Quasi Likelihood Expansion for Panel Regression Models with Factors and its Applications *

Hyungsik Roger Moon
Department of Economics
University of Southern California

Martin Weidner
Department of Economics
University of Southern California

DRAFT, January 2009

Abstract

In this paper we provide a new methodology to analyze the (Gaussian) profile quasi likelihood function for panel regression models with 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 chi-square distribution.

1 Introduction

This paper studies a panel regression model where the individual fixed effects λ_i, also called factor loadings, interact with common time specific effects f_t, also called factors. Interactive models have been widely used in various economic applications. In analyzing the relationship between wages and hours worked, interactive fixed effect models are considered to account for non-stationary in individual effects (*e.g.*, Holtz-Eakin, Newey, and Rosen (1988)). There λ_i can describe the unobserved earnings abilities of individuals, while f_t can correspond to changes in *e.g.* local working conditions, or macroeconomic states of the economy. In some asset pricing theories, the asset returns are described by an interactive factor model (*e.g.*, Ross (1976) and Chamberlain and Rothschild (1983)). In these theories λ_i measures the sensitivity of the asset to the common factors f_t. Also, the interactive models have been proposed for modeling cross sectional dependence (*e.g.*, Phillips and Sul (2003), Bai and Ng (2004), Moon and Perron (2004), and Pesaran (2006)). For example, in international cross country data analysis, the country specific effect λ_i measures how much a particular country is affected by global shocks of f_t. Here the common shocks of f_t cause correlation in the cross country data.

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 integrated out, becomes the

*Previous versions of this paper were circulated under the title "Asymptotic Analysis of the Quasi-MLE of Panel Regression Models with Interactive Fixed Effects".

sum of the $N - R$ smallest eigenvalues of the sample covariance matrix of the panel, where N is the cross-sectional size of the panel, and R is the number of factors (assumed to be known).

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 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), Hahn and Newey (2004), and Hahn and Moon (2006).

Using the likelihood expansion we 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. We show that these tests are still asymptotically equivalent, but have a non-central chi-square 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 chi-square distribution. We also provide a convenient estimator for the score function at the restricted parameter which features in the LM test and which otherwise would need to be calculated numerically – since no explicit formula for the derivative of the likelihood function is known. Using this score estimator we obtain a modified bias corrected LM test statistics that is easy to compute.

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 asymptotic 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. They also investigate various testing problems. 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.

Another closely related to paper is the work of 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, *e.g.* lagged dependent variables, and we show that they cause *additional* bias terms and how to correct for them. Our treatment of "low-rank regressors" is also 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.

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. Afterwards we conclude. Some technical details have been moved to the appendix, and most of the proofs have been transfered to the supplementary material.

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 $\text{Eigval}_n(B)$. For an $m \times n$ matrix A the Frobenius norm is $\|A\|_F = \sqrt{\text{Tr}(AA')}$, and the operator norm is $\|A\| = \max_{0 \neq v \in \mathbb{R}^n} \frac{\|Av\|}{\|v\|}$, or equivalently $\|A\| = \sqrt{\text{Eigval}_1(A'A)}$. Furthermore, we use $P_A = A(A'A)^{-1}A'$ and $M_A = \mathbb{I} - A(A'A)^{-1}A'$, where $(A'A)^{-1}$ 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 $\mathbb{1}(.)$ for the indicator function.

2 Model, QMLE and Consistency

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

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

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.

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. Our goal is to estimate the parameter $\beta^0 = \left(\beta_1^0, ..., \beta_K^0\right)'$ and to find its limiting distribution when both N and T are large. The estimator we consider in this paper is the QMLE that is defined by

$$\hat{\beta} = \underset{\beta \in \mathbb{B}}{\text{argmin}} \; L_{NT}(\beta) \;, \tag{2.3}$$

where $\mathbb{B} \subset \mathbb{R}^K$ is a compact parameter set that contains the true parameter, *i.e.* $\beta^0 \in \mathbb{B}$. If there are multiple global minima in $\mathbb{B}$ we want $\hat{\beta}$ to be one of them. The objective function is given by

$$\begin{aligned} L_{NT}(\beta) &= \inf_{\lambda, f} \frac{1}{NT} \text{Tr} \left[\left(Y - \sum_{k=1}^{K} \beta_k X_k - \lambda f' \right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k - \lambda f' \right) \right] \\ &= \inf_{f} \frac{1}{NT} \text{Tr} \left[\left(Y - \sum_{k=1}^{K} \beta_k X_k \right) M_f \left(Y - \sum_{k=1}^{K} \beta_k X_k \right)' \right] \\ &= \frac{1}{NT} \sum_{t=R+1}^{T} \text{Eigval}_t \left[\left(Y - \sum_{k=1}^{K} \beta_k X_k \right)' \left(Y - \sum_{k=1}^{K} \beta_k X_k \right) \right] \,. \end{aligned} \tag{2.4}$$

Here we give three expressions for $L_{NT}(\beta)$, which are shown to be equivalent in the supplementary material. **!**

The first expression for $L_{NT}(\beta)$ is the sum of the squares of the residuals $\hat{e}_{it} = Y_{it} - \beta' X_{it} - \lambda_i^{0\prime} f_t^0$ minimized over the parameters λ and f. We have $\sum_i \sum_t \hat{e}_{it}^2 = \text{Tr}(\hat{e}'\hat{e})$, and this trace notation will be used extensively throughout the paper. Note that $L_{NT}(\beta)$ is minus the logarithm of the Gaussian profile likelihood function of model (2.2), and in the following we therefore refer to $L_{NT}(\beta)$ as profile quasi likelihood function. Note also that the minimizing value $L_{NT}(\beta)$ is uniquely defined, although the minimizing parameters $\hat{\lambda}$ and $\hat{f}$ are not uniquely determined, since the objective function is invariant under transformations $\lambda \longrightarrow \lambda A$ and $f \longrightarrow f A^{-1}$, where A is a non-singular $R \times R$ matrix.

The second expression for $L_{NT}(\beta)$ is obtained form the first one by integrating out λ, *i.e.* by eliminating it from the objective function by use of its own first order condition. Analogously one can integrate out f to obtain a formulation where only the parameter λ remains. In the appendix we show that the optimal f is obtained by combining the R eigenvectors that correspond to the R largest eigenvalues of the $T \times T$ matrix $\left(Y - \sum_{k=1}^K \beta_k X_k\right)' \left(Y - \sum_{k=1}^K \beta_k X_k\right)$. Using this result one obtains 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. This expression 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.[1]

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.[2]

Assumption 1. *The probability limits of $\lambda^{0\prime}\lambda^0/N$ and $f^{0\prime} f^0/T$ exist and have full rank, i.e.*
(i) $\text{plim}_{N,T\to\infty} \left(\lambda^{0\prime}\lambda^0/N\right) > 0$, *(ii)* $\text{plim}_{N,T\to\infty} \left(f^{0\prime} f^0/T\right) > 0$.

Assumption 2. *(i)* $\text{plim}_{N,T\to\infty} \left[(NT)^{-1}\text{Tr}(X_k\, e')\right] = 0$, *(ii)* $\text{plim}_{N,T\to\infty} \left[(NT)^{-1}\text{Tr}(\lambda^0\, f^{0\prime} e')\right] = 0$.

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

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 $f_{\cdot r}^0$ and factor loading $\lambda_{\cdot r}^0$ grows at a rate of $\sqrt{T}$ and $\sqrt{N}$, respectively. Assumption 2 demands that the regressors are weakly exogenous and that the combination of factors and factor loadings is "weakly exogenous" in the same sense. 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. *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.* $\text{plim}_{N,T\to\infty} \left[(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T X_{it} X_{it}'\right] > 0$. *In addition, 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 $X_{\text{high},\alpha} = \sum_{m=K_1+1}^K \alpha_m X_m$ of the high-rank regressors X_m for

[1] For numerical purposes one should use the last expression in (2.4) 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 $\left(Y - \sum_{k=1}^K \beta_k X_k\right) \left(Y - \sum_{k=1}^K \beta_k X_k\right)'$.

[2] 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.

K_2-vectors[3] *α with $\|\alpha\| = 1$. We assume*

$$\min_{\{\alpha \in \mathbb{R}^{K_2}, \|\alpha\|=1\}} \sum_{i=2R+K_1+1}^{N} \text{Eigval}_i \left(\frac{X_{\text{high},\alpha} X'_{\text{high},\alpha}}{NT} \right) > 0 \qquad \textit{wpa1}.$$

(ii) For the low-rank regressors we assume $\text{rank}(X_l) = 1$*,* $l = 1, \ldots, K_1$*, 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 (a)* $N^{-1} \lambda^{0\prime} M_v \lambda^0 > B \mathbb{I}_R$ *wpa1, and (b)* $T^{-1} f^{0\prime} M_w f^0 > B \mathbb{I}_R$ *wpa1.*

The distinction between low-rank and high-rank regressors introduced in assumption 4 is essential for showing consistency of the QMLE. The two most prominent examples of low-rank regressors are time-invariant regressors, which satisfy $X_{l,it} = X_{l,i\tau}$ for all i, t, τ, and common (or cross-sectionally invariant) regressors, which satisfy $X_{l,it} = X_{l,jt}$ for all i, j, t. To give another example of a low-rank regressor, let $D_i = \mathbb{1}(i \in \mathbb{A})$ and $\tilde{D}_t = \mathbb{1}(t \in \mathbb{B})$ 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. Interacted dummy variables of this sort appear frequently in "difference in difference" estimation. 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 $N, T \to \infty$.[4]

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 (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. To get an intuition for this assumption, consider the smallest principal angles θ_{w,λ^0} and θ_{v,f^0} between the N-dimensional subspaces spanned by w and λ^0, and the T-dimensional subspaces spanned by v and f^0, respectively.[5] It turns out that assumption 4(ii) is equivalent to demanding that both $\theta_{w,\lambda^0} > c$ and $\theta_{v,f^0} > c$ hold wpa1, for some constant $c > 0$, *i.e.* the smallest angle between the subspaces spanned by w and λ^0 is not allowed to converge to zero, and equivalently for v and f^0.[6]

A typical example of a high-rank regressor is one, where its distribution guarantees that it has full rank asymptotically, *e.g.* $X_{m,it} = 1 + Z_{it}$, where $Z_{it} \sim iid\mathcal{N}(0,1)$. However, a high-rank regressors may still have a significant "low-rank component", *e.g.* $X_{m,it} = 1 + Z_{it} + \lambda^0_{ir} f^0_{tr}$, where Z_{it} as above and λ^0_{ir} and f^0_{tr} are the r'th factor loading and factor.

Let the $K_2 \times K_2$ matrix $\tilde{W}$ be defined by $\tilde{W}_{m_1 m_2} = (NT)^{-1} \text{Tr}(X_{m_1} X'_{m_2})$, *i.e.* it is a sub-matrix of $(NT)^{-1} \sum_{i,t} X_{it} X'_{it}$. The no-collinearity condition $\text{plim}_{N,T\to\infty} \tilde{W} > 0$ would be equivalent to assumption 4 (i) on the high-rank regressors if the sum over the eigenvalues in this assumption would

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

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

[5] The concept of the principal angles between subspaces is a well known mathematical concept, see definition B.4 in the appendix. In the simplest case of only one factor ($R = 1$) and only one low-rank regressor ($K_1 = 1$) we have $\theta_{w,\lambda} = \arccos[w'\lambda/(\|w\|\|\lambda\|)]$ and $\theta_{v,f} = \arccos[v'f/(\|v\|\|f\|)]$.

[6] This statement holds conditional on assumption 1 being satisfied. For details see theorem B.5 and the proof of lemma C.2 in the appendix.

run from $i = 1$ to N, because the sum over all eigenvalues of a matrix is just its trace. Assumption 4 (i) is stricter than that since the first $2R + K_1$ eigenvalues are omitted from the sum. An immediate consequence is that high-rank regressors have to satisfy $\text{rank}(X_m) > 2R + K_1$, which explains their name.

