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

Hyungsik Roger Moon[‡] Martin Weidner[§]

November 13, 2012

Abstract

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

Keywords: Panel data, interactive fixed effects, factor models, likelihood expansion, quasi-MLE, perturbation theory of linear operators, random matrix theory.

JEL-Classification: C23, C33

1 Introduction

Panel data models typically incorporate individual and time effects to control for heterogeneity in cross-section and across time-periods. While often these individual and time effects enter the model additively, they can also be interacted multiplicatively, thus giving rise to so called *interactive effects*, which we also refer to as a *factor structure*. The multiplicative form captures the heterogeneity in the data more flexibly, since it allows for common time-varying shocks (factors) to affect the cross-sectional units with individual specific sensitivities (factor loadings). It is this flexibility that motivated the discussion of interactive effects in the econometrics literature, e.g. Holtz-Eakin, Newey and Rosen (1988),

*We thank the participants of the Cowles Summer Conference "Handling Dependence: Temporal, Cross-sectional, and Spatial" at Yale University for interesting comments. Moon acknowledges the NSF for financial support via SES 0920903.

[‡]Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089-0253. Email: moonr@usc.edu.

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

Ahn, Lee, Schmidt (2001; 2007), Pesaran (2006), Bai (2009b; 2009a), Zaffaroni (2009), Moon and Weidner (2010).

Analogous to the analysis of individual specific effects, one can either choose to model the interactive effects as random (random effects/correlated effects) or as fixed (fixed effects), with each option having its specific merits and drawbacks, that have to be weighed in each empirical application separately. In this paper, we consider the *interactive fixed effect* specification, i.e. we treat the interactive effects as nuisance parameters, which are estimated jointly with the parameters of interest.[1] The advantages of the fixed effects approach are for instance that it is semi-parametric, since no assumption on the distribution of the interactive effects needs to be made, and that the regressors can be arbitrarily correlated with the interactive effect parameters.

Let R^0 be the true number of interactive effects (number of factors) in the data, and let R be the number of interactive effects used by the econometrician in the data analysis. A key restriction in the existing literature on interactive fixed effects is that R^0 is assumed to be known,[2] i.e. $R = R^0$. This is true both for the quasi-differencing analysis in Holtz-Eakin, Newey and Rosen (1988)[3] and for the least squares analysis of Bai (2009b). Assuming R^0 to be known could be quite restrictive, since in many empirical applications there is no consensus about the exact number of factors in the data or in the relevant economic model, so that an estimator which is not robust towards some degree of misspecification of R^0 should not be used. The goal of the present paper is to overcome this problem.

For a linear panel regression model with interactive fixed effects we consider the Gaussian quasi maximum likelihood estimator (QMLE),[4] which jointly minimized the sum of squared residuals over the regression parameters and the interactive fixed effects parameters (see Kiefer (1980), Bai (2009b), and Moon and Weinder (2010)). We employ an asymptotic where both the number of cross-sectional and the number of time-serial dimensions becomes large, while the number of interactive effects R^0 (and also R) is constant.

The main finding of the paper is that under appropriate assumptions the QMLE of the regression parameters has the *same limiting distribution* for all $R \geq R^0$. Thus, the QMLE is robust towards inclusion of *extra* interactive effects in the model, and within the QMLE framework there is *no asymptotic efficiency loss* from choosing R larger than R^0. This result is surprising because the conjecture in the literature is that the QMLE with $R > R^0$ might be consistent but could be less efficient than the QMLE with R^0 (e.g., see Bai (2009b)).[5]

The important empirical implication of our result is that as long as a valid upper bound on the number of factors is known one can use this upper bound to construct the QMLE, and need not worry about consistent estimation of the number of factors. Since

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

[2]In the literature, consistent estimation procedures for R^0 are established only for pure factor models, not for the model with regressors.

[3]Holtz-Eakin, Newey and Rosen (1988) assume just one interactive effect, but their approach could easily be generalized to multiple interactive effects, as long as their number is known

[4]The QMLE is sometimes called concentrated least squares estimator in the literature.

[5]For $R < R^0$ the QMLE could be inconsistent, since then there are interactive fixed effects in the residuals of the model which can be correlated with the regressors but are not controlled for in the estimation.

the limiting distribution of the QMLE with $R > R^0$ is identical to the one with $R = R^0$ the results of Bai (2009b) and Moon and Weidner (2010) regarding inference on the regression parameters become applicable.

In order to derive the asymptotic theory of the QMLE with $R \geq R^0$ we study the properties of the profile likelihood function, which is the quasi likelihood function after integrating out the interactive fixed effect parameters. Concretely, we derive an approximate quadratic expansion of this profile likelihood in the regression parameters. This expansion is difficult to perform, since integrating out the interactive fixed effects results in an eigenvalue problem in the formulation of the profile likelihood. For $R = R^0$ we show how to overcome this difficulty by performing a joint expansion of the profile likelihood in the regression parameters and in the idiosyncratic error terms. Using the perturbation theory of linear operators we prove that the profile quasi likelihood function is analytic in a neighborhood of the true parameter, and we obtain explicit formulas for the expansion coefficients, in particular analytic expressions for the approximated score and the approximated Hessian for $R = R^0$.[6]

To generalize the result to $R > R^0$ we then show that the difference between the profile likelihood for $R = R^0$ and for $R > R^0$ is just a constant term plus a term whose dependence on the regression parameters is sufficiently small to be irrelevant for the asymptotic distribution of the QMLE. Due to the eigenvalue problem in the likelihood function, the derivation of this last result requires some very specific knowledge about the eigenvectors and eigenvalues of the random covariance matrix of the idiosyncratic error matrix. We provide high-level assumptions under which the results hold, and we show that these high-level assumptions are satisfied, when the idiosyncratic errors of the model are independent and identically normally distributed. As we explain in section 4, the justification of our high-level assumptions for more general distribution of the idiosyncratic errors requires some further progress in the Random Matrix Theory of real random covariance matrices, both regarding the properties of their eigenvalues and of their eigenvectors (see Bai (1999) for a review of this literature).

The paper is organized as follows. In Section 2 we introduce the interactive fixed effect model, its Gaussian quasi likelihood function, and the corresponding QMLE, and also discuss consistency of the QMLE. The asymptotic profile likelihood expansion is derived in Section 3. Section 4 provides a justification for the high-level assumptions that we impose, and discusses the relation of these assumptions to the random matrix theory literature. Monte Carlo results which illustrate the validity of our conclusion at finite sample are presented in Section 5, and the conclusions of the paper are drawn in Section 6.

A few words on notation. The transpose of a matrix A is denoted by A'. For a column vectors v its Euclidean norm is defined by $\|v\| = \sqrt{v'v}$. For the n-th largest eigenvalues (counting multiple eigenvalues multiple times) of a symmetric matrix B we write $\mu_n(B)$. For an $m \times n$ matrix A the Frobenius or Hilbert Schmidt norm is $\|A\|_{HS} = \sqrt{\mathrm{Tr}(AA')}$, and the operator or spectral norm is $\|A\| = \max_{0 \neq v \in \mathbb{R}^n} \frac{\|Av\|}{\|v\|}$, or equivalently $\|A\| = \sqrt{\mu_1(A'A)}$. Furthermore, we use $P_A = A(A'A)^{-1}A'$ and $M_A = \mathbb{1} - A(A'A)^{-1}A'$, where $\mathbb{1}$ is the $m \times m$ identity matrix, and $(A'A)^{-1}$ denotes some generalized inverse if A is not of full column rank. For square matrices B, C, we use $B > C$ (or $B \geq C$) to indicate that $B - C$ is positive (semi) definite. We use "wpa1" for "with probability approaching one", and $A =_d B$ to

[6] The likelihood expansion for $R = R^0$ was first presented in Moon and Weidner (2009). We separate and extend the expansion result from the 2009 working paper and present it in this paper. The remaining application results of Moon and Weidner (2009) are now in Moon and Weidner (2010).

indicate that the random variables A and B have the same probability distribution.

