of factors is overestimated and (ii) to estimate the number of factors consistently. In the supplementary material of Bai (2009) the idea of approaching the second goal is explained, but the key step here will be to show $\sqrt{NT}$-consistency of the QMLE when only an upper bound on the number of factors is known. We hope to successfully address these issues in future research.

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}, $i = 1 \dots N$, $t = -\infty \dots T$ 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

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

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

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

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:*

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

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 still satisfied $M_\lambda \lambda = 0$ and $\mathrm{rank}(M_\lambda) = N - \mathrm{rank}(\lambda)$. However, if $\mathrm{rank}(Z) \geq R$ then the optimal λ, f, $\tilde{\lambda}$ and $\tilde{f}$ 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 $\mathrm{Tr}(M_{\tilde{\lambda}}\, Z\, M_{\tilde{f}}\, Z')$ 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*

$$\sup_f \left|\frac{\mathrm{Tr}(X_k\, M_f\, e')}{NT}\right| = o_p(1)\,, \quad \sup_f \left|\frac{\mathrm{Tr}(\lambda^0\, f^{0\prime}\, M_f\, e')}{NT}\right| = o_p(1)\,, \quad \sup_f \left|\frac{\mathrm{Tr}(e\, P_f\, e')}{NT}\right| = o_p(1)\,, \tag{B.2}$$

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

Proof. By assumption 2 we know that the first two equations in Lemma B.2 are satisfied when replacing M_f by the identity matrix. So we are left to show $\max_f \left|\frac{1}{NT}\mathrm{Tr}(\Xi\, P_f\, e')\right| = o_p(1)$, where Ξ is either X_k,

$\lambda^0 f^{0\prime}$, or e. In all three cases we have $\|\Xi\|/\sqrt{NT} = \mathcal{O}_p(1)$, by assumptions 1, 3, and 4, respectively. We therefore find[29]

$$\sup_f \left| \frac{1}{NT} \mathrm{Tr}(\Xi\, P_f\, e') \right| \leq R\, \frac{\|e\|}{\sqrt{NT}}\, \frac{\|\Xi\|}{\sqrt{NT}} = o_p(1)\ . \tag{B.3}$$

∎

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

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

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

$$\begin{aligned} S_{NT}(\beta, f) &= S_{NT}(\beta^0, f^0) + \tilde{S}_{NT}(\beta, f) \\ &\quad + \frac{2}{NT}\ \mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)\ M_f\, e' \right] + \frac{1}{NT}\ \mathrm{Tr}\left(e\, (P_{f^0} - P_f)\, e'\right) \\ &= S_{NT}(\beta^0, f^0) + \tilde{S}_{NT}(\beta, f) + o_p(\|\beta - \beta^0\|) + o_p(1)\ , \end{aligned} \tag{B.5}$$

where we defined

$$\tilde{S}_{NT}(\beta, f) = \frac{1}{NT}\ \mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)\ M_f\ \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)' \right]. \tag{B.6}$$

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 use a different condition on high-rank regressors that does not contain f^0 and λ^0. We split $\tilde{S}_{NT}(\beta, f) = \tilde{S}^{(1)}_{NT}(\beta, f) + \tilde{S}^{(2)}_{NT}(\beta, f)$, where

$$\begin{aligned} \tilde{S}^{(1)}_{NT}(\beta, f) &= \frac{1}{NT}\ \mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)\ M_f\ \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)'\ M_{(\lambda_0, w)} \right] \\ &= \frac{1}{NT}\ \mathrm{Tr}\left[\left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right)\ M_f\ \left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right)'\ M_{(\lambda_0, w)} \right], \\ \tilde{S}^{(2)}_{NT}(\beta, f) &= \frac{1}{NT}\ \mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)\ M_f\ \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k \right)'\ P_{(\lambda_0, w)} \right], \end{aligned} \tag{B.7}$$

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

$$\begin{aligned} \tilde{S}^{(1)}_{NT}(\beta, f) &\geq \frac{1}{NT} \sum_{i=2R+K_1+1}^{N} \mu_i \left[\left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right) \left(\sum_{m=K_1+1}^{K} (\beta_m^0 - \beta_m) X_m \right)' \right] \\ &\geq b\, \left\| \beta^{\text{high}} - \beta_0^{\text{high}} \right\|^2\ , \qquad \text{wpa1}, \end{aligned} \tag{B.8}$$

[29] Here we use $|\mathrm{Tr}\,(C)| \leq \|C\|\, \mathrm{rank}\,(C)$, which holds for all square matrices C.

where in the last step we used the existence of a constant $b > 0$ guaranteed by assumption 4(b)(i), 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 $f = f^0$ to obtain $\tilde{S}_{NT}(\beta^0, f^0) = 0$, *i.e.* the optimal $\beta = \hat{\beta}$ and $f = \hat{f}$ must satisfy $\tilde{S}_{NT}(\hat{\beta}, \hat{f}) \leq 0$. Using this and equation B.5 we find

$$0 \;\geq\; b \left\| \hat{\beta}^{\text{high}} - \beta_0^{\text{high}} \right\|^2 + o_p\left(\left\| \hat{\beta}^{\text{high}} - \beta_0^{\text{high}} \right\| \right) + o_p\left(\left\| \hat{\beta}^{\text{low}} - \beta_0^{\text{low}} \right\| \right) + o_p(1) \; . \tag{B.9}$$

Since we assume that $\hat{\beta}^{\text{low}}$ is bounded the last equation implies that $\left\| \hat{\beta}^{\text{high}} - \beta_0^{\text{high}} \right\| = o_p(1)$, *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(b)(ii) guarantees that there exist finite positive constants a_0, a_1, a_2, a_3 and a_4 such that

$$\begin{aligned} \tilde{S}^{(2)}_{NT}(\beta, f) \geq \; & \frac{a_0 \left\| \beta^{\text{low}} - \beta_0^{\text{low}} \right\|^2}{\left\| \beta^{\text{low}} - \beta_0^{\text{low}} \right\|^2 + a_1 \left\| \beta^{\text{low}} - \beta_0^{\text{low}} \right\| + a_2} \\ & - a_3 \left\| \beta^{\text{high}} - \beta_0^{\text{high}} \right\| - a_4 \left\| \beta^{\text{high}} - \beta_0^{\text{high}} \right\| \left\| \beta^{\text{low}} - \beta_0^{\text{low}} \right\| \; , \qquad \text{wpa1.} \end{aligned} \tag{B.10}$$

Using consistency of $\hat{\beta}^{\text{high}}$ and again boundedness of β^{low} this implies that there exists $a > 0$ such that $\tilde{S}^{(2)}_{NT}(\hat{\beta}, f) \geq a \left\| \hat{\beta}^{\text{low}} - \beta_0^{\text{low}} \right\|^2 + o_p(1)$. With the same argument as for $\hat{\beta}^{\text{high}}$ we therefore find $\left\| \hat{\beta}^{\text{low}} - \beta_0^{\text{low}} \right\| = o_p(1)$, *i.e.* $\hat{\beta}^{\text{low}}$ is consistent. This is what we wanted to show. ∎

C Power Series Expansion of the Profile Quasi Likelihood Function (Proofs of Theorems 3.1, 3.2 and 3.3)

Definition C.1. *For the $N \times R$ matrix λ^0 and the $T \times R$ matrix f^0 we define*

$$\begin{aligned} d_{\max}(\lambda^0, f^0) &= \frac{1}{\sqrt{NT}} \left\| \lambda^0 f^{0\prime} \right\| = \frac{1}{\sqrt{NT}} \sqrt{\mu_1(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \; , \\ d_{\min}(\lambda^0, f^0) &= \frac{1}{\sqrt{NT}} \sqrt{\mu_R(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \; , \end{aligned} \tag{C.1}$$

i.e. $d_{\max}(\lambda^0, f^0)$ and $d_{\min}(\lambda^0, f^0)$ are the square roots of the maximal and the minimal eigenvalue of $\lambda^{0\prime} f^0 f^{0\prime} \lambda^0 / NT$. Furthermore, the convergence radius $r_0(\lambda^0, f^0)$ is defined by

$$r_0(\lambda^0, f^0) = \left(\frac{4 d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} + \frac{1}{2 d_{\max}(\lambda^0, f^0)} \right)^{-1} \; . \tag{C.2}$$