We can now state our consistency result for the QMLE.

Theorem 2.1. *Let the assumptions 1, 2, 3, 4 be satisfied, and let the parameter set $\mathbb{B}$ be compact*[7] *In the limit $N, T \to \infty$ we then have*

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

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 assumption 1 and 2, and he uses a low-level version of assumption 3.[8] Bai (2009) demands the regressors to be strictly exogenous, but for his consistency proof this assumption is not used. 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.[9] 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 common regressors, *i.e.* that does not guarantee consistency for high-rank regressors and for more general low-rank regressors.[10] 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 (2.4) for the profile quasi likelihood function is on the one hand very convenient, because it does not involve any minimization over continuous 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, 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 $\epsilon_k = \beta_k^0 - \beta_k$, $k = 1, \ldots, K$ (the sign convention here is chosen for convenience), 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 (2.4) 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

[7] We assume compactness of $\mathbb{B}$ mainly to guarantee existence of $\hat{\beta}$. 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 (see assumption 4). Assuming boundedness of the parameter set simplifies the structure of the proof significantly, but the proof can be done without this assumption, as long as existence of $\hat{\beta}$ is guaranteed.

[8] We state assumption 3 in a high-level format because the operator norm of e is important for our expansion of L_{NT}.

[9] As argued above, for the low-rank regressors appearance of λ^0 and f^0 in assumption 4 (ii) is necessary to guarantee consistency.

[10] In appendix **?** we give two examples that show that Bai's condition $D(F^0) > 0$ does not guarantee consistency in a more general case.

$\sum_{\kappa=0}^{K} \epsilon_k \, X_\kappa/\sqrt{NT}$ 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

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

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 D. 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 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 estimator of β. This shows how our assumption on the model play together to guarantee that the above likelihood expansion is valid asymptotically.[11]

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 $L^{(2)}\left(\lambda^0, f^0, X_{\kappa_1}, X_{\kappa_2}\right) = \text{Tr}(M_{\lambda^0} X_{\kappa_1} M_{f^0} X'_{\kappa_2})$. The general formula is given in theorem D.2 in the appendix. Using this formula one can derive the following bound

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

where a_{NT} and b_{NT} are functions of λ^0 and f^0 that converge to positive constants in probability, *i.e.* $a_{NT} \to_p a > 0$ and $b_{NT} \to_p b > 0$. 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

[11] Note that all we need for this result is assumption 1, 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 assumption 2 and 4.

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 $N/T \to \kappa^2$, $\infty > \kappa > 0$, 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 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 series* $\eta_{NT} \to 0$ *we have*

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

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* $C^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) = g(4NT)^{-1/2}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_k,\, e,\, e, \ldots, e\right)$, *with* $L^{(g)}$ *defined in theorem D.2 of the appendix. For* $g = 2$ *and* $g = 3$ *we have*

$$\begin{aligned}
C^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) &= \frac{1}{\sqrt{NT}}\,\mathrm{Tr}(M_{f^0}\, e'\, M_{\lambda^0}\, X_k)\;,\\
C^{(3)}\left(\lambda^0,\, f^0,\, X_k,\, e\right) &= -\,\frac{1}{\sqrt{NT}}\Big[\mathrm{Tr}\left(eM_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\, (f^{0\prime} f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right)\\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, e\, M_{f^0}\, X'_k\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right)\\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime} f^0)^{-1}\, f^{0\prime}\right)\Big]\,.
\end{aligned} \tag{3.6}$$

In theorem D.2 of the appendix we give the general expansion of $L_{NT}(\beta)$ up to arbitrary orders in β and e. 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 one 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 needed 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 we can write the score function as*

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

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

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

Theorem 3.2 can be easily proven using our complete likelihood expansion.

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 (2.4). 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 assumption 1, 3, and 4 we have the following expansions*

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

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

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

and we have $\mathrm{rank}(\hat{e}^{(\mathrm{rem})}) \leq 6R$ PUT INTO PROOF!!!!, *and the expansion coefficients are given by*

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

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 D.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 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.

The expansions are very useful. In the present paper we make use of the expansions in theorem 3.3 in order to derive the properties of the variance and bias estimates of the QMLE below, *i.e.* of objects that contain the estimates $M_{\hat\lambda}(\beta)$, $M_{\hat f}(\beta)$, 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 (*i.e.* eigenvectors corresponding to the largest eigenvalues of the sample covariance matrix), 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 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 $\sqrt{NT}\left(\hat\beta_{k_1} - \beta^0_{k_1}\right) = W^{-1}_{NT} C_{NT} + o_p(1)$.*

Andrews (1999) provides a general discussion of the limiting distribution of extremum estimators. Once consistency of the QMLE is established, and the profile quasi likelihood expansion in theorem 3.1 is derived, one obtains the above corollary by applying theorem 3 in Andrews (1999). Defining the unrestricted minimizer of the quadratic approximation of the objective function $\hat{\beta}_q = \operatorname{argmin}_{\beta \in \mathbb{R}^K} L_{q,NT}(\beta)$, we find $\sqrt{NT}(\hat{\beta}_q - \beta^0) = W_{NT}^{-1} C_{NT}$, *i.e.* the statement of the corollary can be rewritten as $\sqrt{NT}\left(\hat{\beta} - \beta^0\right) = \sqrt{NT}\left(\hat{\beta}_q - \beta^0\right) + o_p(1)$. Thus, the bound on the remainder $R_{NT}(\beta)$ in the profile quasi likelihood expansion is such that asymptotic distribution of $\hat{\beta}$ is given by the one of $\hat{\beta}_q$. The assumptions on the model made so far guarantee that $W_{NT}^{-1} = \mathcal{O}_p(1)$, *i.e.* a direct consequence of the corollary is that the QMLE $\hat{\beta}$ is $\sqrt{NT}$-consistent if and only if $C_{NT} = \mathcal{O}_p(1)$.

Asymptotic Bias of the QMLE

Corollary 3.4 is a 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 $N/T \to \kappa^2$, $\infty > \kappa > 0$, 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 we always treat λ^0 and f^0 as non-stochastic.

We start with the discussion of the $C^{(2)}$ term. If the regressors X_k are strictly exogenous we have $\mathbb{E}C^{(2)}\left(\lambda^0, f^0, X_k, e\right) = 0$, *i.e.* no asymptotic bias originates from $C^{(2)}$ in this case. However, if the regressors are weakly exogenous we have[12]

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

To better understand the structure of these bias terms, consider model 2.2 and assume f^0 is known. By multiplying with M_{f^0} from the right we eliminate the factor term and obtain $Y M_{f^0} = \sum_{k=1}^K \beta_k X_k M_{f^0} + e M_{f^0}$. The OLS estimator (which is also the QMLE) of this equation satisfies $\sqrt{NT}(\tilde{\beta}_{k_1} - \beta^0_{k_1}) = \sum_{k_2=1}^K [V_{NT}^{-1}]_{k_1 k_2} (NT)^{-1/2}\mathrm{Tr}(M_{f^0} e' X_{k_2})$, where the $K \times K$ matrix V_{NT} is defined by $V_{NT,k_1k_2} = (NT)^{-1}\mathrm{Tr}(X_{k_1} M_{f^0} X'_{k_2})$. Assuming that V_{NT} converges to a positive definite matrix V in probability, we thus find that under weak exogeneity $\mathbb{E}\left[\sqrt{NT}(\tilde{\beta}_{k_1} - \beta^0_{k_1})\right] = -\sum_{k_2=1}^K [V^{-1}]_{k_1k_2}\mathbb{E}\left[(NT)^{-1/2}\mathrm{Tr}(P_{f^0} e' X_{k_2})\right] + o(1)$. Thus, the first type of bias we found in equation (3.14) also appears in a factor model in which the factors are observed. Such a model is the standard

[12]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.

fixed effect model if $R = 1$ and $f^0 = (1, 1, \ldots, 1)'$. For a dynamic fixed effect model this bias of the OLS estimator is well known (for fixed T asymptotics it causes inconsistency), and the standard remedy is to use IV and GMM estimators, see *e.g.* Arellano and Bond (1991). Hahn and Kuersteiner (2002) use the asymptotics $N, T \to \infty$ to characterize this bias in dynamic fixed effect models and in order to work out a biased corrected estimator. We follow the same strategy for the interactive fixed effect model.

The expectation value of the first term in (3.14) has a non-zero probability limit if $X_{k,it}$ is correlated with $e_{i\tau}$ for $t > \tau$. For example, in the AR(1) model $Y_{it} = \beta Y_{i,t-1} + \lambda_i f_t' + e_{it}$, assuming one factor ($R = 1$) that is constant, *i.e.* $f^0 = (1, 1, \ldots, 1)'$, and $-1 < \beta < 1$, and e_{it} independent across i and t with mean zero and variance σ_e^2, we find $\mathrm{Tr}\left[P_{f^0}\mathbb{E}\left(N^{-1}e' Y_-\right)\right] = \sigma_e^2\beta(1-\beta)^{-1} + o(1)$. In addition, there is the pre-factor $\sqrt{N/T}$. Thus, if T grows at a faster rate than N this asymptotic bias due to weak exogeneity vanishes; if N grows at a faster rate than T then the QMLE is not $\sqrt{NT}$-consistent, unless strict exogeneity is assumed; and if N and T grow at the same rate the QMLE is $\sqrt{NT}$-consistent, but biased.

The model is symmetric under $N \leftrightarrow T$, $\lambda^0 \leftrightarrow f^0$, $X_k' \leftrightarrow X_k$ and $e \leftrightarrow e'$. Theoretically, the discussion of the second term in (3.14) is therefore analogous to that of the first term, *i.e.* the second term also describes a bias that is due to weak exogeneity, but that is increasing in $\sqrt{T/N}$. However, practically this bias is probably less relevant, since for most applications it does not seem reasonable to assume that e_{it} is uncorrelated with X_{it} (weak exogeneity) but correlated with X_{jt} for $i \neq j$ (which is the source of this second type of bias due to weak exogeneity). Nevertheless, in some applications this my be the case, *e.g.* when the dependent variable Y_{it} for unit i appears as a regressors in the equation for Y_{jt} of unit $j \neq i$.[13]

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[14]

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

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 $\lambda^{0'} M_{\lambda^0} = 0$. 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] 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 $i > j$.

[14] Here we assume that $\mathrm{Tr}\left(eP_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\, (f^{0\prime}f^0)^{-1}\, (\lambda^{0\prime}\lambda^0)^{-1}\, \lambda^{0\prime}\right) = o_p(1)$, $\mathrm{Tr}\left(e' P_{\lambda^0}\, e\, M_{f^0}\, X_k'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime}\right) = o_p(1)$, and $\mathrm{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0\, (\lambda^{0\prime}\lambda^0)^{-1}\, (f^{0\prime}f^0)^{-1}\, f^{0\prime}\right) = o_p(1)$. 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 $\|P_{\lambda^0} e P_{\lambda^0}\| = \mathcal{O}_p(1)$, and $\|P_{\lambda^0} e X_k'\| = \mathcal{O}_p(\sqrt{NT})$.

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 our the asymptotic distribution of the QMLE $\hat{\beta}$, and to correct for the asymptotic bias. For this purpose the assumptions 1 to 4 made on the model so far are too weak, *i.e.* more specific assumptions on λ_i^0, f_t^0, X_k and e have to be made, and also the asymptotics $N, T \to \infty$ has to be specified further. These additional specifications can be made very differently, depending on the particular empirical application one has in mind. The assumptions we make in the following are clearly restrictive, but they still capture a large class of relevant models.

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 more specifically 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 $\mathbb{E}(X_{k,it}^{\text{str}})^8$ is bounded uniformly over i, t and N, T.