2 Model, QMLE and Consistency

A linear panel regression model with cross-sectional dimension N, time-serial dimension T, and interactive fixed effects of dimension R^0, is given by

$$Y = \sum_{k=1}^{K} \beta_k^0 \, X_k \, + \varepsilon \, , \qquad\qquad \varepsilon = \lambda^0 \, f^{0\,\prime} + e \, , \tag{2.1}$$

where Y, X_k, ε and e are $N \times T$ matrices, λ^0 is a $N \times R^0$ matrix, f^0 is a $T \times R^0$ matrix, and the regression parameters β_k^0 are scalars — the superscript zero indicates the true value of the parameters. We write β for the K-vector of regression parameters, and introduce the notation $\beta \cdot X \equiv \sum_{k=1}^{K} \beta_k \, X_k$. All matrices, vectors and scalars in this paper are real valued. A choice for the number of interactive effects R used in the estimation needs to be made, and we may have $R \neq R^0$ since the true number of factors R^0 may not be known accurately. Given the choice R, the quasi maximum likelihood estimator (QMLE) for the parameters β^0, λ^0 and f^0 is given by[7]

$$\left(\hat{\beta}_R, \, \hat{\Lambda}_R, \, \hat{F}_R\right) = \underset{\{\beta \in \mathbb{R}^K, \, \Lambda \in \mathbb{R}^{N \times R}, \, F \in \mathbb{R}^{T \times R}\}}{\operatorname{argmin}} \left\| Y \, - \, \beta \cdot X \, - \, \Lambda \, F' \right\|_{HS}^2 \, . \tag{2.2}$$

The square of the Hilbert-Schmidt norm is simply the sum of the squared elements of the argument matrix, i.e. the QMLE is defined by minimizing the sum of squared residuals, which is equivalent to minimizing the likelihood function for *iid* normal idiosyncratic errors. The estimator is the *quasi* MLE since the idiosyncratic errors need not be *iid* normal and since R might not equal R^0. The QMLE for β^0 can equivalently be defined by minimizing the profile quasi likelihood function, namely

$$\hat{\beta}_R = \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \, \mathcal{L}_{NT}^R(\beta) \, , \tag{2.3}$$

where

$$\begin{aligned} \mathcal{L}_{NT}^R(\beta) &= \min_{\{\Lambda \in \mathbb{R}^{N \times R}, \, F \in \mathbb{R}^{T \times R}\}} \frac{1}{NT} \left\| Y \, - \, \beta \cdot X \, - \, \Lambda \, F' \right\|_{HS}^2 \\ &= \min_{F \in \mathbb{R}^{T \times R}} \frac{1}{NT} \operatorname{Tr} \left[(Y - \beta \cdot X) \, M_F \, (Y - \beta \cdot X)' \right] \\ &= \frac{1}{NT} \sum_{t=R+1}^{T} \mu_t \left[(Y - \beta \cdot X)' \, (Y - \beta \cdot X) \right] \, . \end{aligned} \tag{2.4}$$

Here, we first concentrated out Λ by use of its own first order condition. The resulting optimization problem for F is a principal components problem, so that the the optimal F is

[7] The optimal $\hat{\Lambda}_R$ and $\hat{F}_R$ in (2.2) are not unique, since the objective function is invariant under right-multiplication of Λ with a non-degenerate $R \times R$ matrix S, and simultaneous right-multiplication of F with $(S^{-1})'$. However, the column spaces of $\hat{\Lambda}_R$ and $\hat{F}_R$ are uniquely determined.

given by the R largest principal components of the $T \times T$ matrix $(Y - \beta \cdot X)' (Y - \beta \cdot X)$. At the optimum the projector M_F therefore exactly projects out the R largest eigenvalues of this matrix, which gives rise to the final formulation of the profile likelihood function as the sum over its $T - R$ smallest eigenvalues.[8]

This last formulation of $\mathcal{L}^R_{NT}(\beta)$ is very convenient since it does not involve any explicit optimization over nuisance parameters. Numerical calculation of eigenvalues is very fast, so that the numerical evaluation of $\mathcal{L}^R_{NT}(\beta)$ is unproblematic for moderately large values of T. The function $\mathcal{L}^R_{NT}(\beta)$ is not convex in β and might have multiple local minima, which have to be accounted for in the numerical calculation of $\hat{\beta}_R$. We write $\mathcal{L}^0_{NT}(\beta)$ for $\mathcal{L}^{R^0}_{NT}(\beta)$, which is the profile likelihood obtain from the true number of factors. In order to show consistency of $\hat{\beta}_R$ we impose the following assumptions.

Assumption 1.

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

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

One can justify Assumption 1(i) by use of the norm inequality $\|X_k\| \leq \|X_k\|_{HS}$ and the fact that $\|X_k\|^2_{HS} = \sum_{i,t} X^2_{k,it} = \mathcal{O}_p(NT)$, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and the last step follows e.g. if $X_{k,it}$ has a uniformly bounded second moment. Assumption 1(ii) is a condition on the largest eigenvalue of the random covariance matrix $e'e$, which is often studied in the literature on random matrix theory, e.g. Geman (1980), Bai, Silverstein, Yin (1988), Yin, Bai, and Krishnaiah (1988), Silverstein (1989). The results in Latala (2005) show that $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$ if e has independent entries with mean zero and uniformly bounded fourth moment. Some weak dependence of the entries e_{it} across i and t is also permissible (see, e.g., Moon and Weidner (2010)).

Assumption 2.

(i) $\frac{1}{\sqrt{NT}} \mathrm{Tr}(X_k e') = \mathcal{O}_p(1), \quad k = 1, \ldots, K.$

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

$$\min_{\{\alpha \in \mathbb{R}^K,\, \|\alpha\|=1\}} \sum_{t=R+R^0+1}^{T} \mu_t \left(\frac{X'_\alpha X_\alpha}{NT} \right) \geq b\,, \qquad \textit{wpa1.}$$

Assumption 2(i) requires weak exogeneity of the regressors X_k. Assumption 2(ii) is a generalization of the usual non-collinearity condition on the regressors. It requires $X'_\alpha X_\alpha$ to be non-degenerate even after elimination of the largest $R + R^0$ eigenvalues (the sum in the assumption only runs over the smallest $T - R - R^0$ eigenvalues of this matrix, while running over all eigenvalues would give the trace operator, and thus the usual non-colinearity condition). In particular, this assumption is violated if there exists a linear combination of the regressors with $\|\alpha\| = 1$ and $\mathrm{rank}(X_\alpha) \leq R + R^0$, i.e. the assumption rules out "low-rank regressors" like time invariant regressors or cross-sectionally invariant

[8] Since the model is symmetric under $N \leftrightarrow T$, $\Lambda \leftrightarrow F$, $Y \leftrightarrow Y'$, $X_k \leftrightarrow X'_k$ there also exists a dual formulation of $\mathcal{L}^R_{NT}(\beta)$ that involves solving an eigenvalue problem for an $N \times N$ matrix.

regressors. These low-rank regressors require a special treatment in the interactive fixed effect model (see Bai (2009b) and Moon and Weidner (2010)), and we ignore them in the present paper. If one is not interested explicitly in their regression coefficients, one can always eliminate the low-rank regressors by an appropriate projection of the data, e.g. subtraction of the time (or cross-sectional) means from the data eliminates all time-invariant (or cross-sectionally invariant) regressors.

Theorem 2.1. *Let Assumption 1 and 2 be satisfied and let $R \geq R^0$. For $N, T \to \infty$ we then have* $\sqrt{\min(N,T)}\left(\hat{\beta}_R - \beta^0\right) = \mathcal{O}_p(1)$.