Theorem C.2. *If the following condition is satisfies*

$$\sum_{k=1}^{K} \left| \beta_k^0 - \beta_k \right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} \;<\; r_0(\lambda^0, f^0) \; , \tag{C.3}$$

then

(i) the profile quasi likelihood function can be written as a power series in the $K+1$ parameters $\epsilon_0 = \|e\|/\sqrt{NT}$ and $\epsilon_k = \beta_k^0 - \beta_k$, namely

$$L_{NT}\left(\beta\right) = \frac{1}{NT} \sum_{g=2}^{\infty} \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \dots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1}\, \epsilon_{\kappa_2}\, \dots\, \epsilon_{\kappa_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) \,, \quad \text{(C.4)}$$

where the expansion coefficients are given by[30]

$$\begin{aligned} L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) &= \tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{(\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g)}\right) \\ &= \frac{1}{g!}\left[\tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) + \textit{all permutations of } \kappa_1, \dots, \kappa_g\right] \,, \end{aligned} \quad \text{(C.5)}$$

i.e. $L^{(g)}$ is obtained by total symmetrization of the last g arguments of[31]

$$\begin{aligned} &\tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) \\ &\qquad = \sum_{p=1}^{g} (-1)^{p+1} \sum_{\substack{\nu_1 + \ldots + \nu_p = g \\ l_1 + \ldots + l_{p+1} = p-1 \\ 2 \geq \nu_j \geq 1\,,\; l_j \geq 0}} \mathrm{Tr}\left(S^{(l_1)}\, \mathcal{T}^{(\nu_1)}_{\kappa_1 \dots}\, S^{(l_2)}\, \dots\, S^{(l_p)}\, \mathcal{T}^{(\nu_p)}_{\dots \kappa_g}\, S^{(l_{p+1})}\right) \,, \end{aligned} \quad \text{(C.6)}$$

with

$$\begin{aligned} &S^{(0)} = -M_{\lambda^0} \,, && S^{(l)} = \left[\lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime}\right]^l \,, \quad \textit{for } l \geq 1, \\ &\mathcal{T}^{(1)}_{\kappa} = \lambda^0\, f^{0\prime}\, X'_{\kappa} + X_{\kappa}\, f^0\, \lambda^{0\prime} \,, \quad \mathcal{T}^{(2)}_{\kappa_1 \kappa_2} = X_{\kappa_1}\, X'_{\kappa_2} \,, && \textit{for } \kappa, \kappa_1, \kappa_2 = 0 \dots K \,, \\ &X_0 = \frac{\sqrt{NT}}{\|e\|}\, e \,, && X_{\kappa} = X_k \,, \qquad \textit{for } \kappa = k = 1 \dots K \,. \end{aligned} \quad \text{(C.7)}$$

(ii) the projector $M_{\hat{\lambda}}(\beta)$ can be written as a power series in the same parameters ϵ_κ ($\kappa = 0, \dots, K$), namely

$$M_{\hat{\lambda}}\left(\beta\right) = \sum_{g=0}^{\infty} \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \dots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1}\, \epsilon_{\kappa_2}\, \dots\, \epsilon_{\kappa_g}\, M^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) \,, \quad \text{(C.8)}$$

where the expansion coefficients are given by $M^{(0)}(\lambda^0,\, f^0) = M_{\lambda^0}$, and for $g \geq 1$

$$\begin{aligned} M^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) &= \tilde{M}^{(g)}\left(\lambda^0,\, f^0,\, X_{(\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g)}\right) \\ &= \frac{1}{g!}\left[\tilde{M}^{(g)}\left(X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) + \textit{all permutations of } \kappa_1, \dots, \kappa_g\right] \,, \end{aligned} \quad \text{(C.9)}$$

i.e. $M^{(g)}$ is obtained by total symmetrization of the last g arguments of

$$\begin{aligned} &\tilde{M}^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) \\ &\qquad = \sum_{p=1}^{g} (-1)^{p+1} \sum_{\substack{\nu_1 + \ldots + \nu_p = g \\ l_1 + \ldots + l_{p+1} = p \\ 2 \geq \nu_j \geq 1\,,\; l_j \geq 0}} S^{(l_1)}\, \mathcal{T}^{(\nu_1)}_{\kappa_1 \dots}\, S^{(l_2)}\, \dots\, S^{(l_p)}\, \mathcal{T}^{(\nu_p)}_{\dots \kappa_g}\, S^{(l_{p+1})} \,, \end{aligned} \quad \text{(C.10)}$$

where $S^{(k)}$, $\mathcal{T}^{(1)}_{\kappa}$, $\mathcal{T}^{(2)}_{\kappa_1\kappa_2}$, and X_κ are given above.

[30] Here we use the round bracket notation $(\kappa_1, \kappa_2, \dots, \kappa_g)$ for total symmetrization of these indices.

[31] One finds $\tilde{L}^{(1)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \dots, X_{\kappa_g}\right) = 0$, which is why the sum in the power series of L_{NT} starts from $g = 2$ instead of $g = 1$.

(iii) The coefficients $L^{(g)}$ in the series expansion of $L_{NT}(\beta)$ are bounded as follows

$$\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2},\ldots,X_{\kappa_g}\right)\right| \\ \leq \frac{R\,g\,d^2_{\min}(\lambda^0,f^0)}{2}\left(\frac{16\;d_{\max}(\lambda^0,f^0)}{d^2_{\min}(\lambda^0,f^0)}\right)^g \frac{\|X_{\kappa_1}\|}{\sqrt{NT}}\,\frac{\|X_{\kappa_2}\|}{\sqrt{NT}}\,\cdots\,\frac{\|X_{\kappa_g}\|}{\sqrt{NT}} \tag{C.11}$$

Under the stronger condition

$$\sum_{k=1}^{K}\left|\beta^0_k-\beta_k\right|\frac{\|X_k\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}} \;<\; \frac{d^2_{\min}(\lambda^0,f^0)}{16\;d_{\max}(\lambda^0,f^0)}\;, \tag{C.12}$$

we therefore have the following bound on the remainder when the series expansion for $L_{NT}(\beta)$ is truncated at order $G\geq 2$:

$$\left|L_{NT}\left(\beta\right)-\frac{1}{NT}\sum_{g=2}^{G}\sum_{\kappa_1=0}^{K}\cdots\sum_{\kappa_g=0}^{K}\epsilon_{\kappa_1}\,\ldots\,\epsilon_{\kappa_g}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_{\kappa_1},\,X_{\kappa_2},\ldots,X_{\kappa_g}\right)\right| \\ \leq \frac{R\,(G+1)\,\alpha^{G+1}\,d^2_{\min}(\lambda^0,f^0)}{2(1-\alpha)^2}\;, \tag{C.13}$$

where

$$\alpha \;=\; \frac{16\;d_{\max}(\lambda^0,f^0)}{d^2_{\min}(\lambda^0,f^0)}\left(\sum_{k=1}^{K}\left|\beta^0_k-\beta_k\right|\frac{\|X_k\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)\;<1\;. \tag{C.14}$$

(iv) The operator norm of the coefficient $M^{(g)}$ in the series expansion of $M_{\hat{\lambda}}\left(\beta\right)$ is bounded as follows, for $g\geq 1$

$$\left\|M^{(g)}\left(\lambda^0,\,f^0,\,X_{\kappa_1},\,X_{\kappa_2},\ldots,X_{\kappa_g}\right)\right\| \leq \frac{g}{2}\left(\frac{16\;d_{\max}(\lambda^0,f^0)}{d^2_{\min}(\lambda^0,f^0)}\right)^g \frac{\|X_{\kappa_1}\|}{\sqrt{NT}}\,\frac{\|X_{\kappa_2}\|}{\sqrt{NT}}\,\cdots\,\frac{\|X_{\kappa_g}\|}{\sqrt{NT}}\;. \tag{C.15}$$

Under the condition (C.12) *we therefore have the following bound on operator norm of the remainder of the series expansion of $M_{\hat{\lambda}}\left(\beta\right)$, for $G\geq 0$*

$$\left\|M_{\hat{\lambda}}\left(\beta\right)-\sum_{g=0}^{G}\sum_{\kappa_1=0}^{K}\cdots\sum_{\kappa_g=0}^{K}\epsilon_{\kappa_1}\,\ldots\,\epsilon_{\kappa_g}\,M^{(g)}\left(\lambda^0,\,f^0,\,X_{\kappa_1},\,X_{\kappa_2},\ldots,X_{\kappa_g}\right)\right\| \leq \frac{(G+1)\,\alpha^{G+1}}{2(1-\alpha)^2}\;. \tag{C.16}$$

Proof.