(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_i^0 and f_t^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 appendix 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}.[15] An example where this is satisfied is if the interactive fixed effect model is just one equation of a vector auto-regression for each cross-sectional unit, *e.g.* for the VAR(1) case we would have

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

[15] 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 a $r \times 1$ vector of additional variables, B is $(r+1)\times(r+1)$ matrix of parameters, the $r \times 1$ vectors d_{it} and u_{it} are independent of e_{it}, and Γ is a $r \times r$ 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}.[16] 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 straitforward.

Assumption 5 is not yet sufficient do guarantee existence of a limiting distribution of the QMLE $\hat{\beta}$. What is missing is the following condition that guarantees that the limiting variance and the asymptotic bias converge to constant values.

Assumption 6. *Let $\mathfrak{X}_k = M_{\lambda^0}\, X_k^{\rm str}\, M_{f^0} + X_k^{\rm weak}$ 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:*

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

We now have all assumptions that are needed to derive the asymptotic distribution of $\hat{\beta}$.

Theorem 4.1. *Let assumption 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 $X_k^{\rm weak}$ is uncorrelated with λ^0 and f^0 it does not matter whether the corresponding subspaces (of fixed dimension) are projected out of $X_k^{\rm weak}$ (whose dimension grows to infinity). For the strictly exogenous part of the regressors this is different, because $X_k^{\rm str}$ 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 $X_k^{\rm str} = 0$.) 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

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

asymptotic variance, but the weakly exogenous part contributes to both, namely to the asymptotic variance via the term $\mathrm{Tr}(e' X_k^{\mathrm{weak}})$ and to the bias B_1 via the term $\mathrm{Tr}(P_{f^0}\, e'\, X_k^{\mathrm{weak}})$. 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 3.3 we also discussed a bias to 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 $C^{(3)}$ 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 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 a well-behaved Kernel function*[17] *with $\Gamma(0) = 1$ **?**, and 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 $A^{\mathrm{truncD}} = \sum_{i=1}^{N} \eta_i\, \eta_i'\, A\, \eta_i\, \eta_i'$, $B^{\mathrm{truncD}} = \sum_{t=1}^{T} \eta_t\, \eta_t'\, B\, \eta_t\, \eta_t'$.

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

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 $\hat B_1(\beta)$, $\hat B_2(\beta)$ and $\hat B_3(\beta)$ as follows*

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

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 assumption 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. We can now present our bias corrected estimator and its limiting distribution.

[17] For the proofs of the theorems we use the truncation Kernel which is defined by $\Gamma(x) = 1$ for $\|x\| \leq 1$, and $\Gamma(x) = 0$ otherwise, but this is only to simplify notation. Other Kernel functions could be used.

[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 assumption 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 locally there are no parameter restriction on β. In the present subsection we discuss situations where β^0 is on the boundary of the $\mathbb{B}$, *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 the 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. $N/T \to \text{const.}$). 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.

(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 $\Lambda \subset \mathbb{R}^K$. 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.[19] We therefore define an alternative "bias corrected" estimator by[20]

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

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 $\sqrt{NT}\left(\hat{\beta} - \beta^0\right) \to_d \Phi$, $\sqrt{NT}\left(\hat{\beta}^{**} - \beta^0\right) \to_d \Phi^{**}$, $NT\left[L_{NT}\left(\hat{\beta}\right) - L_{NT}\left(\beta^0\right)\right] \to_d l_q(\Phi)$, $NT\left[L^{**}_{NT}\left(\hat{\beta}^{**}\right) - L^{**}_{NT}\left(\beta^0\right)\right] \to_d l_q(\Phi^{**} + W^{-1}B) - l_q(W^{-1}B)$.*

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

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

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 Hypothesis Testing

For our interactive fixed effect model, we now want to discuss the three classical test statistics for testing a general linear restriction on β, *i.e.* the null-hypothesis is $H_0 :\ H\beta^0 = h$, and the alternative

[19] Even when β^0 is an interior point of $\mathbb{B}$ one may not have $\hat{\beta}^* \in \mathbb{B}$ at finite sample, *i.e.* the estimator $\hat{\beta}^{**}$ can be useful also in this case.

[20] Alternatively one could define the new "bias corrected" estimator by using the bias corrected QMLE $\hat{\beta}^*$ that is obtained without imposing any local restrictions on β, and whose limiting distribution is given in corollary 4.5 above (the parameter set $\mathbb{B}$ used in the corollary is different from the one we consider now). By defining the new estimator as a minimizer of $(\hat{\beta}^* - \beta)'\hat{W}(\hat{\beta}^*)(\hat{\beta}^* - \beta)$ subject to $\beta \in \mathbb{B}$, one obtains an estimator that has the same limiting distribution as $\hat{\beta}^**$.

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. For easy of exposition we restrict the presentation to 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 can also discuss testing when the true parameter is on the boundary, see Andrews (2001).

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. The limiting distribution of the unrestricted estimator $\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{B}} L_{NT}(\beta)$ was given in theorem 4.1 for a specific set of assumptions on λ^0, f^0, X_k and e. But the result $\sqrt{NT}\left(\hat{\beta} - \beta^0\right) \to_d \mathcal{N}\left(W^{-1}B,\; W^{-1}\,\Omega\, W^{-1}\right)$ holds whenever assumption 7 is satisfied, because the unrestricted case is the special case of theorem 4.6 for which $\Lambda = \mathbb{R}^K$, *i.e.* $\Phi = W^{-1}C$.

We define the restricted estimator by $\tilde{\beta} = \operatorname{argmin}_{\beta \in \tilde{\mathbb{B}}} L_{NT}(\beta)$, where $\tilde{\mathbb{B}} = \{\beta \in \mathbb{B} |\, H\beta = h\}$ is the restricted parameter set. 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.* one can apply theorem 4.6 to obtain the limiting distribution of $\tilde{\beta}$. One finds $\sqrt{NT}(\tilde{\beta} - \beta^0) \to_d \tilde{\Phi}$, with $\tilde{\Phi} = \operatorname{argmin}_{\Lambda} l_q(\phi) = \mathfrak{W}^{-1}C$, and $\mathfrak{W}^{-1} = W^{-1} - W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}$.[21] Therefore $\sqrt{NT}(\tilde{\beta} - \beta^0) \to_d \mathcal{N}\left(\mathfrak{W}^{-1}B,\; \mathfrak{W}^{-1}\,\Omega\,\mathfrak{W}^{-1}\right)$.[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

$$WD^*_{NT} = \left[\sqrt{NT}\left(H\hat{\beta} - h\right) - H\hat{W}^{-1}\hat{B}\right]'\left(H\hat{W}^{-1}\,\hat{\Omega}\,\hat{W}^{-1}H'\right)^{-1}\left[\sqrt{NT}\left(H\hat{\beta} - h\right) - H\hat{W}^{-1}\hat{B}\right], \tag{4.9}$$

and under the null hypothesis we find $WD^*_{NT} \to_d \chi^2_r$ 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 assumption 5 and 6 hold, and if $\mathbb{E}e^2_{it} = \sigma^2_e = c$, *i.e.* if there is no heteroscedasticity. A consistent estimator for c in this context is $\hat{c} = \hat{\sigma}^2_e = (NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T \hat{e}_{it}$.

The likelihood ratio test statistics is defined by $LR_{NT} = \hat{c}^{-1}\, NT\left[L_{NT}\left(\tilde{\beta}\right) - L_{NT}\left(\hat{\beta}\right)\right]$. Applying theorem 4.6 we find that under assumption 7 we have

$$\begin{aligned} LR_{NT} \;\underset{d}{\longrightarrow}\; & c^{-1}\left[l(\tilde{\Phi}) - l(\Phi)\right] = c^{-1}\left[l(\mathfrak{W}^{-1}C) - l(W^{-1}C)\right] \\ & = c^{-1}C'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}C\,. \end{aligned} \tag{4.10}$$

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 define a bias correction for LR test in order to achieve a χ^2 limiting distribution.

[21] By definition we have $\tilde{\Phi} = M_{W,H'}W^{-1}C$, 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), $\mathrm{Tr}(M_{W,H'}) = K - r$ (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 $W = \mathbb{I}_K$.

[22] For the $K \times K$ covariance matrix given here we have $\mathrm{rank}(\mathfrak{W}^{-1}\,\Omega\,\mathfrak{W}^{-1}) = K - r$, because $H\mathfrak{W}^{-1} = 0$. The asymptotic distribution of $\sqrt{NT}(\tilde{\beta} - \beta^0)$ is therefore $K - r$ dimensional with support Λ.

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

$$LR^*_{NT} = \hat{c}^{-1}\left[\min_{\{\beta\in\mathbb{B}|\, H\beta=h\}} L_{NT}\left(\beta+(NT)^{-1/2}\hat{W}^{-1}\hat{B}\right) - \min_{\beta\in\mathbb{B}} L_{NT}\left(\beta+(NT)^{-1/2}\hat{W}^{-1}\hat{B}\right)\right] , \tag{4.11}$$

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

$$\begin{aligned} LR^*_{NT} \underset{d}{\longrightarrow}\ & c^{-1}\left[\min_{\{\phi\in\mathbb{R}^K|\, H\phi=0\}} l\left(\phi+W^{-1}B\right) - l\left(\Phi\right)\right] \\ &= c^{-1}\left[l\left(\mathfrak{W}^{-1}(C-B)+W^{-1}B\right) - l\left(W^{-1}C\right)\right] \\ &= c^{-1}(C-B)'W^{-1}H'(HW^{-1}H')^{-1}HW^{-1}(C-B) , \end{aligned} \tag{4.12}$$

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 we therefore find $LR^*_{NT}\to_d \chi^2_r$. One can show that $LR^*_{NT} = WD^*_{NT} + o_p(1)$.

Lagrange Multiplier Test

Using the bound on ∇R_{NT} given in theorem 3.2 and the fact that the restricted estimator $\tilde{\beta}$ is $\sqrt{NT}$-consistent, we immediately find $\sqrt{NT}\nabla L_{NT}(\tilde{\beta}) = \sqrt{NT}\nabla L_{q,NT}(\tilde{\beta})+o_p(1)$. We have $\sqrt{NT}\nabla L_{q,NT}(\tilde{\beta}) = 2\sqrt{NT}W_{NT}(\tilde{\beta}-\beta^0)-2C_{NT}$, and under assumption 7 we showed that $\sqrt{NT}(\tilde{\beta}-\beta^0)\to_d \mathfrak{W}^{-1}C$. Therefore we find $\sqrt{NT}\nabla L_{NT}(\tilde{\beta})\to_d -2H'(HW^{-1}H')^{-1}HW^{-1}C$, and $\sqrt{NT}HW^{-1}\nabla L_{NT}(\tilde{\beta})\to_d -2HW^{-1}C$.

The LM test statistics is given by $LM_{NT} = NT/4\,\nabla L_{NT}(\tilde{\beta})'\tilde{W}^{-1}\tilde{H}'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}HW^{-1}\nabla L_{NT}(\tilde{\beta})$, where $\tilde{B}=\hat{B}(\tilde{\beta})$, $\tilde{W}=\hat{W}(\tilde{\beta})$ and $\tilde{\Omega}=\hat{\Omega}(\tilde{\beta})$. One can be 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[24]

$$LM^*_{NT} = \frac{1}{4}\,(\sqrt{NT}\,\nabla L_{NT}(\tilde{\beta})+\tilde{B})'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}(\sqrt{NT}\,\nabla L_{NT}(\tilde{\beta})+\tilde{B}) , \tag{4.13}$$

Under the null hypothesis, if assumption 7 holds, and if $R_{NT}(\beta)$ satisfies the bound in equation 3.7 we have $LM^*_{NT}\to_d \chi^2_r$.