Remarks.

(i) The Theorem guarantees consistency of $\hat{\beta}_R$, $R \geq R^0$, in an arbitrary limit $N, T \to \infty$. In the rest of this paper we consider asymptotics where N and T grow at the same rate, i.e. $N/T \to \kappa^2$, for some positive constant κ. For these restricted asymptotics the theorem already guarantees $\sqrt{N}$ (or equivalently $\sqrt{T}$) consistency of $\hat{\beta}_R$, which is a useful intermediate result.

(ii) The $\sqrt{\min(N,T)}$ convergence rate in Theorem 2.1 can be generalized further. If we generalize Assumption 1(ii) and Assumption 2(i) to Assumption 1(ii^*) $\frac{1}{\sqrt{NT}}\|e\| = \mathcal{O}_p(\xi_{NT})$, and Assumption 2$(i^*)$ $\frac{1}{NT}\mathrm{Tr}(X_k e') = \mathcal{O}_p(\xi_{NT})$, $k = 1, \ldots, K$, where $\xi_{NT} \to 0$, then it is possible to establish that $\sqrt{\xi_{NT}}\big(\hat{\beta}_R - \beta^0\big) = \mathcal{O}_p(1)$.

The proof of Theorem 2.1 is presented in the appendix. The theorem imposes no restriction at all on f^0 and λ^0, apart from the condition $R \geq R^0$. To derive the results in the rest of the paper we do however make the following *strong factor* assumption.[9]

Assumption 3.

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

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

The main result of this paper is that the inclusion of unnecessary factors in the estimation does not change the asymptotic distribution of the QMLE for β^0. Before deriving this result rigorously, we want to provide an intuitive explanation for it. As already mentioned above, the estimator $\hat{F}_R$ is given by the first R principal components of the matrix $(Y - \hat{\beta}_R \cdot X)'(Y - \hat{\beta}_R \cdot X)$. We have

$$Y - \hat{\beta}_R \cdot X = \lambda^0 f^{0\prime} + e - (\hat{\beta}_R - \beta^0) \cdot X. \tag{2.5}$$

For asymptotics, where N and T grow at the same rate, we find that Assumption 1 and the result of Theorem 2.1 guarantee that $\|e - (\hat{\beta}_R - \beta^0) \cdot X\| = \mathcal{O}_p(\sqrt{N})$. The strong factor assumption implies that the norms of the columns of λ^0 and f^0 each grow at a rate of $\sqrt{N}$ (or equivalently $\sqrt{T}$), so that the spectral norm of $\lambda^0 f^{0\prime}$ grows at the rate $\sqrt{NT}$. The strong factor assumption therefore guarantees that $\lambda^0 f^{0\prime}$ is the dominant component

[9] The strong factor assumption is regularly imposed in the literature on large N and T factor models, including Bai and Ng (2002), Stock and Watson (2002) and Bai (2009b). Onatski (2006) discussed an alternative "weak factor" assumption for the purpose of estimating the number of factors in a pure factor model, and a more general discussion of strong and weak factors is given in Chudik, Pesaran and Tosetti ().

of $Y - \hat{\beta}_R \cdot X$, which implies that the first R^0 principal components of $(Y - \hat{\beta}_R \cdot X)'(Y - \hat{\beta}_R \cdot X)$ are close to f^0, i.e. the true factors are correctly picked up by the principal component estimator. The additional $R - R^0$ principal components that are estimated for $R > R^0$ cannot pick up anymore true factors and are thus mostly determined by the remaining term $e - (\hat{\beta}_R - \beta^0) \cdot X$. Our results below show that $\hat{\beta}_R$ is not only $\sqrt{N}$ consistent, but actually $\sqrt{NT}$ consistent, so that $\|(\hat{\beta}_R - \beta^0) \cdot X\| = \mathcal{O}_p(1)$, which makes the idiosyncratic error matrix e the dominant part of $e - (\hat{\beta}_R - \beta^0) \cdot X$, i.e. the $R - R^0$ additional principal components in $\hat{F}_R$ are mostly determined by e, and more precisely are close to the $R - R^0$ principal components of $e'e$. This means that they are essentially random and close to uncorrelated with the regressors X_k. Including unnecessary factors in the QMLE calculation is therefore analogous to including irrelevant regressors in a linear regression which are uncorrelated with the relevant regressors X_k. From the second line in equation (2.4) we see that these additional random components of $\hat{F}_R$ project out the corresponding $R - R^0$ dimensional subspace of the T-dimensional space spanned by the observations over time, thus effectively reducing the number of time dimensions by $R - R^0$. This usually results in a somewhat increased finite sample variance of the QMLE, but has no influence asymptotically as T goes to infinity, so that the asymptotic distributions of $\hat{\beta}_{R^0}$ and $\hat{\beta}_R$ are identical for $R \geq R^0$.

3 Asymptotic Profile Likelihood Expansion

To derive the asymptotics of $\hat{\beta}_R$, we study the asymptotic properties of the profile likelihood function $\mathcal{L}^R_{NT}(\beta)$ around β^0. First we notice that the expression cannot easily be discussed by analytic means, since there is no explicit formula for the eigenvalues of a matrix. In particular, a standard Taylor expansion of $\mathcal{L}^0_{NT}(\beta)$ around β^0 cannot easily be derived. In Section 3.1 we show how to overcome this problem when the true number of factors is known, i.e. $R = R^0$, and in Section 3.2 we generalize the results to $R > R^0$.

When the true R^0 is known, the approach we choose is to perform a joint expansion in the regression parameters and in the idiosyncratic error terms. To perform this joint expansion we apply the perturbation theory of linear operators (e.g., Kato (1980)). We thereby obtain an approximate quadratic expansion of $\mathcal{L}^0_{NT}(\beta)$ in β, which can be used to derive the first order asymptotic theory of the QMLE $\hat{\beta}_{R^0}$.

To carry the results for $R = R^0$ over to $R > R^0$, we first note that equation (2.4) implies that

$$\mathcal{L}^R_{NT}(\beta) = \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{t=R^0+1}^{R} \mu_t \left[(Y - \beta \cdot X)' (Y - \beta \cdot X) \right]. \tag{3.1}$$

The extra term $\frac{1}{NT} \sum_{t=R^0+1}^{R} \mu_t \left[(Y - \beta \cdot X)' (Y - \beta \cdot X) \right]$ is due to overfitting on the extra factors. We show that the β-dependence of this term is sufficiently small, so that apart from a constant the approximate quadratic expansions of $\mathcal{L}^R_{NT}(\beta)$ and $\mathcal{L}^0_{NT}(\beta)$ around β^0 are identical. To obtain this result we first strengthen Theorem 2.1 and show that $\hat{\beta}_R$ converges to β^0 at a rate of at least $N^{3/4}$, so that we only have to discuss the β-dependence of the extra term in $\mathcal{L}^R_{NT}(\beta)$ within an $N^{3/4}$ shrinking neighborhood of β^0.

From the analysis of $\mathcal{L}^R_{NT}(\beta)$, we can then deduce the main result of the paper, namely

$$\sqrt{NT}\left(\hat{\beta}_R-\beta^0\right)=\sqrt{NT}\left(\hat{\beta}_{R^0}-\beta^0\right)+o_p(1). \tag{3.2}$$

This implies that the limiting distributions of $\hat{\beta}_R$ and $\hat{\beta}_{R^0}$ are identical, and that over-estimating the number of factors results in no efficiency loss in terms of the asymptotic variance of the QMLE.