(i,ii) We apply perturbation theory in Kato (1980). The unperturbed operator is $\mathcal{T}^{(0)}=\lambda^0\lambda^{0\prime}$, the perturbed operator is $\mathcal{T}=\mathcal{T}^{(0)}+\mathcal{T}^{(1)}+\mathcal{T}^{(2)}$ (*i.e.* the parameter κ that appears in Kato is set to 1), where $\mathcal{T}^{(1)}=\sum_{\kappa=0}^{K}\epsilon_\kappa X_\kappa f^0\lambda^{0\prime}+\lambda^0 f^{0\prime}\sum_{\kappa=0}^{K}\epsilon_\kappa X'_\kappa$, and $\mathcal{T}^{(2)}=\sum_{\kappa_1=0}^{K}\sum_{\kappa_2=0}^{K}\epsilon_{\kappa_1}\epsilon_{\kappa_2}X_{\kappa_1}X'_{\kappa_2}$. The matrices $\mathcal{T}$ and $\mathcal{T}^0$ are real and symmetric (which implies that they are normal operators), and positive semi-definite. We know that $\mathcal{T}^{(0)}$ has an eigenvalue 0 with multiplicity $N-R$, and the separating distance of this eigenvalue is $d=NTd^2_{\min}(\lambda^0,f^0)$. The bound (C.3) guarantees that

$$\|\mathcal{T}^{(1)}+\mathcal{T}^{(2)}\|\leq\frac{NT}{2}\,d^2_{\min}(\lambda^0,f^0)\;, \tag{C.17}$$

by Weyl's inequality we therefore find that the $N-R$ smallest eigenvalues of $\mathcal{T}$ (also counting multiplicity) are all smaller than $\frac{NT}{2}d^2_{\min}(\lambda^0,f^0)$, and they "originate" from the zero-eigenvalue

of $\mathcal{T}^{(0)}$, with the power series expansion for $L_{NT}(\beta)$ given in (2.22) and (2.18) at p.77/78 of Kato, and the expansion of $M_{\hat{\lambda}}$ given in (2.3) and (2.12) at p.75,76 of Kato. We still need to justify the convergence radius of this series. Since we set the complex parameter κ in Kato to 1, we need to show that the convergence radius (r_0 in Kato's notation) is at least 1. The condition (3.7) in Kato p.89 reads $\|\mathcal{T}^{(n)}\| \leq ac^{n-1}$, $n = 1, 2, \ldots$, and it is satisfied for $a = 2\sqrt{NT}d_{\max}(\lambda^0, f^0)\sum_{\kappa=0}^{K}|\epsilon_\kappa|\|X_\kappa\|$ and $c = \sum_{\kappa=0}^{K}|\epsilon_\kappa|\|X_\kappa\|/\sqrt{NT}/2/d_{\max}(\lambda^0, f^0)$. According to equation (3.51) in Kato p.95, we therefore find that the power series for $L_{NT}(\beta)$ and $M_{\hat{\lambda}}$ are convergent ($r_0 \geq 1$ in his notation) if $1 \leq \left(\frac{2a}{d} + c\right)^{-1}$, and this becomes exactly our condition (C.3).

When $L_{NT}(\beta)$ is approximated up to order $G \in \mathbb{N}$, Kato's equation (3.6) at p.89 gives the following bound on the remainder

$$\left| L_{NT}(\beta) - \frac{1}{NT}\sum_{g=2}^{G}\sum_{\kappa_1=0}^{K}\ldots\sum_{\kappa_g=0}^{K}\epsilon_{\kappa_1}\ldots\epsilon_{\kappa_g}L^{(g)}\left(\lambda^0, f^0, X_{\kappa_1}, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)\right| \leq \frac{(N-R)\gamma^{G+1}d^2_{\min}(\lambda^0, f^0)}{4(1-\gamma)}, \tag{C.18}$$

where

$$\gamma = \frac{\sum_{k=1}^{K}\left|\beta^0_k - \beta_k\right|\frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}}{r_0(\lambda^0, f^0)} < 1. \tag{C.19}$$

This bound again shows convergence of the series expansion, since $\gamma^{G+1} \to 0$ as $G \to \infty$. Unfortunately, for our purposes this is not a good bound since it still involves the factor $N - R$ (in Kato this factor is hidden since his $\hat{\lambda}(\kappa)$ is the average of the eigenvalues, not the sum), but as we show below this can be avoided.

(iii,iv) We have $\|S^{(k)}\| = \left(NTd^2_{\min}(\lambda^0, f^0)\right)^{-k}$, $\|\mathcal{T}^{(1)}_\kappa\| \leq 2\sqrt{NT}d_{\max}(\lambda^0, f^0)\|X_\kappa\|$, and $\|\mathcal{T}^{(2)}_{\kappa_1\kappa_2}\| \leq \|X_{\kappa_1}\|\|X_{\kappa_2}\|$. Therefore

$$\left\|S^{(l_1)}\mathcal{T}^{(\nu_1)}_{\kappa_1\ldots}S^{(l_2)}\ldots S^{(l_p)}\mathcal{T}^{(\nu_p)}_{\ldots\kappa_g}S^{(l_{p+1})}\right\| \leq \left(NTd^2_{\min}(\lambda^0, f^0)\right)^{-\sum l_j}\left(2\sqrt{NT}d_{\max}(\lambda^0, f^0)\right)^{2p-\sum\nu_j}\|X_{\kappa_1}\|\|X_{\kappa_2}\|\ldots\|X_{\kappa_g}\|. \tag{C.20}$$

We have

$$\sum_{\substack{\nu_1+\ldots+\nu_p=g\\ 2\geq\nu_j\geq 1}} 1 \leq 2^p,$$

$$\sum_{\substack{l_1+\ldots+l_{p+1}=p-1\\ l_j\geq 0}} 1 \leq \sum_{\substack{l_1+\ldots+l_{p+1}=p\\ l_j\geq 0}} 1 = \frac{(2p)!}{(p!)^2} \leq 4^p. \tag{C.21}$$

Using this we find[32]

$$\left\|M^{(g)}\left(\lambda^0, f^0, X_{\kappa_1}, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)\right\| \leq \left(2\sqrt{NT}d_{\max}(\lambda^0, f^0)\right)^{-g}\|X_{\kappa_1}\|\|X_{\kappa_2}\|\ldots\|X_{\kappa_g}\|\sum_{p=\lceil g/2\rceil}^{g}\left(\frac{32\,d^2_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)}\right)^p \leq \frac{g}{2}\left(\frac{16\,d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)}\right)^g\frac{\|X_{\kappa_1}\|}{\sqrt{NT}}\frac{\|X_{\kappa_2}\|}{\sqrt{NT}}\ldots\frac{\|X_{\kappa_g}\|}{\sqrt{NT}}. \tag{C.22}$$

[32]The sum over p only starts from $\lceil g/2\rceil$, the smallest integer larger or equal $g/2$, because $\nu_1 + \ldots + \nu_p = g$ can not be satisfied for smaller p, since $\nu_j \leq 2$.