In order to apply LM^*_{NT} as defined above one needs the gradient of $L_{NT}(\beta)$ at $\tilde{\beta}$. Since no explicit expression for the derivatives of $L_{NT}(\beta)$ exists, the gradient needs to be calculated numerically, which may be inconvenient. We therefore propose to use an approximation of the gradient that is much easier to compute. Define the K-vector $\tilde{\nabla}L_{NT}(\beta)$ by $(\tilde{\nabla}L_{NT}(\beta))_k = -2\mathrm{Tr}(X_k'\tilde{e}(\beta))$ for $k=1,\ldots,K$, and define the modified bias corrected LM test as

$$LM^{**}_{NT} = \frac{1}{4}\,(\sqrt{NT}\,\tilde{\nabla}L_{NT}(\tilde{\beta})+\tilde{B})'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}(\sqrt{NT}\,\tilde{\nabla}L_{NT}(\tilde{\beta})+\tilde{B}) . \tag{4.14}$$

[23] Here one could also use $\hat{B}(\tilde{\beta})$ and $\hat{W}(\tilde{\beta})$ as estimates B and W, but these estimators are not consistent if the null hypothesis is false. Since we want to test properties of the regression parameters and not properties of the estimators $\hat{B}$ and $\hat{W}$, it is reasonable to use estimators that are robust towards violation of H_0.

[24] Alternatively, one could define the bias corrected LM test as $LM^*_{NT} = NT/4\,\nabla L_{NT}(\tilde{\beta}+\tilde{W}^{-1}\tilde{B})'\tilde{W}^{-1}H'(H\tilde{W}^{-1}\tilde{\Omega}\tilde{W}^{-1}H')^{-1}H\tilde{W}^{-1}\nabla L_{NT}(\tilde{\beta}+\tilde{W}^{-1}\tilde{B})$ and would obtain the same limiting distribution.

Theorem 4.7. *Let assumption* **?** *be satisfied, and let $\tilde{\beta}$ be an estimator that satisfies $\sqrt{NT}(\tilde{\beta}-\beta^0) = o_p(1)$. Then we have $\sqrt{NT}\nabla L_{NT}(\tilde{\beta}) = \sqrt{NT}\hat{\nabla} L_{NT}(\tilde{\beta}) + o_p(1)$.*

This theorem guarantees that LM^{**}_{NT} has the same limiting distribution as LM^{*}_{NT}, *i.e.* $LM^{**}_{NT} \to_d \chi^2_r$.

5 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 that 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 estimators of the asymptotic bias of β.

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 chi-square due to the asymptotic bias of the score and of the QMLE. We provide bias corrected test statistics and show that their asymptotic distribution is chi-squared.

In future work, the most important extension of the present paper will be to study situations where the true number of factors R is not known but has to be estimated. There is a sizable literature on the estimation of the number of factors in approximated factor models, *e.g.* Bai and Ng (2002), Onatski (2005), but none of these papers estimates the number of factors jointly with additional regression coefficients.

A Examples of Error Distributions

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

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

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

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

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

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

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

(iii) The error matrix e is generated as $e = \sigma^{1/2} u \Sigma^{1/2}$, where u is an $N \times T$ matrix with independently distributed entries u_{it} and $\mathbb{E}u_{it} = 0$, $\mathbb{E}u_{it}^2 = 1$, and $\mathbb{E}u_{it}^4$ is bounded uniformly across i, t and

N, T. Here σ is the $N \times N$ cross-sectional covariance matrix, and Σ is $T \times T$ time-serial covariance matrix, and they satisfy

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

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

B Some Matrix Algebra

Theorem B.1 (Special Case of Weyl's Inequalities)**.** *Let A and B be real symmetric matrices of dimensions n. For all $i = 1, \cdots, n$ we then have*

$$\mathrm{Eigval}_i(A) + \mathrm{Eigval}_n(B) \; \leq \; \mathrm{Eigval}_i(A+B) \; \leq \; \mathrm{Eigval}_i(A) + \mathrm{Eigval}_1(B)$$

For the proof see *e.g.* Bhatia (1997). Two immediate implications of theorem B.1 are the following. First, if B is positive semi-definite, then $\mathrm{Eigval}_i(A) \leq \mathrm{Eigval}_i(A+B)$. Second, since the operator norm of B is always larger or equal then the absolute value of any eigenvalue of B we have $\mathrm{Eigval}_i(A) - \|B\| \leq \mathrm{Eigval}_i(A+B) \leq \mathrm{Eigval}_i(A) + \|B\|$.

Now let A be an arbitrary $n \times m$ matrix. In addition to the operator (or spectral) norm $\|A\|$ and to the Frobenius (or Hilbert-Schmidt) norm $\|A\|_F$, it is also convenient to define the 1-norm, the ∞-norm, and the max-norm as

$$\|A\|_1 \; = \max_{j=1\ldots m} \sum_{i=1}^{n} |A_{ij}| \; , \quad \|A\|_\infty \; = \max_{i=1\ldots n} \sum_{j=1}^{m} |A_{ij}| \; , \quad \|A\|_{\max} \; = \max_{i=1\ldots n} \max_{j=1\ldots m} |A_{ij}| \; . \tag{B.1}$$

Theorem B.2 (Some useful Inequalities)**.** *Let A be a $n \times m$ matrix, B be a $m \times p$ matrix, and C and D be $n \times n$ matrices Then we have*

$$\begin{array}{ll}
(i) & \|A\| \; \leq \; \|A\|_F \; \leq \; \|A\| \, \mathrm{rank}\,(A)^{1/2} \; , \\
(ii) & \|AB\| \; \leq \; \|A\| \, \|B\| \; , \\
(iii) & \|AB\|_F \; \leq \; \|A\|_F \, \|B\| \; \leq \; \|A\|_F \, \|B\|_F \; , \\
(iv) & \mathrm{Tr}(AB) \; \leq \; \|A\|_F \, \|B\|_F \; , \qquad \text{if } n = p, \\
(v) & |\mathrm{Tr}\,(C)| \leq \|C\| \, \mathrm{rank}\,(C) \; , \\
(vi) & \|C\| \leq \mathrm{Tr}\,(C) \; , \qquad \text{if } C \text{ symmetric and } C \geq 0, \\
(vii) & \|A\|^2 \; \leq \; \|A\|_1 \, \|A\|_\infty \; , \\
(viii) & \|A\|_{\max} \; \leq \; \|A\| \; \leq \; \sqrt{nm} \, \|A\|_{\max} \; , \\
(ix) & \|A'CA\| \leq \|A'DA\| \; , \qquad \text{if } C \text{ is symmetric and } C \leq D.
\end{array}$$

Theorem B.3. *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. Define

$$
\begin{aligned}
S_1(Z) &= \inf_{f,\lambda} \operatorname{Tr}\left[\left(Z-\lambda f'\right)\left(Z'-f\lambda'\right)\right] \ , \\
S_2(Z) &= \inf_{f} \operatorname{Tr}(Z\, M_f\, Z') \ , \\
S_3(Z) &= \inf_{\lambda} \operatorname{Tr}(Z'\, M_\lambda\, Z) \ , \\
S_4(Z) &= \inf_{\tilde{\lambda},\tilde{f}} \operatorname{Tr}(M_{\tilde{\lambda}}\, Z\, M_{\tilde{f}}\, Z') \ , \\
S_5(Z) &= \sum_{i=1}^{N-R-1} \operatorname{Eigval}_i(Z'Z) \ , \\
S_6(Z) &= \sum_{i=1}^{N-R-1} \operatorname{Eigval}_i(ZZ') \ ,
\end{aligned} \tag{B.2}
$$

Then we have

$$
S_1(Z) \ = \ S_2(Z) \ = \ S_3(Z) \ = \ S_4(Z) \ = \ S_5(Z) \ = \ S_6(Z) \ , \tag{B.3}
$$

Note that we do not have to restrict ourselves to matrices λ, f, $\tilde{\lambda}$ and $\tilde{f}$ of full rank in the above minimization problems. 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 $\operatorname{rank}(M_\lambda) + \operatorname{rank}(\lambda) = N$. If $\operatorname{rank}(Z) \geq R$ then the optimal λ, f, $\tilde{\lambda}$ and $\tilde{f}$ have full rank.

Definition B.4. *Let A be an $n \times r_1$ matrix and B be an $n \times r_2$ matrix with* $\operatorname{rank}(A) = r_1$ *and* $\operatorname{rank}(B) = r_2$. *The smallest principal angle $\theta_{A,B} \in [0, \pi/2]$ between the linear subspaces* $\operatorname{span}(A) = \{Aa |\, a \in \mathbb{R}^{r_1}\}$ *and* $\operatorname{span}(B) = \{Bb |\, b \in \mathbb{B}^{r_2}\}$ *of $\mathbb{R}^n$ is defined by*

$$
\cos(\theta_{A,B}) = \max_{0 \neq a \in \mathbb{R}^{r_1}} \max_{0 \neq b \in \mathbb{R}^{r_2}} \frac{a'A'Bb}{\|Aa\| \|Bb\|} \ . \tag{B.4}
$$

Theorem B.5. *Let A be an $n \times r_1$ matrix and B be an $n \times r_2$ matrix with* $\operatorname{rank}(A) = r_1$ *and* $\operatorname{rank}(B) = r_2$. *Then we have the following alternative characterizations of the smallest principal angle between* $\operatorname{span}(A)$ *and* $\operatorname{span}(B)$

$$
\begin{aligned}
\sin(\theta_{A,B}) &= \min_{0 \neq a \in \mathbb{R}^{r_1}} \frac{\|M_B\, A\, a\|}{\|A\, a\|} \\
&= \min_{0 \neq b \in \mathbb{R}^{r_2}} \frac{\|M_A\, B\, b\|}{\|A\, b\|} \ .
\end{aligned} \tag{B.5}
$$

Proof. Since $\|M_B\, A\, a\|^2 + \|P_B\, A\, a\|^2 = \|A\, a\|^2$ and $\sin(\theta_{A,B})^2 + \cos(\theta_{A,B})^2 = 1$, we find that proving the theorem is equivalent to proving

$$
\cos(\theta_{A,B}) = \min_{0 \neq a \in \mathbb{R}^{r_1}} \frac{\|P_B\, A\, a\|}{\|A\, a\|} = \min_{0 \neq b \in \mathbb{R}^{r_2}} \frac{\|P_A\, B\, b\|}{\|A\, b\|} \ . \tag{B.6}
$$

This result is theorem 8 in Galantai, Hegedus (2006), and the proof can, for example, be found there. ∎

C Proof of Consistency (Theorem 2.1)

For series (of random variables) $a = a_{NT}$ and $b = b_{NT}$ we use the notation $a = \mathcal{O}_{p,+}(b)$ if $a = \mathcal{O}_p(b)$ and $a > 0$. Analogously we define $o_{p,+}(b)$. In addition we use the notation $a = \mathcal{O}_{p,++}(b)$ if $c_1 > a/b > c_2$ wpa1, for some positive constant c_1 and c_2.

Following Bai (2009), we first show

Lemma C.1. *Under the assumptions of theorem 2.1 we have*

$$
\begin{aligned}
\max_f \left| \frac{1}{NT} \text{Tr}(X_k \, M_f \, e') \right| &= o_p(1) \, , \\
\max_f \left| \frac{1}{NT} \text{Tr}(\lambda^0 \, f^{0\prime} \, M_f \, e') \right| &= o_p(1) \, , \\
\max_f \left| \frac{1}{NT} \text{Tr}(e \, P_f \, e') \right| &= o_p(1) \, ,
\end{aligned} \tag{C.1}
$$

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

Proof. By assumption 2 we know that the first two equations in lemma C.1 are satisfied when replacing M_f by the identity matrix. So we are left to show $\max_f \left| \frac{1}{NT} \text{Tr}(\Xi \, P_f \, e') \right| = o_p(1)$, 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 assumption 1, 3, and 4, respectively. Therefore

$$
\max_f \left| \frac{1}{NT} \text{Tr}(\Xi \, P_f \, e') \right| \leq R \, \frac{\|e\|}{\sqrt{NT}} \, \frac{\|\Xi\|}{\sqrt{NT}} = o_p(1) \, . \tag{C.2}
$$

∎

Lemma C.2. *Under assumption 1 and 4 there exists a constant* $B_0 > 0$ *such that*

$$
\begin{aligned}
w' \, M_{\lambda^0} \, w \; - \; B_0 \, w' \, v \geq 0 \, , \qquad & \textit{wpa1,} \\
v' \, M_{f^0} \, v \; - \; B_0 \, v' \, v \geq 0 \, , \qquad & \textit{wpa1.}
\end{aligned} \tag{C.3}
$$

Proof. Note that w may not have full rank, *e.g.* because $w_2 = w_3$. We can decompose $w = \tilde{w} \, \bar{w}$, where $\tilde{w}$ is a $N \times \text{rank}(w)$ matrix and $\bar{w}$ is a $\text{rank}(w) \times K_1$ matrix. Note that $\tilde{w}$ has full rank, and $M_w = M_{\tilde{w}}$.