3.1 When $R = R^0$

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

$$\frac{1}{\sqrt{NT}}\left(Y-\beta\cdot X\right) \;=\; \frac{1}{\sqrt{NT}}\left[\lambda^0 f^{0\prime}+(\beta^0-\beta)\cdot X+e\right] \;=\; \frac{\lambda^0 f^{0\prime}}{\sqrt{NT}}+\sum_{k=0}^{K}\epsilon_k\,\frac{X_k}{\sqrt{NT}}\,. \tag{3.3}$$

According to equation (2.4) the profile likelihood $\mathcal{L}^0_{NT}(\beta)$ can be written as the sum over the $T-R^0$ smallest eigenvalues of the matrix in (3.3) multiplied by its transposed. We consider $\sum_{k=0}^{K}\epsilon_k\,X_k/\sqrt{NT}$ as a small perturbation of the unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$, and thus expand $\mathcal{L}^0_{NT}(\beta)$ in the perturbation parameters $\epsilon = (\epsilon_0,\ldots,\epsilon_K)$ around $\epsilon = 0$, namely

$$\mathcal{L}^0_{NT}(\beta)=\frac{1}{NT}\sum_{g=0}^{\infty}\sum_{k_1,\ldots,k_g=0}^{K}\epsilon_{k_1}\,\epsilon_{k_2}\,\ldots\,\epsilon_{k_g}\,L^{(g)}\left(\lambda^0,\,f^0,\,X_{k_1},\,X_{k_2},\ldots,X_{k_g}\right)\,, \tag{3.4}$$

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

The unperturbed matrix $\lambda^0 f^{0\prime}/\sqrt{NT}$ has rank R^0, so that the $T-R^0$ smallest eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ are all zero, i.e. $\mathcal{L}^0_{NT}(\beta)=0$ for $\epsilon = 0$ and thus $L^{(0)}\left(\lambda^0,\,f^0\right)=0$. Due to Assumption 3 the R^0 non-zero eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ converge to positive constants as $N,T\to\infty$. This means that the "separating distance" of the $T-R^0$ zero-eigenvalues of the unperturbed $T\times T$ matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$ converges to a positive constant, i.e. the next largest eigenvalue is well separated. This is exactly the technical condition under which the perturbation theory of linear operators guarantees that the above expansion of $\mathcal{L}^0_{NT}$ in ϵ exists and is convergent as long as the spectral norm of the perturbation $\sum_{k=0}^{K}\epsilon_k\,X_k/\sqrt{NT}$ is smaller than a particular convergence radius $r_0(\lambda^0,f^0)$, which is closely related to the separating distance of the zero-eigenvalues. For details on that see Kato (1980) and Appendix A.2, where we define $r_0(\lambda^0,f^0)$ and show that it converges to a positive constant as $N,T\to\infty$. Note that for the expansion (3.4) it is crucial that we have $R = R^0$, since the perturbation theory of linear operators describes the perturbation of the sum of *all* zero-eigenvalues of the unperturbed matrix $f^0\lambda^{0\prime}\lambda^0 f^{0\prime}/NT$. For $R > R^0$ the sum in $\mathcal{L}^R_{NT}(\beta)$ leaves out the $R-R^0$ largest of these perturbed zero-eigenvalues, which results in a much more complicated mathematical problem, since the structure and ranking among these perturbed zero-eigenvalues needs to be discussed.

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

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

$$\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right)\right| \;\leq\; a_{NT}\,(b_{NT})^g\,\frac{\|X_{k_1}\|}{\sqrt{NT}}\,\frac{\|X_{k_2}\|}{\sqrt{NT}}\,\ldots\,\frac{\|X_{k_g}\|}{\sqrt{NT}}\,, \qquad (3.5)$$

where a_{NT} and b_{NT} are functions of λ^0 and f^0 that converge to finite positive constants in probability. This bound on the coefficients $L^{(g)}$ allows us to derive a bound on the remainder term, when the profile likelihood expansion is truncated at a particular order. The likelihood expansion can be applied under more general asymptotics, but here we only consider the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, i.e. N and T grow at the same rate. Then, the relevant coefficients of the expansion, which are not treated as part of the remainder term, are

$$\begin{aligned}
\mathcal{L}^0_{NT}(\beta^0) &= \frac{1}{NT}\sum_{g=2}^{\infty} \epsilon_0^g L^{(g)}\left(\lambda^0, f^0, X_0, X_0, \ldots, X_0\right) = \frac{1}{NT}\sum_{g=2}^{\infty} L^{(g)}\left(\lambda^0, f^0, e, e, \ldots, e\right)\,, \\
W_{k_1 k_2} &= \frac{1}{NT} L^{(2)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}\right) = \frac{1}{NT}\,\mathrm{Tr}(M_{\lambda^0}\, X_{k_1}\, M_{f^0}\, X'_{k_2})\,, \\
C^{(1)}_k &= \frac{1}{\sqrt{NT}} L^{(2)}\left(\lambda^0,\, f^0,\, X_k,\, U\right) = \frac{1}{\sqrt{NT}}\,\mathrm{Tr}(M_{\lambda^0}\, X_k\, M_{f^0}\, e')\,, \\
C^{(2)}_k &= \frac{3}{2\sqrt{NT}} L^{(3)}\left(\lambda^0,\, f^0,\, X_k,\, e,\, U\right) \\
&= -\frac{1}{\sqrt{NT}}\Big[\mathrm{Tr}\left(e M_{f^0}\, e'\, M_{\lambda^0}\, X_k\, f^0\,(f^{0\prime} f^0)^{-1}\,(\lambda^{0\prime}\lambda^0)^{-1}\,\lambda^{0\prime}\right) \\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, e\, M_{f^0}\, X'_k\, \lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime} f^0)^{-1}\, f^{0\prime}\right) \\
&\qquad + \mathrm{Tr}\left(e' M_{\lambda^0}\, X_k\, M_{f^0}\, e'\, \lambda^0\,(\lambda^{0\prime}\lambda^0)^{-1}\,(f^{0\prime} f^0)^{-1}\, f^{0\prime}\right)\Big]\,.
\end{aligned} \qquad (3.6)$$

In the first line above we used the fact that $L^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, \ldots, X_{k_g}\right)$ is linear in the arguments X_{k_1} to X_{k_g} and that $\epsilon_0 X_0 = e$. The $K \times K$ matrix W with elements $W_{k_1 k_2}$ is the approximated Hessian of the profile likelihood function $\mathcal{L}^0_{NT}(\beta)$. The K-vectors $C^{(1)}$ and $C^{(2)}$ with elements $C^{(1)}_k$ and $C^{(2)}_k$ constitute the approximated score of $\mathcal{L}^0_{NT}(\beta)$. From the expansion (3.4) and the bound (3.5) we obtain the following theorem, whose proof is provided in the appendix.

Theorem 3.1. *Let Assumptions 1 and 3 be satisfied. Suppose that $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Then we have*

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

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

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

Corollary 3.2. *Let Assumptions 1, 2, and 3 be satisfied. Furthermore assume that $C^{(1)} = \mathcal{O}_p(1)$. In the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have*

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

Since the estimator $\hat{\beta}_{R^0}$ minimizes $\mathcal{L}^0_{NT}(\beta)$ it must in particular satisfy $\mathcal{L}^0_{NT}(\hat{\beta}_{R^0}) \le \mathcal{L}^0_{NT}\big(\beta^0 + W^{-1}\left(C^{(1)} + C^{(2)}\right)/\sqrt{NT}\big)$. The Corollary follows from applying Theorem 3.1 to this inequality and using the consistency of $\hat{\beta}_{R^0}$. Details are given in the appendix. Using Theorem 3.1, the corollary is also directly obtained from the results in Andrews (1999). Our assumptions already guarantee $C^{(2)} = \mathcal{O}_p(1)$ and $W^{-1} = \mathcal{O}_p(1)$, so that only $C^{(1)} = \mathcal{O}_p(1)$ needs to be assumed explicitly in the Corollary.