For $g \geq 3$ there always appears at least one factor $S^{(l)}$, $l \geq 1$, inside the trace of the terms that contribute to $L^{(g)}$, and we have $\mathrm{rank}(S^{(l)}) = R$ for $l \geq 1$. Using $\mathrm{Tr}(A) \leq \mathrm{rank}(A)\|A\|$, and the equations (C.20) and (C.21), we therefore find[33] for $g \geq 3$

$$\begin{aligned}
&\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)\right| \\
&\quad \leq R\, d_{\min}^2(\lambda^0, f^0)\, \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0)\right)^{-g} \|X_{\kappa_1}\| \|X_{\kappa_2}\| \ldots \|X_{\kappa_g}\| \sum_{p=\lceil g/2 \rceil}^{g} \left(\frac{32\, d_{\max}^2(\lambda^0, f^0)}{d_{\min}^2(\lambda^0, f^0)}\right)^p \\
&\quad \leq \frac{R\, g\, d_{\min}^2(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d_{\min}^2(\lambda^0, f^0)}\right)^g \frac{\|X_{\kappa_1}\|}{\sqrt{NT}} \frac{\|X_{\kappa_2}\|}{\sqrt{NT}} \cdots \frac{\|X_{\kappa_g}\|}{\sqrt{NT}} \,. \qquad \text{(C.23)}
\end{aligned}$$

This implies for $g \geq 3$

$$\begin{aligned}
&\frac{1}{NT}\left|\sum_{\kappa_1=0}^{K}\sum_{\kappa_2=0}^{K}\cdots\sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1}\, \epsilon_{\kappa_2}\, \ldots\, \epsilon_{\kappa_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)\right| \\
&\qquad\qquad \leq \frac{R\, g\, d_{\min}^2(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d_{\min}^2(\lambda^0, f^0)}\right)^g \left(\sum_{\kappa=0}^{K} \frac{\|\epsilon_\kappa\, X_\kappa\|}{\sqrt{NT}}\right)^g \,. \qquad \text{(C.24)}
\end{aligned}$$

Therefore for $G \geq 2$ we have

$$\begin{aligned}
&\left|L_{NT}\left(\beta\right) - \frac{1}{NT}\sum_{g=2}^{G}\sum_{\kappa_1=0}^{K}\cdots\sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1}\, \ldots\, \epsilon_{\kappa_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)\right| \\
&\qquad = \frac{1}{NT}\sum_{g=G+1}^{\infty}\sum_{\kappa_1=0}^{K}\sum_{\kappa_2=0}^{K}\cdots\sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1}\, \epsilon_{\kappa_2}\, \ldots\, \epsilon_{\kappa_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \\
&\qquad \leq \sum_{g=G+1}^{\infty} \frac{R\, g\, \alpha^g\, d_{\min}^2(\lambda^0, f^0)}{2} \\
&\qquad \leq \frac{R\,(G+1)\,\alpha^{G+1}\, d_{\min}^2(\lambda^0, f^0)}{2(1-\alpha)^2} \,, \qquad \text{(C.25)}
\end{aligned}$$

where

$$\begin{aligned}
\alpha \;&= \frac{16\, d_{\max}(\lambda^0, f^0)}{d_{\min}^2(\lambda^0, f^0)} \sum_{\kappa=0}^{K} \frac{\|\epsilon_\kappa\, X_\kappa\|}{\sqrt{NT}} \\
&= \frac{16\, d_{\max}(\lambda^0, f^0)}{d_{\min}^2(\lambda^0, f^0)} \left(\sum_{k=1}^{K} \left|\beta_k^0 - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right) \;< 1 \,. \qquad \text{(C.26)}
\end{aligned}$$

Using the same argument we can start from equation (C.22) to obtain the bound (C.16) for the remainder of the series expansion for $M_{\hat{\lambda}}\left(\beta\right)$.

Note that compared to the bound (C.18) on the remainder, the new bound (C.25) only shows convergence of the power series within the the smaller convergence radius $\frac{d_{\min}^2(\lambda^0, f^0)}{16\, d_{\max}(\lambda^0, f^0)} < r_0(\lambda^0, f^0)$. However, the factor $N - R$ does not appear in this new bound, which is crucial for our approximations.

∎

We can now proof the quadratic expansion of $L_{NT}(\beta)$ given in the main text.

[33] The calculation for the bound of $L^{(g)}$ is almost identical to the one for $M^{(g)}$. But now there appears an additional factor R from the rank, and since $\sum l_j = p - 1$ (not p as before), there is also an additional factor $NT d_{\min}^2(\lambda^0, f^0)$.

Proof of theorem 3.1. Assumption 1 implies that

$$d_{\max}(\lambda^0, f^0) \underset{p}{\longrightarrow} d^{\infty}_{\max} > 0 \,, \qquad\qquad d_{\min}(\lambda^0, f^0) \underset{p}{\longrightarrow} d^{\infty}_{\min} > 0 \,. \tag{C.27}$$

Therefore also $r_0(\lambda^0, f^0) \to_p r_0^{\infty} > 0$. Assumptions 1, 3, and 4 furthermore imply that

$$\begin{aligned} \frac{\|\lambda^0\|}{\sqrt{N}} &= \mathcal{O}_p(1) \,, & \frac{\|f^0\|}{\sqrt{T}} &= \mathcal{O}_p(1) \,, \\ \left\| \left(\frac{\lambda^{0\prime}\lambda^0}{N} \right)^{-1} \right\| &= \mathcal{O}_p(1) \,, & \left\| \left(\frac{f^{0\prime}f^0}{T} \right)^{-1} \right\| &= \mathcal{O}_p(1) \,, \\ \frac{\|X_k\|}{\sqrt{NT}} &= \mathcal{O}_p(1) \,, & \frac{\|e\|}{\sqrt{NT}} &= o_p(1) \,. \end{aligned} \tag{C.28}$$

Since $\left\|\beta - \beta^0\right\| \leq \eta_{NT}$ and $\eta_{NT} = o(1)$ we find $\alpha \to 0$ as $N, T \to \infty$, *i.e.* the condition to apply theorem C.2 part (iii) is asymptotically satisfied. Using the inequality (C.11), the linearity of $L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right)$ in the arguments X_κ, and the fact that $\epsilon_0 X_0 = e$ we find

$$\frac{1}{NT}(\epsilon_0)^{g-r}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1}, \ldots,\, X_{k_r},\, X_0, \ldots,\, X_0\right) = \mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}} \right)^{g-r} \right) \,. \tag{C.29}$$

Applying the inequality (C.13) for $G = G_e$ then gives

$$\begin{aligned} & L_{NT}\left(\beta\right) \\ =& \frac{1}{NT} \sum_{g=2}^{G_e} \sum_{\kappa_1=0}^{K} \cdots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \ldots \epsilon_{\kappa_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) + \mathcal{O}_p\left(\alpha^{G+1}\right) \\ =& \frac{1}{NT} \sum_{g=2}^{G_e} \epsilon_0^g\, L^{(g)}\left(\lambda^0,\, f^0,\, X_0,\, X_0, \ldots, X_0\right) \\ & + \frac{1}{NT} \sum_{g=2}^{G_e} g \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k\right) \epsilon_0^{g-1}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, X_0, \ldots, X_0\right) \\ & + \frac{1}{NT} \sum_{g=2}^{G_e} g\,(g-1) \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \left(\beta_{k_1}^0 - \beta_{k_1}\right) \left(\beta_{k_2}^0 - \beta_{k_2}\right) \epsilon_0^{g-2} L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2},\, X_0, \ldots, X_0\right) \\ & + \frac{1}{NT} \sum_{g=3}^{G_e} \sum_{r=3}^{g} \mathcal{O}_p\left[\left\|\beta^0 - \beta\right\|^r \epsilon_0^{g-r} L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1}, \ldots,\, X_{k_r},\, X_0, \ldots, X_0\right) \right] \\ & + \mathcal{O}_p\left[\left(\sum_{k=1}^{K} \left|\beta_k^0 - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} \right)^{G_e+1} \right] \\ =& \frac{1}{NT} \sum_{g=2}^{G_e} g \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k\right) L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e, \ldots, e\right) \\ & + \frac{2}{NT} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \left(\beta_{k_1}^0 - \beta_{k_1}\right) \left(\beta_{k_2}^0 - \beta_{k_2}\right) L^{(2)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}\right) \\ & + \frac{1}{NT}\, I_{NT} + \frac{1}{NT}\, R_{NT}(\beta) \,, \end{aligned} \tag{C.30}$$

where

$$
\begin{aligned}
I_{NT} =& \sum_{g=2}^{G_e} L^{(g)}\left(\lambda^0,\, f^0,\, e,\, e, \ldots, e\right) + NT\, \mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e+1}\right) \,, \\
R_{NT}(\beta) =& R_{1,NT}(\beta) + R_{2,NT}(\beta) \,, \\
R_{1,NT}(\beta) =& NT \sum_{g=3}^{G_e+1} \sum_{r=2}^{g} \mathcal{O}_p\left(\|\beta^0 - \beta\|^r \left(\frac{\|e\|}{\sqrt{NT}}\right)^{g-r}\right) \,, \\
R_{2,NT}(\beta) =& NT\, \mathcal{O}_p\left(\|\beta^0 - \beta\| \left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e}\right) \,. \qquad \text{(C.31)}
\end{aligned}
$$