By assumption 1 (i) we know that $\lambda^{0\prime}\lambda^0/N$ has a probability limit, i.e there exists some $B_1 > 0$ such that $\lambda^{0\prime}\lambda^0/N < B_1 \mathbb{I}_R$ wpa1. Using this and assumption 4 we find that for any $R \times 1$ vector $a \neq 0$ we have

$$
\frac{\|M_v \, \lambda^0 \, a\|^2}{\|\lambda^0 \, a\|^2} \; = \; \frac{a' \, \lambda^{0\prime} \, M_v \, \lambda^0 \, a}{a' \, \lambda^{0\prime} \, \lambda^0 \, a} > \frac{B}{B_1} \qquad \text{wpa1.} \tag{C.4}
$$

Applying theorem B.5 we find

$$
\min_{0 \neq b \in \mathbb{R}^{\text{rank}(w)}} \frac{b' \, \tilde{w}' \, M_{\lambda^0} \, \tilde{w} \, b}{b' \, \tilde{w}' \, \tilde{w} \, b} \; = \; \min_{0 \neq a \in \mathbb{R}^R} \frac{a' \, \lambda^{0\prime} \, M_w \, \lambda^0 \, a}{a' \, \lambda^{0\prime} \, \lambda^0 \, a} > \frac{B}{B_1} \qquad \text{wpa1.} \tag{C.5}
$$

Therefore we have wpa1

$$
b' \left(\tilde{w}' \, M_{\lambda^0} \, \tilde{w} - \frac{B}{B_1} \tilde{w}' \tilde{w} \right) b > 0 \, , \tag{C.6}
$$

for every $\text{rank}(w) \times 1$ vector b, and thus

$$
\tilde{w}' \, M_{\lambda^0} \, \tilde{w} - \frac{B}{B_1} \, \tilde{w}' \, \tilde{w} > 0 \, , \qquad \text{wpa1.} \tag{C.7}
$$

Therefore also (multiplying from the left with $\bar{w}'$ and from the right with $\bar{w}$)

$$
w' \, M_{\lambda^0} \, w - \frac{B}{B_1} \, w' \, w \geq 0 \, , \qquad \text{wpa1.} \tag{C.8}
$$

Analogously we can show

$$
v' \, M_{f^0} \, v - B_0 \, v' \, v > 0 \, , \qquad \text{wpa1,} \tag{C.9}
$$

for some positive constant B_0. ∎

As a consequence of the last results we obtain some properties of the low-rank regressors summarized in the following lemma.

Lemma C.3. *Let the assumptions 1 and 4 be satisfied and let $X_{\mathrm{low},\alpha} = \sum_{l=1}^{K_1} \alpha_l X_l$ be linear combinations of the low-rank regressors. Then there exists some constant $B > 0$ such that*

$$\min_{\{\alpha \in \mathbb{R}^{K_1}, \|\alpha\|=1\}} \frac{\left\| X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha} \right\|}{NT} > B \,, \qquad \textit{wpa1},$$

$$\min_{\{\alpha \in \mathbb{R}^{K_1}, \|\alpha\|=1\}} \frac{\left\| M_{\lambda^0}\, X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha}\, M_{\lambda^0} \right\|}{NT} > B \,, \qquad \textit{wpa1}. \tag{C.10}$$

Proof. Note that $\left\| M_{\lambda^0}\, X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha}\, M_{\lambda^0} \right\| \leq \left\| X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha} \right\|$, because $\|M_{\lambda^0}\| = 1$, *i.e.* if we can show the second inequality of the lemma we have also shown the first inequality.

We can write $X_{\mathrm{low},\alpha} = w\, \mathrm{diag}(\alpha')\, v'$. Using lemma C.2 and part (v), (vi) and (xi) of theorem B.2 we find

$$\begin{aligned}
&\left\| M_{\lambda^0}\, X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha}\, M_{\lambda^0} \right\| \\
&\quad = \left\| M_{\lambda^0}\, w\, \mathrm{diag}(\alpha')\, v'\, M_{f^0}\, v\, \mathrm{diag}(\alpha')\, w' M_{\lambda^0} \right\| \\
&\quad \geq B_0\, \left\| M_{\lambda^0}\, w\, \mathrm{diag}(\alpha')\, v'\ v\, \mathrm{diag}(\alpha')\, w' M_{\lambda^0} \right\| \\
&\quad \geq \frac{B_0}{K_1}\, \mathrm{Tr}\left[M_{\lambda^0}\, w\, \mathrm{diag}(\alpha')\, v'\ v\, \mathrm{diag}(\alpha')\, w' M_{\lambda^0} \right] \\
&\quad = \frac{B_0}{K_1}\, \mathrm{Tr}\left[v\, \mathrm{diag}(\alpha')\, w' M_{\lambda^0} w\, \mathrm{diag}(\alpha')\, v' \right] \\
&\quad \geq \frac{B_0}{K_1}\, \left\| v\, \mathrm{diag}(\alpha')\, w' M_{\lambda^0} w\, \mathrm{diag}(\alpha')\, v' \right\| \\
&\quad \geq \frac{B_0^2}{K_1}\, \left\| v\, \mathrm{diag}(\alpha')\, w' w\, \mathrm{diag}(\alpha')\, v' \right\| \\
&\quad \geq \frac{B_0^2}{K_1^2}\, \mathrm{Tr}\left[v\, \mathrm{diag}(\alpha')\, w' w\, \mathrm{diag}(\alpha')\, v' \right] \\
&\quad = \frac{B_0^2}{K_1^2}\, \mathrm{Tr}\left[X_{\mathrm{low},\alpha} X'_{\mathrm{low},\alpha} \right]
\end{aligned} \tag{C.11}$$

Thus we have

$$\frac{\left\| M_{\lambda^0}\, X_{\mathrm{low},\alpha}\, M_{f^0}\, X'_{\mathrm{low},\alpha}\, M_{\lambda^0} \right\|}{NT} \geq \frac{B_0^2}{K_1^2}\, \alpha'\, W^{\mathrm{low}}\, \alpha \,, \tag{C.12}$$

where the $K_1 \times K_1$ matrix W^{low} is defined by $W^{\mathrm{low}}_{l_1 l_2} = (NT)^{-1} \mathrm{Tr}\left(X_{l_1} X'_{l_2} \right)$. Since by assumption W^{low} converges to a positive definite matrix, the above inequality proves the lemma. ∎

For the second version of the profile quasi likelihood function in equation (2.4) we write

$$L_{NT}(\beta) = \inf_f\, S_{NT}(\beta, f) \,, \tag{C.13}$$

where

$$S_{NT}(\beta, f) = \frac{1}{NT}\, \mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k + e \right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K} (\beta_k^0 - \beta_k) X_k + e \right)' \right] , \tag{C.14}$$

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

$$\begin{aligned} S_{NT}(\beta, f) &= S_{NT}(\beta^0, f^0) + \tilde{S}_{NT}(\beta, f) \\ &\quad + \frac{2}{NT}\,\mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)\, M_f\, e'\right] + \frac{1}{NT}\,\mathrm{Tr}\left(e\,(P_{f^0} - P_f)\, e'\right) \\ &= S_{NT}(\beta^0, f^0) + \tilde{S}_{NT}(\beta, f) + o_p(\|\beta - \beta^0\|) + o_p(1)\ , \end{aligned} \tag{C.15}$$

where we defined

$$\tilde{S}_{NT}(\beta, f) = \frac{1}{NT}\,\mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)'\right]. \tag{C.16}$$

According to assumption 4 we can split the regressors into low- and high-rank regressors which gives $\sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k = \sum_{l=1}^{K_1}(\beta_l^0 - \beta_l)X_l + \sum_{m=K_1+1}^{K}(\beta_m^0 - \beta_m)X_m$. Since the low-rank regressors have finite rank we write $X_l = w_l\, v_l'$, where w_l is a $N \times \mathrm{rank}(X_l)$ matrix and v_l is a $T \times \mathrm{rank}(X_l)$ matrix. We also define the $N \times K_1$ matrix $W = (w_1, w_2, \ldots, w_{K_1})$ which combines all first components of the low-rank regressors. We then have $M_\Xi = M_W$.

We split $\tilde{S}_{NT}(\beta, f) = \tilde{S}^{(1)}_{NT}(\beta, f) + \tilde{S}^{(2)}_{NT}(\beta, f)$, where

$$\begin{aligned} \tilde{S}^{(1)}_{NT}(\beta, f) &= \frac{1}{NT}\,\mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)'\, M_{(\lambda_0, W)}\right] \\ &= \frac{1}{NT}\,\mathrm{Tr}\left[\left(\sum_{m=K_1+1}^{K}(\beta_m^0 - \beta_m)X_m\right)\, M_f\, \left(\sum_{m=K_1+1}^{K}(\beta_m^0 - \beta_m)X_m\right)'\, M_{(\lambda_0, W)}\right]. \\ \tilde{S}^{(2)}_{NT}(\beta, f) &= \frac{1}{NT}\,\mathrm{Tr}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)\, M_f\, \left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta_k^0 - \beta_k)X_k\right)'\, P_{(\lambda_0, W)}\right]\ , \end{aligned} \tag{C.17}$$

and (λ_0, W) is the $N \times (R + K_1)$ matrix that is composed out of λ_0 and W.

Applying theorem B.3 and using the definitions in assumption 4(i) we find

$$\begin{aligned} \tilde{S}^{(1)}_{NT}(\beta, f) &\geq \frac{1}{NT}\sum_{i=2R+K_1+1}^{N}\mathrm{Eigval}_i\left[\left(\sum_{m=K_1+1}^{K}(\beta_m^0 - \beta_m)X_m\right)\left(\sum_{m=K_1+1}^{K}(\beta_m^0 - \beta_m)X_m\right)'\right] \\ &\geq\ \mathcal{O}_{p,++}\left(\left\|\beta^{\mathrm{high}} - \beta_0^{\mathrm{high}}\right\|^2\right)\ , \end{aligned} \tag{C.18}$$

where β^{high} refers to the $K_2 \times 1$ parameter vector that corresponds to the high-rank regressors, and similarly we use β^{low} for the $K_1 \times 1$ parameter vector of low-rank regressors.