Corollary 3.2 allows to replicate the result in Bai (2009b). Furthermore, the assumptions in the corollary do not restrict the regressor to be strictly exogenous, and the techniques developed here are applied in Moon and Weidner (2009) to discuss pre-determined regressors in the linear factor regression model with $R = R^0$, in which case the score term $C^{(1)}$ contributes an additional incidental parameter bias to the asymptotic distribution of $\hat{\beta}_R$.

Remark. If we weaken Assumption 1(ii) to $\|e\| = o_p(N^{2/3})$, then Theorem 3.1 still continues to hold. If we assume that $C^{(2)} = \mathcal{O}_p(1)$, then Corollary 3.2 also holds under this weaker condition on $\|e\|$.

3.2 When $R > R^0$

We now extend the likelihood expansion to the case $R > R^0$. Let $\hat{\lambda}(\beta)$ and $\hat{f}(\beta)$ be the minimizing parameters in the first line of equation (2.4) for $R = R^0$. These are the first R^0 principal components of $(Y - \beta \cdot X)(Y - \beta \cdot X)'$ and $(Y - \beta \cdot X)'(Y - \beta \cdot X)$, respectively. For the corresponding orthogonal projectors we use the notation $M_{\hat{\lambda}}(\beta) \equiv M_{\hat{\lambda}(\beta)}$ and $M_{\hat{f}}(\beta) \equiv M_{\hat{f}(\beta)}$. For the residuals after taking out these first R^0 principal components we write $\hat{e}(\beta) \equiv Y - \beta \cdot X - \hat{\lambda}(\beta)\hat{f}'(\beta)$.

Analogous to the expansion of $\mathcal{L}^0_{NT}(\beta)$ the perturbation theory of linear operators also provides an expansion for $M_{\hat{\lambda}}(\beta)$, $M_{\hat{f}}(\beta)$ and $\hat{e}(\beta)$ in $(\beta-\beta^0)$ and $\|e\|$, i.e. in addition to describing the sum of the perturbed eigenvalues $\mathcal{L}^0_{NT}(\beta)$ it also describes the structure of the corresponding perturbed eigenvectors. For example, we have $\hat{e}(\beta) = M_{\lambda^0} e M_{f^0} - \sum_k (\beta_k - \beta^0_k) M_{\lambda^0} X_k M_{f^0}$+higher order terms. The details of these expansions are presented in Lemma A.1 and A.2 in the appendix. These expansions are crucial when generalizing the likelihood expansion to $R > R^0$. Equation (3.1) can equivalently be written as

$$\mathcal{L}^R_{NT}(\beta) = \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{t=1}^{R-R^0} \mu_t \left[\hat{e}'(\beta)\hat{e}(\beta)\right] \, . \tag{3.7}$$

Here we used that $\hat{e}'(\beta)\hat{e}(\beta)$ is the residual of $(Y-\beta\cdot X)'(Y-\beta\cdot X)$ after subtracting the first R^0 principal components, which implies that the eigenvalues of these two matrices are the same, except from the R^0 largest ones which are missing in $\hat{e}'(\beta)\hat{e}(\beta)$. By applying the expansion of $\hat{e}(\beta)$ to this expression for $\mathcal{L}^R_{NT}(\beta)$ one obtains the following.

Theorem 3.3. *Under Assumption 1 and 3 and for $R > R^0$ we have*

(i) $\mathcal{L}^R_{NT}(\beta) = \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{t=1}^{R-R^0} \mu_t \left[A(\beta)\right] + \mathcal{L}^{R,\mathrm{rem},1}_{NT}(\beta),$

where $A(\beta) = M_{f^0}\left[e - (\beta-\beta^0)\cdot X\right]' M_{\lambda^0} \left[e - (\beta-\beta^0)\cdot X\right] M_{f^0},$

and for any constant $c > 0$

$$\sup_{\{\beta:\sqrt{N}\|\beta-\beta^0\|\leq c\}} \frac{\left|\mathcal{L}^{R,\mathrm{rem},1}_{NT}(\beta)\right|}{\sqrt{N} + \sqrt{NT}\,\|\beta-\beta^0\|} = \mathcal{O}_p\left(\frac{1}{NT}\right).$$

(ii) $\mathcal{L}^R_{NT}(\beta) = \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{t=1}^{R-R^0} \mu_t \left[B(\beta) + B'(\beta)\right] + \mathcal{L}^{R,\mathrm{rem},2}_{NT}(\beta),$

where

$$\begin{aligned}
B(\beta) = {} & \tfrac{1}{2}A(\beta) - M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \\
& + M_{f^0}\left[(\beta-\beta^0)\cdot X - e\right]' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
& + M_{f^0} e' M_{\lambda^0} \left[(\beta-\beta^0)\cdot X\right] f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
& + M_{f^0} e' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} \left[(\beta-\beta^0)\cdot X\right] M_{f^0} \\
& + B^{(eeee)} + M_{f^0} B^{(\mathrm{rem},1)}(\beta) P_{f^0} + P_{f^0} B^{(\mathrm{rem},2)} P_{f^0},
\end{aligned}$$

and

$$
\begin{aligned}
B^{(eeee)} &= -M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
&\quad + M_{f^0} e' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
&\quad - \tfrac{1}{2} M_{f^0} e' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} e' M_{\lambda^0} e M_{f^0} \\
&\quad + \tfrac{1}{2} M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} e' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \, .
\end{aligned}
$$

Here, $B^{(\mathrm{rem},1)}(\beta)$ and $B^{(\mathrm{rem},2)}$ are $T \times T$ matrices, $B^{(\mathrm{rem},2)}$ is independent of β and satisfies $\|B^{(\mathrm{rem},2)}\| = \mathcal{O}_p(1)$, and for any constant $c > 0$

$$
\sup_{\{\beta : \sqrt{N}\|\beta - \beta^0\| \le c\}} \frac{\|B^{(\mathrm{rem},1)}(\beta)\|}{1 + \sqrt{NT}\,\|\beta - \beta^0\|} = \mathcal{O}_p(1) \, ,
$$

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

Here, the remainder terms $\mathcal{L}^{R,\mathrm{rem},1}_{NT}(\beta)$ and $\mathcal{L}^{R,\mathrm{rem},2}_{NT}(\beta)$ stem from terms in $\hat{e}'(\beta)\hat{e}(\beta)$ whose spectral norm is smaller than $\mathcal{O}_p(1)$ and $o_p(1)$, respectively, within a $\sqrt{N}$ shrinking neighborhood of β after dividing by $\sqrt{N} + \sqrt{NT}\,\|\beta - \beta^0\|$ and $1 + \sqrt{NT}\,\|\beta - \beta^0\|$, respectively. Using Weyl's inequality those terms can be separated from the eigenvalues $\mu_t\,[\hat{e}'(\beta)\hat{e}(\beta)]$. The expression for $B(\beta)$ looks complicated, in particular the terms in $B^{(eeee)}$. Note however, that $B^{(eeee)}$ is β-independent and satisfies $\|B^{(eeee)}\| = \mathcal{O}_p(1)$ under our assumptions, so that it is relatively easy to deal with these terms. Note furthermore that the structure of $B(\beta)$ is closely related to the expansion of $\mathcal{L}^0_{NT}(\beta)$, since by definition we have $\mathcal{L}^0_{NT}(\beta) = (NT)^{-1}\mathrm{Tr}(\hat{e}'(\beta)\hat{e}(\beta))$, which can be approximated by $(NT)^{-1}\mathrm{Tr}(B(\beta) + B'(\beta))$. Plugging the definition of $B(\beta)$ into $(NT)^{-1}\mathrm{Tr}(B(\beta) + B'(\beta))$ one indeed recovers the terms of the approximated Hessian and score provided by Theorem 3.1, which is a convenient consistency check. We do not give explicit formulas for $B^{(\mathrm{rem},1)}(\beta)$ and $B^{(\mathrm{rem},2)}$, because those terms enter $B(\beta)$ projected by P_{f^0}, which makes them orthogonal to the leading term $A(\beta)$, so that they can only have limited influence on the eigenvalues of $B(\beta) + B'(\beta)$. The bounds on the norms of $B^{(\mathrm{rem},1)}(\beta)$ and $B^{(\mathrm{rem},2)}$ provided in the theorem are sufficient for all conclusions on the properties of $\mu_t\,[B(\beta) + B'(\beta)]$ below. The proof of the theorem can be found in the appendix.