We find that I_{NT} is independent of β, while $R_{1,NT}(\beta)$ and $R_{2,NT}(\beta)$ satisfy

$$
\begin{aligned}
\sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{1,NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} \leq & \sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{1,NT}(\beta)|}{NT\,\left\|\beta-\beta^0\right\|^2} \\
&= \mathcal{O}_p\left(\frac{\|e\|}{\sqrt{NT}}\right) + \mathcal{O}_p\left(\eta_{NT}\right) = o_p(1) \,, \\
\sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{2,NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} \leq & \sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{2,NT}(\beta)|}{2\sqrt{NT}\,\left\|\beta-\beta^0\right\|} \\
&= \sqrt{NT}\,\mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e}\right) = o_p(1) \,, \qquad \text{(C.32)}
\end{aligned}
$$

where we used $\eta_{NT} \to 0$ and assumption 3^* to show that the terms are $o_p(1)$. Since the condition (3.4) is satisfied for $R_{1,NT}(\beta)$ and $R_{2,NT}(\beta)$ separately, it is also satisfied for the total remainder $R_{NT}(\beta)$. ∎

Proof of Theorem 3.2. Using the expansion of $L_{NT}(\beta)$ in theorem C.2 we find for the derivative (the sign convention $\epsilon_k = \beta_k^0 - \beta_k$ results in the minus sign below)

$$
\begin{aligned}
\frac{\partial L_{NT}}{\partial \beta_k} =& -\frac{1}{NT} \sum_{g=2}^{\infty} g \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \cdots \sum_{\kappa_{g-1}=0}^{K} \epsilon_{\kappa_1}\, \epsilon_{\kappa_2}\, \ldots\, \epsilon_{\kappa_{g-1}}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, X_{\kappa_1}, \ldots, X_{\kappa_{g-1}}\right) \\
=& \left[2W_{NT}(\beta-\beta^0)\right]_k - \frac{2}{\sqrt{NT}} C_{NT,k} + \frac{1}{NT}\nabla R_{1,NT,k} + \frac{1}{NT}\nabla R_{2,NT,k} \,, \qquad \text{(C.33)}
\end{aligned}
$$

where

$$
\begin{aligned}
W_{NT,k_1k_2} =& \frac{1}{NT}\, L^{(2)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}\right) \,, \\
C_{NT,k} =& \frac{1}{2\sqrt{NT}} \sum_{g=2}^{G_e} g\,(\epsilon_0)^{g-1}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, X_0, \ldots, X_0\right) \\
=& \sum_{g=2}^{G_e} \frac{g}{2\sqrt{NT}}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e, \ldots, e\right) \,, \qquad \text{(C.34)}
\end{aligned}
$$

and

$$\begin{aligned}
\nabla R_{1,NT,k} &= -\sum_{g=G_e+1}^{\infty} g\,(\epsilon_0)^{g-1}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_0,\ldots,X_0\right)\,,\\
&= -\sum_{g=G_e+1}^{\infty} g\;L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,e,\ldots,e\right)\,,\\
\nabla R_{2,NT,k} &= -\sum_{g=3}^{\infty} g \sum_{r=1}^{g-1}\binom{g-1}{r}\sum_{k_1=1}^{K}\cdots\sum_{k_r=1}^{K}\epsilon_{k_1}\,\ldots\,\epsilon_{k_r}\,(\epsilon_0)^{g-r-1}\\
&\qquad\qquad L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_{k_1},\ldots,X_{k_r},X_0,\ldots,X_0\right)\,.\\
&= -\sum_{g=3}^{\infty} g \sum_{r=1}^{g-1}\binom{g-1}{r}\sum_{k_1=1}^{K}\cdots\sum_{k_r=1}^{K}(\beta^0_{k_1}-\beta_{k_1})\,\ldots\,(\beta^0_{k_r}-\beta_{k_r})\\
&\qquad\qquad L^{(g)}\left(\lambda^0,\,f^0,\,X_k,\,X_{k_1},\ldots,X_{k_r},e,\ldots,e\right)\,. \qquad \text{(C.35)}
\end{aligned}$$

The above expressions for W_{NT} and C_{NT} are equivalent to their definitions given in theorem 3.1. Using the bound on $L^{(g)}$ in formula (C.11) we find[34]

$$\begin{aligned}
|\nabla R_{1,NT,k}| &\leq c_0\,NT\,\frac{\|X_k\|}{\sqrt{NT}}\sum_{g=G_e+1}^{\infty} g^2\left(\frac{c_1\|e\|}{\sqrt{NT}}\right)^{g-1}\\
&\leq 2\,c_0\,(1+G_e)^2\,NT\,\frac{\|X_k\|}{\sqrt{NT}}\left(\frac{c_1\|e\|}{\sqrt{NT}}\right)^{G_e}\left[1-\left(\frac{c_1\|e\|}{\sqrt{NT}}\right)\right]^{-3}=o_p(\sqrt{NT})\,,\\
|\nabla R_{2,NT,k}| &\leq c_0\,NT\,\frac{\|X_k\|}{\sqrt{NT}}\sum_{g=3}^{\infty} g^2\sum_{r=1}^{g-1}\binom{g-1}{r}c_1^{g-1}\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_k|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right)\\
&\qquad\qquad\times\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_k|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)^{g-2}\\
&\leq c_0\,NT\,\frac{\|X_k\|}{\sqrt{NT}}\sum_{g=3}^{\infty} g^3\,(4c_1)^{g-1}\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_k|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right)\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_{\tilde{k}}|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)^{g-2}\\
&\leq c_2\,NT\,\frac{\|X_k\|}{\sqrt{NT}}\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_k|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}\right)\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_{\tilde{k}}|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)\,, \qquad \text{(C.36)}
\end{aligned}$$

where $c_0 = 8Rd_{\max}(\lambda^0,f^0)/2$ and $c_1 = 16d_{\max}(\lambda^0,f^0)/d^2_{\min}(\lambda^0,f^0)$ both converge to a constants as $N,T\to\infty$, and the very last inequality is only true if $4c_1\left(\sum_{\tilde{k}=1}^{K}|\beta_{\tilde{k}}-\beta^0_{\tilde{k}}|\frac{\|X_{\tilde{k}}\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)<1$, 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 $\eta_{NT}\to\infty$

$$\begin{aligned}
\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}}\frac{\|\nabla R_{1,NT}(\beta)\|}{\sqrt{NT}} &= o_p\,(1)\,,\\
\sup_{\{\beta:\|\beta-\beta^0\|\leq\eta_{NT}\}}\frac{\|\nabla R_{2,NT}(\beta)\|}{NT\,\left\|\beta-\beta^0\right\|} &= o_p\,(1)\,. \qquad \text{(C.37)}
\end{aligned}$$

Thus $R_{NT}(\beta)=R_{1,NT}(\beta)+R_{2,NT}(\beta)$ satisfies the bound in equation (3.7). ∎

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

Proof of Theorem 3.3. The general expansion of $M_{\hat{\lambda}}$ is given in theorem C.2, and in the theorem we just make this expansion explicit up to a particular order. To obtain the bound on the remainder we make us of equation (C.22) in the proof of theorem C.2. The result for $M_{\hat{f}}$ 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

$$\begin{aligned}\hat{e} &= M_{\hat{\lambda}} \left(Y - \sum_{k=1} \hat{\beta}_k \, X_k \right) \\ &= M_{\hat{\lambda}} \left[e - \sum_{k=1}^{K} \left(\hat{\beta}_k - \beta_k^0 \right) X_k + \lambda^0 f^{0\prime} \right] , \end{aligned} \tag{C.38}$$

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 $A_3 = -\hat{e}_e^{(1)}$. We find $\mathrm{rank}(A_1) \leq 2R$, $\mathrm{rank}(A_2) \leq 2R$, $\mathrm{rank}(A_3) \leq 3R$, and thus $\mathrm{rank}(\hat{e}^{(\mathrm{rem})}(\beta)) \leq 7R$, as stated in the theorem. ∎