Applying theorems B.3 and B.1 we find

$$
\begin{aligned}
\tilde{S}^{(2)}_{NT}(\beta, f) &\geq \frac{1}{NT}\,\mathrm{Eigval}_{R+1}\left[\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta^0_k - \beta_k) X_k\right)'\, P_{(\lambda_0,W)}\,\left(\lambda^0\, f^{0\prime} + \sum_{k=1}^{K}(\beta^0_k - \beta_k) X_k\right)\right] \\
&= \frac{1}{NT}\,\mathrm{Eigval}_{R+1}\Bigg[\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)'\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right) \\
&\qquad + \left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)' P_{(\lambda_0,W)} \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m \\
&\qquad + \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m' P_{(\lambda_0,W)}\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right) \\
&\qquad + \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m' P_{(\lambda_0,W)} \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m\Bigg] \\
&\geq \frac{1}{NT}\,\mathrm{Eigval}_{R+1}\Bigg[\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)'\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right) \\
&\qquad + \left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)' P_{(\lambda_0,W)} \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m \\
&\qquad + \sum_{m=K_1}^{K}(\beta^0_m - \beta_m) X_m' P_{(\lambda_0,W)}\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)\Bigg] \\
&\geq \frac{1}{NT}\,\mathrm{Eigval}_{R+1}\left[\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)'\left(\lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_k - \beta_k) w_l\, v_l'\right)\right] \\
&\quad - \mathcal{O}_{p,+}\left(\left\|\beta^{\mathrm{high}} - \beta^{\mathrm{high}}_0\right\|\right) \qquad \text{(C.19)}
\end{aligned}
$$

where we used $(NT)^{-1}\, K_2\, \|\lambda^0\, f^{0\prime}\| \max_m\, \|X_m\| = \mathcal{O}_{p,+}(1)$, $(NT)^{-1}\, K_1\, K_2\, \max_m\, \|X_m\|\, \max_l\, \|w_l\, v_l'\| = \mathcal{O}_{p,+}(1)$, and we have $\mathcal{O}_{p,+}\left(\left\|\beta^{\mathrm{low}} - \beta^{\mathrm{low}}_0\right\|\right) = \mathcal{O}_{p,+}(1)$ because by assumption $\left\|\beta^{\mathrm{low}} - \beta^{\mathrm{low}}_0\right\|$ is bounded.

We define

$$
A \equiv \lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_l - \beta_l)\, w_l\, v_l' = A_1 + A_2 + A_3\ , \qquad \text{(C.20)}
$$

where

$$
\begin{aligned}
A_1 &= M_W\, A\, P_{f^0} = M_W\, \lambda^0\, f^{0\prime} \\
A_2 &= P_W\, A\, M_{f^0} = \sum_{l=1}^{K_1}(\beta^0_l - \beta_l)\, w_l\, v_l'\, M_{f^0} \\
A_3 &= P_W\, A\, P_{f^0} = P_W\, \lambda^0\, f^{0\prime} + \sum_{l=1}^{K_1}(\beta^0_l - \beta_l)\, w_l\, v_l'\, P_f \qquad \text{(C.21)}
\end{aligned}
$$

We find

$$
\begin{aligned}
A'A \;&\geq\; A'A - (a^{1/2} A_3' + a^{-1/2} A_2')(a^{1/2} A_3 + a^{-1/2} A_2) \\
&= [A_1' A_1 - (a-1)\, A_3' A_3] \;+\; (1 - a^{-1}) A_2' A_2 \qquad \text{(C.22)}
\end{aligned}
$$

where $\geq$ for matrices refers to the difference being positive definite, and a is a positive number, namely

$$a = 1 + \frac{\text{Eigval}_R(A_1' A_1)}{2\,\|A_3\|^2} \ . \tag{C.23}$$

The reason for this choice becomes clear below.

Note that $[A_1' A_1 - (a-1)\, A_3' A_3]$ has at most rank R (asymptotically it has exactly rank R). The non-zero eigenvalues of $A'A$ are therefore given by the (at most) R non-zero eigenvalues of $[A_1' A_1 - (a-1)\, A_3' A_3]$ and the non-zero eigenvalues of $(1-a^{-1})A_2' A_2$, the largest one of the latter being given given by the operator norm $(1-a^{-1})\|A_2' A_2\|$. We therefore find

$$\begin{aligned}\frac{1}{NT}\,\text{Eigval}_{R+1}\,(A'A) &\geq \frac{1}{NT}\,\text{Eigval}_{R+1}\left[(A_1' A_1 - (a-1)\, A_3' A_3) \ + \ (1-a^{-1})A_2' A_2\right] \\ &\geq \frac{1}{NT}\,\min\left\{(1-a^{-1})\|A_2\|^2 \ , \ \ \text{Eigval}_R\,[A_1' A_1 - (a-1)\, A_3' A_3]\right\} \ . \end{aligned} \tag{C.24}$$

Using theorem B.1 and our particular choice of a we find

$$\begin{aligned}\text{Eigval}_R\ [A_1' A_1 - (a-1)\, A_3' A_3] &\geq \text{Eigval}_R(A_1' A_1) - \|(a-1)A_3' A_3\| \\ &\geq \frac{1}{2}\,\text{Eigval}_R(A_1' A_1) \ . \end{aligned} \tag{C.25}$$

Therefore

$$\begin{aligned}\tilde{S}^{(2)}_{NT}(\beta, f) &\geq \frac{1}{2\,NT}\,\text{Eigval}_R(A_1' A_1)\,\min\left\{1 \ , \ \ \frac{2\,\|A_2\|^2}{2\,\|A_3\|^2 + \text{Eigval}_R(A_1' A_1)}\right\} - \mathcal{O}_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) \\ &\geq \frac{1}{NT}\,\frac{\|A_2\|^2\,\text{Eigval}_R(A_1' A_1)}{2\,\|A\|^2 + \text{Eigval}_R(A_1' A_1)} - \mathcal{O}_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) \end{aligned} \tag{C.26}$$

where we used $\|A\| \geq \|A_3\|$ and $\|A\| \geq \|A_1\|$. We have

$$\begin{aligned}\frac{\|A\|}{\sqrt{NT}} &\leq \frac{\|\lambda^0\, f^{0\prime}\|}{\sqrt{NT}} + \sum_{l=1}^{K_1} |\beta_l^0 - \beta_l|\,\frac{\|w_l\, v_l'\|}{\sqrt{NT}} \leq \mathcal{O}_{p,++}(1) + \mathcal{O}_{p,+}\left(\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\|\right) = \mathcal{O}_{p,++}(1) \ , \\ \frac{\text{Eigval}_R(A_1' A_1)}{\sqrt{NT}} &= \frac{\text{Eigval}_R\left(f^0\, \lambda^{0\prime}\, M_W\, \lambda^0\, f^{0\prime}\right)}{\sqrt{NT}} = \mathcal{O}_{p,++}(1) \ , \\ \frac{\|A_2\|^2}{NT} &= \text{Eigval}_1\left[\sum_{l_1=1}^{K_1}(\beta^0_{l_1} - \beta_{l_1})\, w_{l_1}\, v_{l_1}'\, M_{f^0}\, \sum_{l_2=1}^{K_1}(\beta^0_{l_2} - \beta_{l_2})\, v_{l_2}\, w_{l_2}'\right] \\ &\geq \mathcal{O}_{p,++}\left(\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\|^2\right) = \mathcal{O}_{p,++}(1) \end{aligned} \tag{C.27}$$

Here we used assumption 4 (ii) and again the boundedness of β^{low}. We thus have

$$\tilde{S}^{(2)}_{NT}(\beta, f) \geq \mathcal{O}_{p,++}\left(\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\|^2\right) - \mathcal{O}_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) \ , \tag{C.28}$$

Since $\beta = \beta^0$ and $f = f^0$ are possible choices for $\hat{\beta}$ and $\hat{f}$, we find that the optimal $\hat{\beta}$ and $\hat{f}$ must satisfy $S_{NT}(\hat{\beta}, \hat{f}) \leq S_{NT}(\beta^0, f^0)$. From (C.15) we thus find

$$\begin{aligned} 0 \ &\geq \ \tilde{S}_{NT}(\hat{\beta}, \hat{f}) + o_p\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) + o_p(1) \\ &\geq \ \tilde{S}_{NT}(\hat{\beta}, \hat{f}) - o_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) - o_{p,+}(1) \end{aligned} \tag{C.29}$$

Using (C.18), (C.28) and $\tilde{S}^{(2)}_{NT}(\beta, f) \geq 0$ we thus find

$$\begin{aligned} 0 \ \geq \ &\mathcal{O}_{p,++}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|^2\right) - o_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) - o_{p,+}(1) \\ &+ \max\left\{0\,, \ \mathcal{O}_{p,++}\left(\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\|^2\right) - \mathcal{O}_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right)\right\} \end{aligned} \tag{C.30}$$

In particular this implies

$$0 \geq \mathcal{O}_{p,++}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|^2\right) - o_{p,+}\left(\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\|\right) - o_{p,+}(1) \;, \tag{C.31}$$

from which we can deduce[25]

$$\left\|\beta^{\text{high}} - \beta_0^{\text{high}}\right\| = o_p(1) \;. \tag{C.34}$$

Once we have this we find that (C.30) becomes

$$0 \geq \mathcal{O}_{p,++}\left(\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\|^2\right) - o_{p,+}(1) \;, \tag{C.35}$$

and therefore

$$\left\|\beta^{\text{low}} - \beta_0^{\text{low}}\right\| = o_p(1) \;. \tag{C.36}$$

D Power Series Expansion of the Profile Quasi Likelihood Function (Proofs of Theorems 3.1, 3.3 and Corollary 3.4)

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

$$\begin{aligned} d_{\max}(\lambda^0, f^0) &= \frac{1}{\sqrt{NT}} \left\|\lambda^0 f^{0\prime}\right\| = \frac{1}{\sqrt{NT}} \sqrt{\text{Eigval}_1(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \;, \\ d_{\min}(\lambda^0, f^0) &= \frac{1}{\sqrt{NT}} \sqrt{\text{Eigval}_R(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \;, \end{aligned} \tag{D.1}$$

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 given by

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

Why $r_0(\lambda^0, f^0)$ is called convergence radius will become clear immediately.

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

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

then

[25] To give a proper proof we note that (C.31) implies that there exists $w > 0$ such that for all $\varepsilon > 0$, $\delta > 0$ and $\kappa > 0$ we have

$$\lim_{N,T \to \infty} \Pr\left[0 \;\geq\; (w - \varepsilon)\|\hat{\beta}^{\text{high}} - \beta_0^{\text{high}}\|^2 - \delta\,\|\beta - \beta^0\| - \kappa\right] \;=\; 1 \;, \tag{C.32}$$

so we find

$$\Rightarrow \quad \lim_{N,T \to \infty} \Pr\left[\|\hat{\beta}^{\text{high}} - \beta_0^{\text{high}}\| \leq \frac{\delta}{2\,(w - \varepsilon)} + \sqrt{\frac{\kappa}{w - \varepsilon} + \frac{\delta^2}{4\,(w - \varepsilon)^2}}\right] \;=\; 1 \;, \tag{C.33}$$

where we have already chosen $\varepsilon < w$. In addition, we can choose δ and κ arbitrarily small, so we have shown $\hat{\beta}^{\text{high}} \to_p \beta_0^{\text{high}}$.

(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

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

where the expansion coefficients are given by[26]

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

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

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

with

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

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

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

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

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

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

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

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

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

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

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

Under the stronger condition

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

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

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

where

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

(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 $g \geq 1$*

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

Under the condition (D.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 $G \geq 0$*

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

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 $d = NT d_{\min}^2(\lambda^0, f^0)$. The bound (D.3) guarantees that

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

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

$$1 \leq \left(\frac{2a}{d} + c \right)^{-1} \tag{D.18}$$

and this becomes exactly our condition (D.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

$$\left| L_{NT}(\beta) - \frac{1}{NT} \sum_{g=1}^{G} \sum_{\kappa_1=0}^{K} \cdots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \ldots \epsilon_{\kappa_g} L^{(g)} \left(\lambda^0, f^0, X_{\kappa_1}, X_{\kappa_2}, \ldots, X_{\kappa_g} \right) \right| \\ \leq \frac{(N-R)\gamma^{G+1} d_{\min}^2(\lambda^0, f^0)}{4(1-\gamma)}, \tag{D.19}$$

where

$$\gamma = \frac{\sum_{k=1}^{K} \left| \beta_k^0 - \beta_k \right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}}{r_0(\lambda^0, f^0)} < 1 . \tag{D.20}$$

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), which in our particular case turns out to be unnecessary.