The first part of Theorem 3.3 is useful to show that $\hat{\beta}_R$ converges to β^0 at a rate of at least $N^{3/4}$. The purpose of the second part is to show that $\hat{\beta}_R$ has the same limiting distribution as $\hat{\beta}_{R^0}$. To actually obtain these two results one requires further conditions on the β-dependence of the largest few eigenvalues of $A(\beta)$ and $B(\beta) + B'(\beta)$.

Assumption 4. *For all constants $c > 0$*

$$
\sup_{\{\beta : \sqrt{N}\|\beta - \beta^0\| \le c\}} \frac{\sum_{t=1}^{R-R^0} \left\{ \mu_t\,[A(\beta)] - \mu_t\,[A(\beta^0)] - \mu_t\left[\tilde{A}(\beta)\right] \right\}}{\sqrt{N} + N^{5/4}\|\beta - \beta^0\| + N^2\|\beta - \beta^0\|^2/\log(N)} \le \mathcal{O}_p(1) \, ,
$$

where $\quad \tilde{A}(\beta) = M_{f^0}\left[(\beta - \beta^0)\cdot X\right]' M_{\lambda^0}\left[(\beta - \beta^0)\cdot X\right] M_{f^0}.$

Corollary 3.4. *Let $R > R^0$, let Assumptions 1, 2, 3 and 4 be satisfied and furthermore assume that $C^{(1)} = \mathcal{O}_p(1)$. In the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have*

$$N^{3/4}\left(\hat{\beta}_R - \beta^0\right) = \mathcal{O}_p(1).$$

The corollary follows from the inequality $\mathcal{L}^R_{NT}(\hat{\beta}_R) \le \mathcal{L}^R_{NT}(\beta^0)$ by applying the first part of Theorem 3.3, Assumption 4, and our expansion of $\mathcal{L}^0_{NT}(\beta)$. The justfication of Assumption 4 is discussed in the next section. Knowing that $\hat{\beta}_R$ converges to β^0 at a rate of at least $N^{3/4}$ is a convenient intermediate result. It implies that we only have to study the properties of $\mathcal{L}^R_{NT}(\beta)$ within a $N^{3/4}$ shrinking neighborhood of β^0, which is reflected in the formulation of the following assumption.

Assumption 5. *For all constants $c > 0$*

$$\sup_{\{\beta: N^{3/4}\|\beta-\beta^0\|\le c\}} \frac{\left|\sum_{t=1}^{R-R^0}\left\{\mu_t\left[B(\beta)+B'(\beta)\right] - \mu_t\left[B(\beta^0)+B'(\beta^0)\right]\right\}\right|}{(1+\sqrt{NT}\|\beta-\beta^0\|)^2} = o_p(1).$$

Combining the first part of Theorem 3.3, Assumption 5, and Theorem 3.1, we find that the profile likelihood for $R > R^0$ can be written as

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

with a remainder term that satisfies for all constants $c > 0$

$$\sup_{\{\beta: N^{3/4}\|\beta-\beta^0\|\le c\}} \frac{\left|\mathcal{L}^{R,\mathrm{rem}}_{NT}(\beta)\right|}{\left(1+\sqrt{NT}\;\|\beta-\beta^0\|\right)^2} = o_p\left(\frac{1}{NT}\right).$$

This result, together with $N^{3/4}$-consistency of $\hat{\beta}_R$, gives rise to the following corollary.

Corollary 3.5. *Let $R > R^0$, let Assumptions 1, 2, 3, 4 and 5 be satisfied and furthermore assume that $C^{(1)} = \mathcal{O}_p(1)$. In the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have*

$$\sqrt{NT}\left(\hat{\beta}_R - \beta^0\right) = W^{-1}\left(C^{(1)}+C^{(2)}\right) + o_p(1) = \mathcal{O}_p(1).$$

The proof of Corollary 3.5 is analogous to that of Corollary 3.2. The combination of both corollaries shows that our main result in equation (3.2) holds, i.e. the limiting distributions of $\hat{\beta}_R$ and $\hat{\beta}_{R^0}$ are indeed identical. What is left to do is to justify the high-level assumptions 4 and 5.

4 Justification of Assumptions 4 and 5

We start with the justification of Assumption 4. We have $A(\beta) = A(\beta^0) + \tilde{A}(\beta) - A_{\mathrm{mixed}}(\beta)$, where $A_{\mathrm{mixed}}(\beta) = M_{f^0} e' M_{\lambda^0}\left[(\beta-\beta^0)\cdot X\right] M_{f^0} +$ the same term transposed. By applying

Weyl's inequality[10] we thus find

$$\sum_{t=1}^{R-R^0} \left\{ \mu_t \left[A(\beta)\right] - \mu_t \left[A(\beta^0)\right] - \mu_t \left[\tilde{A}(\beta)\right] \right\} \leq \sum_{t=1}^{R-R^0} \mu_t \left[A_{\rm mixed}(\beta)\right]$$
$$\leq 2\,(R-R^0)\, K \, \|e\| \, \|\beta - \beta^0\| \max_k \|M_{\lambda^0} X_k M_{f^0}\|. \quad (4.1)$$

For asymptotics with N and T growing at the same rate Assumption 1(ii) guarantees $\|e\| = \mathcal{O}_p(\sqrt{N})$. Using this and inequality 4.1, we find that $\|M_{\lambda^0} X_k M_{f^0}\| = \mathcal{O}_p(N^{3/4})$ is a sufficient condition for Assumption 4. This condition can be justified by assuming that $X_k = \Gamma_k f^{0\prime} + \tilde{X}_k$, where Γ_k is an $N \times R^0$ matrix and $\tilde{X}_k$ is an $N \times T$ matrix with $\|\tilde{X}\| = \mathcal{O}_p(N^{3/4})$, i.e. X_k has an approximate factor structure with the same factors that enter into the equation for Y and an idiosyncratic component $\tilde{X}_k$. Analogous to our discussion of Assumption 1(ii) we can obtain the bound on the norm of $\tilde{X}_k$ by assuming that its entries $\tilde{X}_{k,it}$ are mean zero, have bounded fourth moment and are only weakly correlated across i and t.

We have thus provided a way to justify Assumption 4 without imposing any additional condition on the error matrix e, but by restricting the data generating process for the regressors X_k. Alternatively, one can derive the statement in the assumption by imposing weaker restrictions on X_k, but making further assumptions on the error matrix e. An example of this is provided by Theorem 4.1 below, where we only assume that $X_k = \overline{X}_k + \tilde{X}_k$, with $\text{rank}(\overline{X}_k)$ being bounded, but without assuming that $\overline{X}_k$ is generated by the factors f^0.

The discussion of Assumption 5 is more complicated. By Weyl's inequality we know that the absolute value of $\mu_t\left[B(\beta) + B'(\beta)\right] - \mu_t\left[B(\beta^0) + B'(\beta^0)\right]$ is bounded by the spectral norm of $B(\beta) + B'(\beta) - B(\beta^0) - B'(\beta^0)$, which is of order $\mathcal{O}_p(N^{3/2})\|\beta - \beta^0\| + \mathcal{O}_p(N^2)\|\beta - \beta^0\|^2$. This bound is obviously too crude to justify the assumption. What we need here is a bound that not only takes into account the spectral norm of the difference between $B(\beta) + B'(\beta)$ and $B(\beta^0) + B'(\beta^0)$, but also the structure of the eigenvectors of the various matrices involved.