D Proof of Theorem 4.1

Lemma D.1. *Under assumption 5 we have*

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

Lemma D.2. *Under assumption 5 we have*

$$
\begin{aligned}
&(a) && \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X_{k_1}^{\mathrm{weak}\prime}\, P_{\lambda^0}\, X_{k_2}^{\mathrm{weak}}) = o_p(1)\,,\\
&(b) && \frac{1}{NT}\mathrm{Tr}(P_{f^0}\, X_{k_1}^{\mathrm{weak}\prime}\, X_{k_2}^{\mathrm{weak}}) = o_p(1)\,,\\
&(c) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e\, M_{f^0}\, e'\, M_{\lambda^0}\, X_k^{\mathrm{weak}}\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) = o_p(1)\,,\\
&(d) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, M_{\lambda^0}\, e\, M_{f^0}\, X_k^{\mathrm{weak}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) = o_p(1)\,,\\
&(e) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e' M_{\lambda^0}\, X_k^{\mathrm{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) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(P_{f^0}\, e'\, P_{\lambda^0}\, X_k^{\mathrm{weak}}\right) = o_p(1)\,,\\
&(g) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(eP_{f^0}\, e'\, 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) = o_p(1)\,,\\
&(h) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, P_{\lambda^0}\, e\, M_{f^0}\, X_k^{\mathrm{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) = o_p(1)\,,\\
&(i) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e' M_{\lambda^0}\, X_k^{\mathrm{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) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, P_{\lambda^0}\, X_k^{\mathrm{weak}}\right) = o_p(1)\,,\\
&(k) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{\left[ee' - \mathbb{E}\,(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\} = o_p(1)\,,\\
&(l) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{\left[e'e - \mathbb{E}\,(e'e)\right]\, M_{f^0}\, X_k^{\mathrm{str}\,\prime}\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right\} = o_p(1)\,,\\
&(m) && \frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{P_{f^0}\, \left[e'\, X_k^{\mathrm{weak}} - \mathbb{E}\left(e'\, X_k^{\mathrm{weak}}\right)\right]\right\} = o_p(1)\,. \qquad \text{(D.2)}
\end{aligned}
$$

Lemma D.3. *Under assumptions 5 and 6 we have*

$$
\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\mathfrak{X}_{it} \underset{d}{\to} \mathcal{N}\left(0, \Omega\right)\,. \qquad \text{(D.3)}
$$

The proofs of the three preceding lemmas are given in the supplementary material. Lemma D.3 is a straightforward application of the Lindberg Feller CLT, using our assumption of cross-sectional independence (for e_{it} and X_{it}^{weak}). With these lemmas we can now proof the theorem on the limiting distribution of $\hat{\beta}$ in the main text.

Proof of Theorem 4.1. We have $\|e\| = \mathcal{O}_p(N^{1/2})$, i.e. assumption 3* is satisfied with $G_e = 3$. When applying corollary 3.4 to calculate the limiting distribution of $\hat{\beta}$ we therefore only have to calculate the limit of the denominator terms $\frac{1}{\sqrt{NT}}C^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e\right)$ for $g = 2$ and $g = 3$. Using Lemma D.2 and assumption 6 we find for the matrix in the numerator

$$
\begin{aligned}
W_{NT,k_1k_2} &= \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X_{k_1}'\, M_{\lambda^0}\, X_{k_2})\\
&= \frac{1}{NT}\mathrm{Tr}(\mathfrak{X}_{k_1}'\mathfrak{X}_{k_2}) - \frac{1}{NT}\mathrm{Tr}(P_{f^0}\, X_{k_1}^{\mathrm{weak}\prime}\, X_{k_2}^{\mathrm{weak}}) - \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X_{k_1}^{\mathrm{weak}\prime}\, P_{\lambda^0}\, X_{k_2}^{\mathrm{weak}})\\
&= \frac{1}{NT}\mathrm{Tr}(\mathfrak{X}_{k_1}'\mathfrak{X}_{k_2}) + o_p(1)\,.\\
&= W + o_p(1)\,. \qquad \text{(D.4)}
\end{aligned}
$$

Using Lemmas D.2 and D.3 and assumption 6 we find for the denominator terms

$$\begin{aligned}
\frac{1}{\sqrt{NT}}C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) =& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(M_{f^0}\, e'\, M_{\lambda^0}\, X_k\right) \\
=& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\, X_k^{\mathrm{weak}}\right)\right] \\
& - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, P_{\lambda^0}\, X_k^{\mathrm{weak}}\right) + \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(P_{f^0}\, e'\, P_{\lambda^0}\, X_k^{\mathrm{weak}}\right) \\
& - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left\{P_{f^0}\,\left[e'\, X_k^{\mathrm{weak}} - \mathbb{E}\left(e'\, X_k^{\mathrm{weak}}\right)\right]\right\} \\
=& \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\mathfrak{X}_k\right) - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\, X_k^{\mathrm{weak}}\right)\right] + o_p(1)\,. \\
\underset{d}{\to}& \mathcal{N}\left(-\kappa B_1,\, \Omega\right)\,, \qquad \text{(D.5)}
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{\sqrt{NT}}C^{(3)}\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) \\
& + \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) \\
& + \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^{\mathrm{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(eP_{f^0}\, e'\, 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}{\sqrt{NT}}\mathrm{Tr}\left\{\left[ee' - \mathbb{E}\left(ee'\right)\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}{\sqrt{NT}}\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] \\
& - \frac{1}{\sqrt{NT}}\mathrm{Tr}\left(e'\, M_{\lambda^0}\, e\, M_{f^0}\, X_k^{\mathrm{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^{\mathrm{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^{\mathrm{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^{\mathrm{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^{\mathrm{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^{\mathrm{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^{\mathrm{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^{\mathrm{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{(D.6)}
\end{aligned}$$

Combining these results we obtain

$$\sqrt{NT}\left(\hat{\beta}-\beta^0\right) = W_{NT}^{-1}\left(\frac{1}{\sqrt{NT}}C^{(2)}+\frac{1}{\sqrt{NT}}C^{(2)}\right) , \\ \underset{d}{\to} \mathcal{N}\left(-W^{-1}\left(\kappa B_1+\kappa^{-1}B_2+\kappa B_3\right), W^{-1}\,\Omega\,W^{-1}\right) , \tag{D.7}$$

which is what we wanted to show. ∎

E Proof of Theorem 4.4

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 4.1.

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

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

Proof. Using $\|e\|=\mathcal{O}_p(N^{1/2})$ and $\|X_k\|=\mathcal{O}_p(N)$ we find that the expansion terms in theorem 3.3 satisfy

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

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*

$$A_1 \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2\left(\mathfrak{x}_{it}\mathfrak{x}_{it}'-\hat{\mathfrak{x}}_{it}\hat{\mathfrak{x}}_{it}'\right)=o_p(1) , \quad A_2 \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(e_{it}^2-\hat{e}_{it}^2\right)\hat{\mathfrak{x}}_{it}\hat{\mathfrak{x}}_{it}'=o_p(1) . \tag{E.3}$$

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 4.4, there exists an* $R\times R$ *matrices* $H=H_{N,T}$ *such that*[35]

$$\left\|\hat{f}-f^0\,H\right\| = O_p\left(1\right) , \qquad \left\|\hat{\lambda}-\lambda^0\,\left(H'\right)^{-1}\right\| = O_p\left(1\right) . \tag{E.4}$$

Furthermore

$$\left\|\hat{\lambda}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\left(\hat{f}'\hat{f}\right)^{-1}\hat{f}'-\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}f^{0\prime}\right\| = O_p\left(N^{-3/2}\right) . \tag{E.5}$$

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

$$\text{(i)} \qquad N^{-1}\left\|\mathbb{E}(e'M_{\lambda^0}X_k)-\left(\hat{e}'\,X_k\right)^{\text{truncR}}\right\| = o_p(1) , \\ \text{(ii)} \qquad N^{-1}\left\|\mathbb{E}(e'M_{\lambda^0}e)-\left(\hat{e}'\,\hat{e}\right)^{\text{truncD}}\right\| = o_p(1) , \\ \text{(iii)} \qquad T^{-1}\left\|\mathbb{E}(eM_{f^0}e')-\left(\hat{e}\,\hat{e}'\right)^{\text{truncD}}\right\| = o_p(1) . \tag{E.6}$$

[35] 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 $H_{N,T}$ to keep notation simple.