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

$$\left\| S^{(l_1)} \mathcal{T}_{\kappa_1 \ldots}^{(\nu_1)} S^{(l_2)} \ldots S^{(l_p)} \mathcal{T}_{\ldots \kappa_g}^{(\nu_p)} S^{(l_{p+1})} \right\| \\ \leq \left(NT d_{\min}^2(\lambda^0, f^0) \right)^{-\sum l_j} \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0) \right)^{2p - \sum \nu_j} \|X_{\kappa_1}\| \|X_{\kappa_2}\| \ldots \|X_{\kappa_g}\| \tag{D.21}$$

We have

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

$$\sum_{\substack{l_1 + \ldots + l_{p+1} = p-1 \\ l_j \geq 0}} 1 \leq \sum_{\substack{l_1 + \ldots + l_{p+1} = p \\ l_j \geq 0}} 1 = \frac{(2p)!}{(p!)^2} \leq 4^p \tag{D.22}$$

Using this we find[27]

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

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 (D.21) and (D.22), we therefore find[28] for $g \geq 3$

$$
\begin{aligned}
&\frac{1}{NT} \left| L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \right| \\
&\quad \leq R\, d^2_{\min}(\lambda^0, f^0) \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0)\right)^{-g} \|X_{\kappa_1}\| \|X_{\kappa_2}\| \ldots \|X_{\kappa_g}\| \sum_{p=\lceil g/2 \rceil}^{g} \left(\frac{32\, d^2_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^p \\
&\quad \leq \frac{R\, g\, d^2_{\min}(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^g \frac{\|X_{\kappa_1}\|}{\sqrt{NT}} \frac{\|X_{\kappa_2}\|}{\sqrt{NT}} \cdots \frac{\|X_{\kappa_g}\|}{\sqrt{NT}}
\end{aligned} \tag{D.24}
$$

This implies for $g \geq 3$

$$
\begin{aligned}
&\frac{1}{NT} \left| \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \cdots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \epsilon_{\kappa_2} \ldots \epsilon_{\kappa_g} L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \right| \\
&\qquad\qquad \leq \frac{R\, g\, d^2_{\min}(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^g \left(\sum_{\kappa=0}^{K} \frac{\|\epsilon_\kappa X_\kappa\|}{\sqrt{NT}} \right)^g
\end{aligned} \tag{D.25}
$$

Therefore for $G \geq 2$ we have

$$
\begin{aligned}
&\left| L_{NT}(\beta) - \frac{1}{NT} \sum_{g=1}^{G} \sum_{\kappa_1=0}^{K} \cdots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \ldots \epsilon_{\kappa_g} L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \right| \\
&\qquad = \frac{1}{NT} \sum_{g=G+1}^{\infty} \sum_{\kappa_1=0}^{K} \sum_{\kappa_2=0}^{K} \cdots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \epsilon_{\kappa_2} \ldots \epsilon_{\kappa_g} L^{(g)}\left(\lambda^0,\, f^0,\, X_{\kappa_1},\, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) \\
&\qquad \leq \sum_{g=G+1}^{\infty} \frac{R\, g\, \alpha^g\, d^2_{\min}(\lambda^0, f^0)}{2} \\
&\qquad \leq \frac{R\,(G+1)\, \alpha^{G+1}\, d^2_{\min}(\lambda^0, f^0)}{2(1-\alpha)^2} \,,
\end{aligned} \tag{D.26}
$$

where

$$
\begin{aligned}
\alpha &= \frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \sum_{\kappa=0}^{K} \frac{\|\epsilon_\kappa X_\kappa\|}{\sqrt{NT}} \\
&= \frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \left(\sum_{k=1}^{K} \left|\beta^0_k - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} \right) < 1 \,,
\end{aligned} \tag{D.27}
$$

[27] 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 $\nu_j \leq 2$.

[28] 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 $NT d^2_{\min}(\lambda^0, f^0)$.

Using the same argument we can start from equation (D.23) to obtain the bound (D.16) for the remainder of the series expansion for $M_{\hat{\lambda}}(\beta)$.

Note that compared to the bound (D.19) on the remainder, the new bound (D.26) only shows convergence of the power series within the the smaller convergence radius $\frac{d^2_{\min}(\lambda^0, f^0)}{16\, d_{\max}(\lambda^0, f^0)} < r_0(\lambda^0, f^0)$. However, the factor $N - R$ does not appear in this new bound, which is crucial for our approximations.

∎

We can now proof the key theorem of the main text.

Proof of theorem 3.1. Assumption 1 implies that

$$d_{\max}(\lambda^0, f^0) \underset{p}{\longrightarrow} d^{\infty}_{\max} > 0 \;, \qquad\qquad d_{\min}(\lambda^0, f^0) \underset{p}{\longrightarrow} d^{\infty}_{\min} > 0 \;. \tag{D.28}$$

Therefore also $r_0(\lambda^0, f^0) \to_p r_0^{\infty} > 0$. Assumptions 1, 2, and 3 furthermore imply that

$$\begin{aligned} \frac{\|\lambda^0\|}{\sqrt{N}} &= \mathcal{O}_p(1) \;, & \frac{\|f^0\|}{\sqrt{T}} &= \mathcal{O}_p(1) \;, \\ \left\| \left(\frac{\lambda^{0\prime}\lambda^0}{N} \right)^{-1} \right\| &= \mathcal{O}_p(1) \;, & \left\| \left(\frac{f^{0\prime} f^0}{T} \right)^{-1} \right\| &= \mathcal{O}_p(1) \;, \\ \frac{\|X_k\|}{\sqrt{NT}} &= \mathcal{O}_p(1) \;, & \frac{\|e\|}{\sqrt{NT}} &= o_p(1) \;. \end{aligned} \tag{D.29}$$

For $\|\beta - \beta^0\| \leq \eta_{NT}$ we have $\|\beta - \beta^0\| = o_p(1)$ and also $|\beta_k - \beta^0_k| = o_p(1)$, $k = 1 \ldots K$. We thus find $\alpha = o_p(1)$, *i.e.* the condition to apply theorem D.2 part (iii) is asymptotically satisfied. Using the inequality (D.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

$$\frac{1}{NT} (\epsilon_0)^{g-r}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1}, \ldots,\, X_{k_r},\, X_0, \ldots,\, X_0\right) = \mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}} \right)^{g-r} \right) \;. \tag{D.30}$$

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

$$
\begin{aligned}
L_{NT}(\beta) =& \frac{1}{NT} \sum_{g=2}^{G_e} \sum_{\kappa_1=0}^{K} \ldots \sum_{\kappa_g=0}^{K} \epsilon_{\kappa_1} \ldots \epsilon_{\kappa_g} L^{(g)}\left(\lambda^0, f^0, X_{\kappa_1}, X_{\kappa_2}, \ldots, X_{\kappa_g}\right) + \mathcal{O}_p\left(\alpha^{G+1}\right) \\
=& \frac{1}{NT} \sum_{g=2}^{G_e} \epsilon_0^g L^{(g)}\left(\lambda^0, f^0, X_0, X_0, \ldots, X_0\right) \\
&+ \frac{1}{NT} \sum_{g=2}^{G_e} g \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k\right) \epsilon_0^{g-1} L^{(g)}\left(\lambda^0, f^0, X_k, X_0, \ldots, X_0\right) \\
&+ \frac{1}{NT} \sum_{g=2}^{G_e} g(g-1) \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \left(\beta_{k_1}^0 - \beta_{k_1}\right) \left(\beta_{k_2}^0 - \beta_{k_2}\right) \epsilon_0^{g-2} L^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, X_0, \ldots, X_0\right) \\
&+ \frac{1}{NT} \sum_{g=3}^{G_e} \sum_{r=3}^{g} \mathcal{O}_p\left[\left\|\beta^0 - \beta\right\|^r \epsilon_0^{g-r} L^{(g)}\left(\lambda^0, f^0, X_{k_1}, \ldots, X_{k_r}, X_0, \ldots, X_0\right)\right] \\
&+ \mathcal{O}_p\left[\left(\sum_{k=1}^{K} \left|\beta_k^0 - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}\right)^{G_e+1}\right] , \\
=& \frac{1}{NT} \sum_{g=2}^{G_e} g \sum_{k=1}^{K} \left(\beta_k^0 - \beta_k\right) L^{(g)}\left(\lambda^0, f^0, X_k, e, \ldots, e\right) \\
&+ \frac{2}{NT} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \left(\beta_{k_1}^0 - \beta_{k_1}\right) \left(\beta_{k_2}^0 - \beta_{k_2}\right) L^{(2)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}\right) \\
&+ \frac{1}{NT} I_{NT} + \frac{1}{NT} R_{NT}
\end{aligned}
\tag{D.31}
$$

where

$$
\begin{aligned}
I_{NT} =& \sum_{g=2}^{G_e} L^{(g)}\left(\lambda^0, f^0, e, e, \ldots, e\right) + NT\, \mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e+1}\right) , \\
R_{NT}(\beta) =& R_{1,NT}(\beta) + R_{2,NT}(\beta) + R_{3,NT}(\beta) , \\
R_{1,NT}(\beta) =& NT \sum_{g=3}^{G_e+1} \mathcal{O}_p\left(\|\beta^0 - \beta\|^2 \left(\frac{\|e\|}{\sqrt{NT}}\right)^{g-2}\right) , \\
R_{2,NT}(\beta) =& NT \sum_{g=3}^{G_e+1} \sum_{r=3}^{g} \mathcal{O}_p\left(\|\beta^0 - \beta\|^r \left(\frac{\|e\|}{\sqrt{NT}}\right)^{g-r}\right) , \\
R_{3,NT}(\beta) =& NT\, \mathcal{O}_p\left(\|\beta^0 - \beta\| \left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e}\right) .
\end{aligned}
\tag{D.32}
$$

We find that I_{NT} is independent of β, while $R_{1,NT}(\beta)$, $R_{2,NT}(\beta)$ and $R_{3,NT}(\beta)$ satisfy

$$\begin{aligned}
\sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{1,NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} &\leq \sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{1,NT}(\beta)|}{NT\,\left\|\beta-\beta^0\right\|^2} \\
&= \mathcal{O}_p\left(\frac{\|e\|}{\sqrt{NT}}\right) = o_p(1)\ , \\
\sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{2,NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} &\leq \sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{2,NT}(\beta)|}{NT\,\left\|\beta-\beta^0\right\|^2} \\
&= \mathcal{O}_p\left(\eta_{NT}\right) = o_p(1)\ , \\
\sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{3,NT}(\beta)|}{\left(1+\sqrt{NT}\,\left\|\beta-\beta^0\right\|\right)^2} &\leq \sup_{\beta:\|\beta-\beta^0\|\leq\eta_{NT}} \frac{|R_{3,NT}(\beta)|}{2\sqrt{NT}\,\left\|\beta-\beta^0\right\|} \\
&= \sqrt{NT}\,\mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}}\right)^{G_e}\right) = o_p(1)\ ,
\end{aligned} \tag{D.33}$$

with $\eta_{NT} \to 0$. In the last line we used assumption 3' to show that the term is $o_p(1)$. Since the condition (3.4) is satisfied for $R_{1,NT}(\beta)$, $R_{2,NT}(\beta)$ and $R_{3,NT}(\beta)$ separately, it is also satisfied for the total remainder $R_{NT}(\beta)$.

∎

Proof of theorem 3.3. The general expansion of $M_{\hat{\lambda}}$ is given in theorem D.2, and here we just make this expansion explicit up to a particular order. To obtain the bound on the remainder we make us of equation (D.23) in the proof of theorem D.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

$$\begin{aligned}
\hat{e} &= M_{\hat{\lambda}}\left(Y - \sum_{k=1} \hat{\beta}_k\, X_k\right) \\
&= M_{\hat{\lambda}}\left[e - \sum_{k=1}^{K}\left(\hat{\beta}_k - \beta_k^0\right) X_k + \lambda^0 f^{0\prime}\right]\ ,
\end{aligned} \tag{D.34}$$

and plugging in the expansion of $M_{\hat{\lambda}}$ gives the expansion of $\hat{e}$. ∎

Proof of Corollary 3.4. Having the expansion of the profile quasi likelihood function in theorem 3.1, and in particular the bounds on the remainder terms given there, we only have to verify that asymptotically the smallest eigenvalue of the $K \times K$ denominator matrix $W(\lambda^0,\, f^0,\, X_k)$ (which appears in the second order term in the likelihood expansion) is bounded from below. This guarantees that $W(\lambda^0,\, f^0,\, X_k)$ is invertible and that the norm of its inverse $\|W^{-1}(\lambda^0,\, f^0,\, X_k)\|$ is bounded from above as $N, T \to \infty$. Once this is verified, we can apply the results in *e.g.* Andrews (1999) to obtain the equation for $\hat{\beta}$ and the corollary is proven.