The assumption only restricts the properties of $B(\beta)$ in an $N^{3/4}$ shrinking neighborhood of β^0. In this shrinking neighborhood the dominant term in $B(\beta) + B'(\beta)$ is $M_{f^0} e' M_{\lambda^0} e M_{f^0}$, since its spectral norm is of order N, while the spectral norm of the remaining terms, e.g. $A_{\rm mixed}(\beta)$ above, is at most of order $N^{3/4}$. Our goal is to show that the largest few eigenvalues of $B(\beta) + B'(\beta)$ only differ by $o_p(1)$ from those of the leading term $M_{f^0} e' M_{\lambda^0} e M_{f^0}$, within the shrinking neighborhood of β^0. To do so, we first need to introduce some notation.

Let $w_t \in \mathbb{R}^T$, $t = 1, \ldots, T - R^0$, be the normalized eigenvectors of $M_{f^0} e' M_{\lambda^0} e M_{f^0}$ with the constraint $f^{0\prime} w_t = 0$, and let ρ_t, $t = 1, \ldots, T - R^0$, be the corresponding eigenvalues. Let $v_i \in \mathbb{R}^N$, $i = 1, \ldots, N - R^0$, be the normalized eigenvectors of $M_{\lambda^0} e M_{f^0} e' M_{\lambda^0}$ with the constraint $\lambda^{0\prime} v_i = 0$, and let ρ_i, $i = 1, \ldots, N - R^0$, be the corresponding eigenvalues.[11]

[10] Weyl's inequality says $\mu_m(G + H) \leq \mu_m(G) + \mu_1(H)$ for arbitrary symmetric $n \times n$ matrices G and H and $1 \leq m \leq n$. Here, we refer to a generalization of this, which reads $\sum_{t=1}^{m} \mu_t(G+H) \leq \sum_{t=1}^{m} \mu_t(G) + \sum_{t=1}^{m} \mu_t(H)$. These inequalities are standard results in linear algebra and are readily derived from the Courant-Fischer-Weyl min-max principle.

[11] For $T < N$ the vectors v_i, $i = T - R^0 + 1, \ldots, N - R^0$, correspond to null-eigenvalues, and if there are multiple null-eigenvalues those v_i are not uniquely defined. In that case we assume that those v_i are drawn randomly from the Haar measure on the unit sphere of the corresponding null-eigenspace. For $T > N$ we assume

We assume that eigenvalues are sorted in decreasing order, i.e. $\rho_1 \geq \rho_2 \geq \ldots$. Note that the eigenvalues ρ_t and ρ_i are identical for $t = i$. Let

$$
\begin{aligned}
&d_{NT}^{(1)} = \max_{i,t,k} |v_i' X_k w_t|, && d_{NT}^{(2)} = \max_{i} \|v_i' e P_{f^0}\|, && d_{NT}^{(3)} = \max_{t} \|w_t' e' P_{\lambda^0}\|, \\
&d_{NT}^{(4)} = N^{-3/4} \max_{i} \|v_i' X_k P_{f^0}\|, && d_{NT}^{(5)} = N^{-3/4} \max_{t} \|w_t' X_k' P_{\lambda^0}\|,
\end{aligned}
$$

where $i = 1 \ldots N - R^0$, $t = 1, \ldots, T - R^0$ and $k = 1 \ldots K$. Furthermore, define $d_{NT} = \max\left(1,\ d_{NT}^{(1)},\ d_{NT}^{(2)},\ d_{NT}^{(3)},\ d_{NT}^{(4)},\ d_{NT}^{(5)}\right)$.

Theorem 4.1. *Let assumptions 1 and 3 hold, let $R > R^0$ and consider a limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Assume that $\rho_{R-R^0} > a\, N$, wpa1, for some constant $a > 0$. Furthermore, let there exists a series of integers $q_{NT} > R - R^0$ such that*

$$
d_{NT}\, q_{NT} = o_p(N^{1/4}) \,, \qquad \text{and} \qquad \frac{1}{q_{NT}} \sum_{t=q_{NT}}^{T-R^0} \frac{1}{\rho_{R-R^0} - \rho_t} = \mathcal{O}_p(1) \,.
$$

Then, for all constants $c > 0$ and $t = 1, \ldots, R - R^0$ we have

$$
\sup_{\{\beta : N^{3/4} \|\beta - \beta^0\| \leq c\}} \left| \mu_t \left(B(\beta) + B'(\beta) \right) - \rho_t \right| = o_p(1),
$$

which implies that Assumption 5 is satisified.

We can now justify Assumption 5 by showing that the conditions of Theorem 4.1 are satisfied. The following discussion is largely heuristic. Since v_i and w_t are the normalized eigenvalues of $M_{f^0} e' M_{\lambda^0} e M_{f^0}$ and $M_{\lambda^0} e M_{f^0} e' M_{\lambda^0}$ we expect them to be essentially uncorrelated with X_k and $e P_{f^0}$, and therefore we expect $v_i' X_k w_t = \mathcal{O}_p(1)$, $\|v_i' e P_{f^0}\| = \mathcal{O}_p(1)$, $\|w_t' e' P_{\lambda^0}\| = \mathcal{O}_p(1)$. We also expect $\|v_i' X_k P_{f^0}\| = \mathcal{O}_p(\sqrt{T})$ and $\|w_t' X_k' P_{\lambda^0}\| = \mathcal{O}_p(\sqrt{N})$, which is different to the preceding terms with e, since X_k can be correlated with f^0 and λ^0. In the definition of d_{NT} the maxima over these terms are taken over i and t, so that we anticipate some weak dependence of d_{NT} on N (or equivalently T). Note that we need $d_{NT} = o_p(N^{1/4})$ since otherwise q_{NT} does not exist. The key to making this discussion rigorous and show that indeed $d_{NT} = o_p(N^{1/4})$, or smaller, is a good knowledge of the properties of the eigenvectors v_i and w_t. If the entries e_{it} are *iid* normal, then the matrix of v_i's and w_t's is Haar-distributed (on the $N - R^0$ and $T - R^0$ dimensional subspaces spanned by M_{λ^0} and M_{f^0}). In that case the formalization of the above discussion becomes relatively easy, and the result is summarized in Theorem 4.2 below.

The conjecture in the random matrix theory literature is that the limiting distribution of the eigenvectors of a random covariance matrix is "distribution free", i.e. is independent of the particular distribution of e_{it} (see, e.g., Silverstein (1990), Bai (1999)). However, we are not aware of a formulation and corresponding proof of this conjecture that is sufficient for our purposes.

the same for w_t, $t = N - R^0, \ldots, T - R^0$. This specification avoids correlation between X_k and those v_i and w_t being caused by a particular choice of the eigenvectors that correspond to degenerate null-eigenvalues.

The second condition in Theorem 4.1 is on the eigenvectors ρ_t of the random covariance matrix $M_{f^0} e' M_{\lambda^0} e M_{f^0}$. Eigenvalues are studied more intensely than eigenvectors in the random matrix theory literature, and it is well-known that the properly normalized empirical distribution of the eigenvalues (the so called empirical spectral distribution) of an *iid* sample covariance matrix converges to the Marčenko-Pastur-law (Marčenko and Pastur (1967)) for asymptotics where N and T grow at the same rate. This means that the sum over the eigenvalues ρ_t in Theorem 4.1 asymptotically becomes an integral over the Marčenko-Pastur limiting spectral distribution.[12] To derive a bound on this sum, one furthermore needs to know the asymptotic properties of ρ_{R-R^0}. For random covariance matrices from *iid* normal errors, it is known from Johnstone (2001) and Soshnikov (2002) that the properly normalized few largest eigenvalues converge to the Tracy-Widom law.[13].