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

$$\begin{aligned} &\text{(i)} && N^{-1}\left\|(\hat{e}'\,X_k)^{\text{truncR}}\right\| = \mathcal{O}_p(MT^{1/8})\,, \\ &\text{(ii)} && N^{-1}\left\|(\hat{e}'\,\hat{e})^{\text{truncD}}\right\| = \mathcal{O}_p(1)\,, \\ &\text{(iii)} && T^{-1}\left\|(\hat{e}\,\hat{e}')^{\text{truncD}}\right\| = \mathcal{O}_p(1)\,. \end{aligned} \tag{E.7}$$

The proof of the above lemmas is given in the supplementary material. Using these lemmas we can now prove theorem 4.4.

Proof of Theorem 4.4, 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

$$\begin{aligned} W_{NT,k_1k_2} &= (NT)^{-1}\text{Tr}(\mathfrak{X}_{k_1}\mathfrak{X}'_{k_2}) \\ &= (NT)^{-1}\text{Tr}(M_{\lambda^0}\,X_{k_1}^{\text{str}}\,M_{f^0}\,X_{k_2}^{\text{str}\,\prime}) + (NT)^{-1}\text{Tr}(M_{\lambda^0}\,X_{k_1}^{\text{str}}\,M_{f^0}\,X_{k_2}^{\text{weak}\,\prime}) \\ &\quad + (NT)^{-1}\text{Tr}(M_{\lambda^0}\,X_{k_1}^{\text{weak}}\,M_{f^0}\,X_{k_2}^{\text{str}\,\prime}) + (NT)^{-1}\text{Tr}(X_{k_1}^{\text{weak}}X_{k_2}^{\text{weak}\,\prime})\,. \end{aligned} \tag{E.8}$$

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 $\hat{W}_{k_1k_2} = (NT)^{-1}\text{Tr}(\hat{\mathfrak{X}}_{k_1}\hat{\mathfrak{X}}'_{k_2}) = (NT)^{-1}\text{Tr}(M_{\hat{\lambda}}\,X_{k_1}\,M_{\hat{f}}\,X'_{k_2})$. Using $|\text{Tr}\,(C)| \leq \|C\|\,\text{rank}\,(C)$ (which is true for every square matrix C, see the supplementary material), corollary E.2, and the result $\|X_{k_1}^{\text{weak}}\| = \mathcal{O}_p(N^{-1})$ from Lemma D.1, we find

$$\begin{aligned} &\left|\hat{W}_{k_1k_2} - W_{NT,k_1k_2}\right| \\ &\quad = \Big|(NT)^{-1}\text{Tr}\left[\left(M_{\hat{\lambda}} - M_{\lambda^0}\right)\,X_{k_1}\,M_{\hat{f}}\,X'_{k_2}\right] + (NT)^{-1}\text{Tr}\left[M_{\lambda^0}\,X_{k_1}\,\left(M_{\hat{f}} - M_{f^0}\right)\,X'_{k_2}\right] \\ &\qquad - (NT)^{-1}\text{Tr}(M_{\lambda^0}X_{k_1}^{\text{weak}}P_{f^0}X_{k_2}^{\text{weak}\,\prime}) - (NT)^{-1}\text{Tr}(P_{\lambda^0}X_{k_1}^{\text{weak}}X_{k_2}^{\text{weak}\,\prime})\Big| \\ &\quad \leq \frac{2R}{NT}\left\|M_{\hat{\lambda}} - M_{\lambda^0}\right\|\|X_{k_1}\|\|X_{k_2}\|\frac{2R}{NT}\left\|M_{\hat{f}} - M_{f^0}\right\|\|X_{k_1}\|\|X_{k_2}\| \\ &\qquad + \frac{R}{NT}\|X_{k_1}^{\text{weak}}\|\|X_{k_2}^{\text{weak}}\| + \frac{R}{NT}\|X_{k_1}^{\text{weak}}\|\|X_{k_2}^{\text{weak}}\| \\ &\quad = \frac{2R}{NT}\mathcal{O}_p(N^{-1})\mathcal{O}_p(NT) + \frac{2R}{NT}\mathcal{O}_p(T^{-1})\mathcal{O}_p(NT) + \frac{2R}{NT}\mathcal{O}_p(N) \\ &\quad = o_p(1)\,. \end{aligned} \tag{E.9}$$

This is what we wanted to show. ∎

Proof of Theorem 4.4, 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 $\text{Var}\,(\Omega_{NT,k_1k_2}) = o(1)$ we are done.

Using cross-sectional independence of e_{it} and X_{it}^{weak}, we find that conditional on X_k^{str} (or alterna-

tively, treating X_k^{str} as non-stochastic) we have

$$\begin{aligned}
\text{Var}\left(\Omega_{NT,k_1k_2}\right) &= \frac{1}{(NT)^2}\sum_{i,j=1}^{N}\sum_{t,\tau=1}^{T}\Big[\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\,e_{j\tau}^2\mathfrak{X}_{k_1,j\tau}\mathfrak{X}_{k_2,j\tau}\right) \\
&\qquad\qquad - \mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\mathbb{E}\left(e_{j\tau}^2\mathfrak{X}_{k_1,j\tau}\mathfrak{X}_{k_2,j\tau}\right)\Big] \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\Big[\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\,e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) \\
&\qquad\qquad - \mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\mathbb{E}\left(e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right)\Big] \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\left\{\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\,e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) - \left[\sum_{t=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\right)\right]^2\right\} \\
&\leq \frac{1}{(NT)^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^2\mathfrak{X}_{k_1,it}\mathfrak{X}_{k_2,it}\,e_{i\tau}^2\mathfrak{X}_{k_1,i\tau}\mathfrak{X}_{k_2,i\tau}\right) \\
&\leq \frac{1}{N}\sqrt{\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(e_{it}^4e_{i\tau}^4\right)\frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t,\tau=1}^{T}\mathbb{E}\left(\mathfrak{X}_{k_1,it}^2\mathfrak{X}_{k_2,it}^2\,\mathfrak{X}_{k_1,i\tau}^2\mathfrak{X}_{k_2,i\tau}^2\right)} \\
&= \frac{1}{N}\,\mathcal{O}(1) = o(1)\,, \qquad \text{(E.10)}
\end{aligned}$$

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 4.4, Part III: show $\hat{B}_1 = B_1 + o_p(1)$.
Let $B_{1,k,NT} = N^{-1}\,\text{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,X_k^{\text{weak}}\right)\right]$, and $\tilde{B}_{1,k,NT} = N^{-1}\,\text{Tr}\left[P_{f^0}\,\mathbb{E}\left(e'\,M_{\lambda^0}\,X_k^{\text{weak}}\right)\right]$. According to assumption 6 we have $B_{1,k} = B_{1,k,NT} + o_p(1)$. Applying part (f) of Lemma D.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 $|\text{Tr}\,(C)| \leq \|C\|\,\text{rank}\,(C)$ (which is true for every square matrix C, see the supplementary material) we find

$$\begin{aligned}
\left|\tilde{B}_{1,k,NT} - \hat{B}_1\right| &= \left|\mathbb{E}\left[\frac{1}{N}\text{Tr}(P_{f^0}\,e'\,M_{\lambda^0}\,X_k)\right] - \frac{1}{N}\text{Tr}\left[P_{\hat{f}}\,(\hat{e}'X_k)^{\text{truncR}}\right]\right| \\
&\leq \left|\frac{1}{N}\text{Tr}\left[\left(P_{f^0} - P_{\hat{f}}\right)(\hat{e}'X_k)^{\text{truncR}}\right]\right| \\
&\quad + \left|\frac{1}{N}\text{Tr}\left\{P_{f^0}\left[\mathbb{E}\,(e'\,M_{\lambda^0}\,X_k) - (\hat{e}'X_k)^{\text{truncR}}\right]\right\}\right| \\
&\leq \frac{2R}{N}\left\|P_{f^0} - P_{\hat{f}}\right\|\left\|(\hat{e}'X_k)^{\text{truncR}}\right\| \\
&\quad + \frac{R}{N}\left\|P_{f^0}\right\|\left\|\mathbb{E}\,(e'\,M_{\lambda^0}\,X_k) - (\hat{e}'X_k)^{\text{truncR}}\right\| \qquad \text{(E.11)}
\end{aligned}$$