Remember that

$$W_{k_1 k_2}(\lambda^0,\, f^0,\, X_k) = \frac{1}{NT}\mathrm{Tr}(M_{f^0}\, X'_{k_1}\, M_{\lambda^0}\, X_{k_2})\ . \tag{D.35}$$

The smallest eigenvalue of the symmetric matrix $W(\lambda^0, f^0, X_k)$ is given my

$$
\begin{aligned}
\text{Eigval}_K \left[W(\lambda^0, f^0, X_k)\right] &= \min_{\{a \in \mathbb{R}^K,\, a \neq 0\}} \frac{a' \, W(\lambda^0, f^0, X_k) \, a}{\|a\|^2} \\
&= \min_{\{a \in \mathbb{R}^K,\, a \neq 0\}} \frac{1}{NT\,\|a\|^2} \text{Tr}\left[M_{f^0} \left(\sum_{k_1=1}^{K} a_{k_1} X'_{k_1} \right) M_{\lambda^0} \left(\sum_{k_2=1}^{K} a_{k_2} X_{k_2} \right) \right] \\
&= \min_{\substack{\{\alpha \in \mathbb{R}^{K_1},\ \mu \in \mathbb{R}^{K_2} \\ \alpha \neq 0,\ \mu \neq 0\}}} \frac{\text{Tr}\left[M_{f^0} \left(X'_{\text{low},\alpha} + X'_{\text{high},\mu} \right) M_{\lambda^0} \left(X_{\text{low},\alpha} + X_{\text{high},\mu} \right) \right]}{NT \, (\|\alpha\|^2 + \|\mu\|^2)} ,
\end{aligned}
\tag{D.36}
$$

where we decomposed $a' = (\alpha', \mu')$, with α and μ being vectors of length K_1 and K_2, respectively, and as in assumption 4 we defined corresponding linear combinations of high- and low-rank regressors[29]

$$
X_{\text{low},\alpha} = \sum_{k=1}^{K_1} \alpha_k \, X_k \ , \qquad\qquad X_{\text{high},\mu} = \sum_{k=K_1+1}^{K} \mu_k \, X_k \ . \tag{D.37}
$$

We have $M_{\lambda^0} = M_{(\lambda^0,\Xi)} + P_{(M_{\lambda^0}\Xi)}$, where $\Xi = (X_1, \ldots, X_{K_1})$ is the matrix that combines all low-rank regressors. Using this we obtain

$$
\begin{aligned}
&\text{Eigval}_K \left[W(\lambda^0, f^0, X_k)\right] \\
&= \min_{\substack{\{\alpha \in \mathbb{R}^{K_1},\ \mu \in \mathbb{R}^{K_2} \\ \alpha \neq 0,\ \mu \neq 0\}}} \frac{1}{NT \, (\|\alpha\|^2 + \|\mu\|^2)} \Big\{ \text{Tr}\left[M_{f^0} \left(X'_{\text{low},\alpha} + X'_{\text{high},\mu} \right) M_{(\lambda^0,\Xi)} \left(X_{\text{low},\alpha} + X_{\text{high},\mu} \right) \right] \\
&\qquad\qquad + \text{Tr}\left[M_{f^0} \left(X'_{\text{low},\alpha} + X'_{\text{high},\mu} \right) P_{(M_{\lambda^0}\Xi)} \left(X_{\text{low},\alpha} + X_{\text{high},\mu} \right) \right] \Big\} \\
&= \min_{\substack{\{\alpha \in \mathbb{R}^{K_1},\ \mu \in \mathbb{R}^{K_2} \\ \alpha \neq 0,\ \mu \neq 0\}}} \frac{1}{NT \, (\|\alpha\|^2 + \|\mu\|^2)} \Big\{ \text{Tr}\left[M_{f^0} \, X'_{\text{high},\mu} \, M_{(\lambda^0,\Xi)} \, X_{\text{high},\mu} \right] \\
&\qquad\qquad + \text{Tr}\left[M_{f^0} \left(X'_{\text{low},\alpha} + X'_{\text{high},\mu} \right) P_{(M_{\lambda^0}\Xi)} \left(X_{\text{low},\alpha} + X_{\text{high},\mu} \right) \right] \Big\} \\
&\geq \min_{\substack{\{\alpha \in \mathbb{R}^{K_1},\ \mu \in \mathbb{R}^{K_2} \\ \alpha \neq 0,\ \mu \neq 0\}}} \frac{1}{\|\alpha\|^2 + \|\mu\|^2} \left\{ c_1 \|\mu\|^2 + \max\left[0,\, c_2 \|\alpha\|^2 - c_3 \|\alpha\| \|\mu\| \right] \right\} \\
&\geq \min\left(\frac{c_2}{2}, \frac{c_1 c_2^2}{c_2^2 + c_3^2} \right)
\end{aligned}
\tag{D.38}
$$

where the inequalities hold in probability as $N, T \to \infty$, and c_1, c_2 and c_3 are appropriate positive constants (independent of N, T). Here we used that as $N, T \to \infty$ we have in probability that

$$
\begin{aligned}
\frac{1}{NT} \text{Tr}\left[M_{f^0} \, X'_{\text{high},\mu} \, M_{(\lambda^0,\Xi)} \, X_{\text{high},\mu} \right] &\geq c_1 \|\mu\|^2 \ , \\
\frac{1}{NT} \text{Tr}\left[M_{f^0} \left(X'_{\text{low},\alpha} + X'_{\text{high},\mu} \right) P_{(M_{\lambda^0}\Xi)} \left(X_{\text{low},\alpha} + X_{\text{high},\mu} \right) \right] &\geq 0 \ , \\
\frac{1}{NT} \text{Tr}\left[M_{f^0} \, X'_{\text{low},\alpha} \, P_{(M_{\lambda^0}\Xi)} \, X_{\text{low},\alpha} \right] &\geq c_2 \, \|\alpha\|^2 \ , \\
\frac{1}{NT} \text{Tr}\left[M_{f^0} \, X'_{\text{low},\alpha} \, P_{(M_{\lambda^0}\Xi)} \, X_{\text{high},\mu} \right] &\geq -\frac{c_3}{2} \, \|\alpha\| \|\mu\| \ , \\
\frac{1}{NT} \text{Tr}\left[M_{f^0} \, X'_{\text{high},\mu} \, P_{(M_{\lambda^0}\Xi)} \, X_{\text{high},\mu} \right] &\geq 0 \ ,
\end{aligned}
\tag{D.39}
$$

[29]As in assumption 4 the components of μ are denoted μ_{K_1+1} to μ_K to simplify notation.

which we want to justify now. The second and the last equation in (D.39) are true because *e.g.* $\operatorname{Tr}\left[M_{f^0}\, X'_{\text{high},\mu}\, P_{(M_{\lambda^0}\Xi)}\, X_{\text{high},\mu}\right] = \operatorname{Tr}\left[M_{f^0}\, X'_{\text{high},\mu}\, P_{(M_{\lambda^0}\Xi)}\, X_{\text{high},\mu}\, M_{f^0}\right]$, and the trace of a symmetric positive semi-definite matrix is non-negative. The first inequality in (D.39) is true because $\operatorname{rank}(f^0) + \operatorname{rank}(\lambda^0, \Xi) = 2R + K_1$ and using theorem B.3 and assumption 4(i) we have

$$\frac{1}{NT\|\mu\|^2}\operatorname{Tr}\left[M_{f^0}\, X'_{\text{high},\mu}\, M_{(\lambda^0,\Xi)}\, X_{\text{high},\mu}\right] \geq \frac{1}{NT\|\mu\|^2}\operatorname{Eigval}_{2R+K_1+1}\left[X_{\text{high},\mu}\, X'_{\text{high},\mu}\right] > 0\ , \quad \text{(D.40)}$$

in probability as $N, T \to \infty$, which justifies the existence of the constant c_1. The third inequality in (D.39) is true because according theorem B.2 (v) we have

$$\begin{aligned}
\frac{1}{NT}\operatorname{Tr}\left[M_{f^0}\, X'_{\text{low},\alpha}\, P_{(M_{\lambda^0}\Xi)}\, X_{\text{high},\mu}\right] &\geq -\frac{K_1}{NT}\, \|X_{\text{low},\alpha}\|\, \|X_{\text{high},\mu}\| \\
&\geq -\frac{K_1}{NT}\, \|X_{\text{low},\alpha}\|_F\, \|X_{\text{high},\mu}\|_F \\
&\geq -K_1\, K_1\, K_2\, \|\alpha\|\, \|\mu\| \max_{k_1=1\ldots K_1}\left\|\frac{X_{k_1}}{\sqrt{NT}}\right\|_F \max_{k_2=K_1+1\ldots K}\left\|\frac{X_{k_2}}{\sqrt{NT}}\right\|_F \\
&\geq -\frac{c_3}{2}\, \|\alpha\|\, \|\mu\|\ , \qquad \text{(D.41)}
\end{aligned}$$

where we used that assumption 2 (iii) implies that $\left\|X_k/\sqrt{NT}\right\|_F < C$ holds in probability for some constant C as $N, T \to \infty$ (which is stronger than $\left\|X_k/\sqrt{NT}\right\|_F = O_p(1)$), and we set $c_3 = K_1\, K_1\, K_2\, C^2$. Finally, we have to argue that the third inequality in (D.39) holds. Note that $X'_{\text{low},\alpha}\, P_{(M_{\lambda^0}\Xi)}\, X_{\text{low},\alpha} = X'_{\text{low},\alpha}\, M_{\lambda^0}\, X_{\text{low},\alpha}$, *i.e.* we need to show that

$$\frac{1}{NT}\operatorname{Tr}\left[M_{f^0}\, X'_{\text{low},\alpha}\, M_{\lambda^0}\, X_{\text{low},\alpha}\right] \geq c_2\, \|\alpha\|^2\ . \qquad \text{(D.42)}$$

Using part (vi) or theorem B.2 we find

$$\begin{aligned}
\frac{1}{NT}\operatorname{Tr}\left[M_{f^0}\, X'_{\text{low},\alpha}\, M_{\lambda^0}\, X_{\text{low},\alpha}\right] &= \frac{1}{NT}\operatorname{Tr}\left[M_{\lambda^0}\, X_{\text{low},\alpha}\, M_{f^0}\, X'_{\text{low},\alpha}\, M_{\lambda^0}\right] \\
&\geq \frac{1}{NT}\left\|M_{\lambda^0}\, X_{\text{low},\alpha}\, M_{f^0}\, X'_{\text{low},\alpha}\, M_{\lambda^0}\right\|\ , \qquad \text{(D.43)}
\end{aligned}$$

and according to lemma C.3 this expression is bounded by some positive constant times $\|\alpha\|^2$ (in the lemma we have $\|\alpha\| = 1$, but all expressions are homogeneous in $\|\alpha\|$). This concludes the proof. ∎

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.

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

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

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

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, 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.

Bhatia, R. (1997). *Matrix Analysis*. Springer-Verlag, New York.

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

Galantai, A. and Hegedus, C. J. (2006). Jordan's principal angles in complex vector spaces. *Num. Lin. Alg. Appl.*, 13:589–598.

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.

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).

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.

Onatski, A. (2005). Determining the number of factors from empirical distribution of eigenvalues. Technical report.

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.

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.