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

$$\left|\tilde{B}_{1,k,NT} - \hat{B}_1\right| = N^{-1}\left(\mathcal{O}_p(N^{-1/2})\mathcal{O}_p(MNT^{1/8}) + o_p(N)\right) = o_p(1)\,. \qquad \text{(E.12)}$$

This is what we wanted to show. ∎

Proof of Theorem 4.4, final part: show $\hat{B}_i = B_i + o_p(1)$, $i = 2, 3$.
Define

$$B_{2,k,NT} = \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(ee'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right] ,$$
$$\tilde{B}_{2,k,NT} = \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(eM_{f^0}e'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right] . \tag{E.13}$$

According to assumption 6 we have $B_{2,k} = B_{2,k,NT} + o_p(1)$. Applying part (g) of Lemma D.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 weakly and strictly exogenous part. As a consequence of Lemma D.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 weakly exogenous part of the regressors does not contribute to $\hat{B}_2$ asymptotically, namely

$$\begin{aligned}\hat{B}_{2,k}^{\mathrm{weak}} &= \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{weak}}\, \hat{f}\left(\hat{f}'\hat{f}\right)^{-1}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\hat{\lambda}'\right] \\ &\leq \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|X_k^{\mathrm{weak}}\right\|\left\|\hat{f}\left(\hat{f}'\hat{f}\right)^{-1}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\hat{\lambda}'\right\| \\ &= \frac{R}{T}\,\mathcal{O}_p(T)\,\mathcal{O}_p(N^{1/2})\,\mathcal{O}_p((NT)^{-1/2}) = o_p(1) .\end{aligned} \tag{E.14}$$

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

$$\begin{aligned}\tilde{B}_{2,k} - \hat{B}_{2,k}^{\mathrm{str}} =& \frac{1}{T}\mathrm{Tr}\left[\mathbb{E}\left(eM_{f^0}e'\right)\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right] \\ &- \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{str}}\, \hat{f}\left(\hat{f}'\hat{f}\right)^{-1}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\hat{\lambda}'\right] \\ =& \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, M_{\hat{\lambda}}\, X_k^{\mathrm{str}}\, \left(f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime} - \hat{f}\left(\hat{f}'\hat{f}\right)^{-1}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\hat{\lambda}'\right)\right] \\ &+ \frac{1}{T}\mathrm{Tr}\left[\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\, \left(M_{\lambda^0} - M_{\hat{\lambda}}\right)\, X_k^{\mathrm{str}}\, f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right] \\ &+ \frac{1}{T}\mathrm{Tr}\left\{\left[\mathbb{E}\left(eM_{f^0}e'\right) - \left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right]\, M_{\lambda^0}\, X_k^{\mathrm{str}}\, f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right\} .\end{aligned} \tag{E.15}$$

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

$$\begin{aligned}\left|\tilde{B}_{2,k} - \hat{B}_{2,k}^{\mathrm{str}}\right| \leq& \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|X_k^{\mathrm{str}}\right\|\left\|f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime} - \hat{f}\left(\hat{f}'\hat{f}\right)^{-1}\left(\hat{\lambda}'\hat{\lambda}\right)^{-1}\hat{\lambda}'\right\| \\ &+ \frac{R}{T}\left\|\left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\left\|M_{\lambda^0} - M_{\hat{\lambda}}\right\|\left\|X_k^{\mathrm{str}}\right\|\left\|f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right\| \\ &+ \frac{R}{T}\left\|\mathbb{E}\left(eM_{f^0}e'\right) - \left(\hat{e}\,\hat{e}'\right)^{\mathrm{truncD}}\right\|\, \left\|X_k^{\mathrm{str}}\right\|\, \left\|f^0\left(f^{0\prime}f^0\right)^{-1}\left(\lambda^{0\prime}\lambda^0\right)^{-1}\lambda^{0\prime}\right\| .\end{aligned} \tag{E.16}$$

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

$$\begin{aligned}\left|\tilde{B}_{2,k} - \hat{B}_{2,k}^{\mathrm{str}}\right| = T^{-1}\Big[&\mathcal{O}_p(T)\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p(N^{-3/2}) \\ &+ \mathcal{O}_p(T)\,\mathcal{O}_p(N^{-1/2})\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p((NT)^{-1/2}) \\ &+ o_p(T)\,\mathcal{O}_p((NT)^{1/2})\,\mathcal{O}_p((NT)^{-1/2})\Big] = o_p(1) .\end{aligned} \tag{E.17}$$

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

F Proof of Theorem 4.7

Proof of Theorem 4.7.

- We have shown that the assumption of the theorem imply that assumption 7 holds. Showing that WD^*_{NT} has the limiting distribution χ^2_r is therefore straightforward.

- 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

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

  Alternatively, one could use the expansion of $\hat{e}$ in theorem 3.3 to show this. From the above result we find

$$\begin{aligned}\left|\hat{c} - \frac{1}{NT}\mathrm{Tr}(ee')\right| &= \frac{1}{NT}\left|\mathrm{Tr}(P_{\lambda^0} e M_{f^0} e') + \mathrm{Tr}(eP_{f^0} e')\right| + o_p(1)\\ &\leq \frac{2R}{NT}\,\|e\|^2 + o_p(1) = o_p(1)\;.\end{aligned} \tag{F.2}$$

  By the weak law of large numbers we thus have

$$\hat{c} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e^2_{it} + o_p(1) = c + o_p(1)\;, \tag{F.3}$$

  i.e. $\hat{c}$ is indeed consistent for c. Having this and using theorem 4.6 one immediately obtains the result for the limiting distribution of LR^*_{NT}, as already discussed in the main text.

- For the LM test we use equation 4.18 and $\tilde{W} = W + o_p(1)$, $\tilde{\Omega} = \Omega + o_p(1)$, and $\tilde{B} = B + o_p(1)$ to obtain

$$LM^*_{NT} \underset{d}{\longrightarrow} (C-B)'W^{-1}H'(HW^{-1}\Omega W^{-1}H')^{-1}HW^{-1}(C-B)\;. \tag{F.4}$$

  Under H_0 and assumption 7 we thus find $LM^*_{NT} \to_d \chi^2_r$.

- In order to show that LM^*_{NT} has the same limiting distribution we only need to show that $\sqrt{NT}\tilde{\nabla}\mathcal{L}_{NT} = \sqrt{NT}\nabla L_{NT}(\tilde{\beta}) + o_p(1)$. Using the expansion of $\hat{e}$ in theorem 3.3 one obtains

$$\begin{aligned}\sqrt{NT}(\tilde{\nabla}\mathcal{L}_{NT})_k &= -\frac{2}{\sqrt{NT}}\,\mathrm{Tr}\,(X_k'\tilde{e})\\ &= \left[2\sqrt{NT}\,W_{NT}\left(\tilde{\beta}-\beta^0\right) + \frac{2}{NT}C^{(2)}(\lambda^0,f^0,X_k,e) + \frac{2}{NT}C^{(3)}(\lambda^0,f^0,X_k,e)\right]_k\\ &\qquad - \frac{2}{\sqrt{NT}}\,\mathrm{Tr}\left(X_k'\tilde{e}^{(\mathrm{rem})}\right)\\ &= \left[2\sqrt{NT}\,W_{NT}\left(\tilde{\beta}-\beta^0\right) + \frac{2}{NT}C_{NT}\right]_k + o_p(1)\;,\\ &= \sqrt{NT}\nabla L_{NT}(\tilde{\beta})_k + o_p(1)\,,\end{aligned} \tag{F.5}$$

  which is what we wanted to show. Here we used that $|\mathrm{Tr}\left(X_k'\tilde{e}^{(\mathrm{rem})}\right)| \leq 7R\|X_k\|\|\tilde{e}^{(\mathrm{rem})}\| = \mathcal{O}_p(N^{3/2})$. Note that $\|X_k\| = \mathcal{O}_p(N)$, and theorem 3.3 and $\sqrt{NT}$-consistency of $\tilde{\beta}$ imply $\|\tilde{e}^{(\mathrm{rem})}\| = \mathcal{O}_p(\sqrt{N})$. We also used the expression for $\nabla L_{NT}(\tilde{\beta})$ given in theorem 3.2, and the bound on $\nabla R_{NT}(\beta)$ given there.

∎

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. *Forthcoming in Econometrica*.

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.

Geran, 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. *unpublished 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.

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

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

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 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.