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

Thus, if the e_{it} are *iid* normal, then we can show that the conditions of Theorem 4.1 are satisfied, and we can therefore verify that the high-level assumptions of the last section hold. This result is summarized in the following theorem.

Theorem 4.2. *Let $R > R^0$, let Assumption 3 hold and consider a limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$. Furthermore, assume that*

(i) For all $k = 1, \ldots, K$ we can decompose $X_k = \overline{X}_k + \tilde{X}_k$, , such that

$$\|\tilde{X}_k\| = \mathcal{O}_p(N^{3/4}), \quad \|\tilde{X}_k\|_{HS} = \mathcal{O}_p(\sqrt{NT}), \quad \|\overline{X}_k\| = \mathcal{O}_p(\sqrt{NT}), \quad \operatorname{rank}(\overline{X}_k) \leq Q_k,$$

where Q_k is independent of N and T. For the $K \times K$ matrix $\tilde{W}$ defined by $\tilde{W}_{k_1 k_2} = \frac{1}{NT} \operatorname{Tr}(\tilde{X}_k \tilde{X}'_k)$ we assume that $\operatorname{plim}_{N,T\to\infty} W_{k_1 k_2} > 0$. In addition, we assume that $\mathbb{E}\left|(M_{\lambda^0} X_k M_{f^0})_{it}\right|^{24+\epsilon}$, $\mathbb{E}\left|(M_{\lambda^0} X_k)_{it}\right|^{6+\epsilon}$ and $\mathbb{E}\left|(X_k M_{f^0})_{it}\right|^{6+\epsilon}$ are bounded uniformly across i, j, N and T for some $\epsilon > 0$. .

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

Then, the Assumptions 1, 2, 4 and 5 are satisfied and we have $C^{(1)} = \mathcal{O}_p(1)$. By Corollary 3.2 and 3.5 we can therefore conclude $\sqrt{NT}\left(\hat{\beta}_R - \beta^0\right) = \sqrt{NT}\left(\hat{\beta}_{R^0} - \beta^0\right) + o_p(1)$.

The proofs for Theorem 4.1 and Theorem 4.2 are provided in the supplementary material to this paper. It seems to be quite challenging to extend Theorem 4.2 to non-*iid*-normal

[12] To make this argument mathematically rigorous one needs to know the convergence rate of the empirical spectral distribution to its limit law, which is a ongoing research subject in the literature, e.g. Bai (1993), Bai, Miao and Yao (2004), Götze and Tikhomirov (2010).

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

	$N=T=50$				$N=T=100$			
	$e_{it}\sim\mathcal{N}(0,1)$		$e_{it}\sim$ t(5)		$e_{it}\sim\mathcal{N}(0,1)$		$e_{it}\sim$ t(5)	
	bias	std	bias	std	bias	std	bias	std
$R=0$	0.42741	0.02710	0.42788	0.02699	0.42806	0.01890	0.42813	0.01884
$R=1$	0.29566	0.05712	0.29633	0.05830	0.29597	0.03725	0.29541	0.03717
$R=2$	0.00047	0.02015	0.00175	0.02722	0.00005	0.00974	0.00057	0.01296
$R=3$	0.00046	0.02101	0.00139	0.02693	0.00007	0.00993	0.00062	0.01314
$R=4$	0.00051	0.02183	0.00140	0.02792	0.00010	0.01012	0.00062	0.01335
$R=5$	0.00042	0.02259	0.00137	0.02888	0.00011	0.01028	0.00061	0.01361

Table 1: Simulation results for the bias and standard error (std) of the QMLE $\hat{\beta}_R$ for different value of R, two different sample sizes N and T, and the two different specifications for e_{it}. The data generating process is described in the main text, in particular the true number of factors here is $R^0 = 2$. We used $10,000$ repetitions in the simulation.

e_{it}, given the present status of the literature on eigenvalues and eigenvectors of large dimensional random covariance matrices, and we would like to leave this as a future research topic.

5 Monte Carlo Simulations

Here, we consider a panel model with one regressor ($K = 1$), two factors ($R^0 = 2$) and the following data generating process (DGP)

$$Y_{it} = \beta^0 X_{it} + \sum_{r=1}^{2} \lambda_{ir} f_{tr} + e_{it}, \qquad X_{it} = 1 + \tilde{X}_{it} + \sum_{r=1}^{2} (\lambda_{ir} + \chi_{ir}) f_{tr}, \tag{5.1}$$

where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The random variables $\tilde{X}_{it}$, λ_{ir}, f_{tr}, χ_{ir} and e_{it} are mutually independent, $\tilde{X}_{it}$ is distributed as $iid\mathcal{N}(1,1)$, and λ_{ir}, f_{tr} and χ_{ir} are all distributed as $iid\mathcal{N}(1,1)$. For e_{it} we also assume that it is iid across i and t, but we consider two different specifications for the marginal distribution, namely either $\mathcal{N}(0,1)$ or a Student's t-distribution with 5 degrees of freedom. We choose $\beta^0 = 1$, and use $10,000$ repetitions in our simulation. For each draw of Y and X we compute the QMLE $\hat{\beta}_R$ according to equation (2.3) for different values of R.

Table 1 reports the bias and standard error of $\hat{\beta}_R$ for sample sizes $N = T = 50$ and $N = T = 100$. For $R = 0$ (OLS estimator) and $R = 1$ we have $R < R^0$, i.e. less factors are used in the estimation than are present in the DGP. As a result of this, the QMLE is heavily biased for these values of R, since the factor structure in the DGP is correlated with the regressors, but is not controlled for in the estimation. In contrast, for all values $R \geq R^0$ the bias of the QMLE is negligible compared to its standard error. Furthermore, the standard error remains almost constant as R increases beyond $R = R^0$; concretely from $R = 2$ to $R = 5$ it increases only by about 7% for $N = T = 50$ and only by 5% for $N = T = 100$.

Table 2 reports quantiles of the appropriately normalized QMLE for $R \geq R^0$ and $N = T = 100$. One finds that the quantiles remain almost constant as R increases. In

$e_{it} \sim$		Quantiles of $\sqrt{NT}\left(\hat{\beta}_R - \beta^0\right)$ 2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
$\mathcal{N}(0,1)$	$R=2$	-1.903	-1.598	-1.239	-0.643	0.008	0.663	1.240	1.616	1.916
	$R=3$	-1.977	-1.625	-1.253	-0.650	0.011	0.658	1.276	1.650	1.952
	$R=4$	-1.998	-1.664	-1.275	-0.666	0.016	0.682	1.296	1.694	1.992
	$R=5$	-2.041	-1.672	-1.284	-0.682	0.019	0.698	1.328	1.723	2.000
t(5)	$R=2$	-2.537	-2.095	-1.614	-0.807	0.072	0.935	1.716	2.188	2.573
	$R=3$	-2.550	-2.116	-1.642	-0.817	0.071	0.946	1.757	2.206	2.626
	$R=4$	-2.592	-2.147	-1.653	-0.829	0.067	0.961	1.796	2.259	2.664
	$R=5$	-2.652	-2.181	-1.688	-0.854	0.071	0.972	1.805	2.296	2.720

Table 2: Simulation results for the quantiles of $\sqrt{NT}\left(\hat{\beta}_R - \beta^0\right)$ for $N = T = 100$, the two different specifications of e_{it}, different values of R, and the data generating process as described in the main text with $R^0 = 2$. We used $10,000$ repetitions in the simulation.

particular, the differences in the quantiles for different values of R are relatively small compared to the differences between the quantiles, so that the size of a test statistics that is based on $\hat{\beta}_R$ is essentially independent of the choice of $R \geq R^0$.

The findings of the Monte Carlo simulations described in the last two paragraph hold just as well for the specification with normally distributed as for the specification where e_{it} has Student's t-distribution. From this finding one may conjecture that Theorem 4.2 also holds for more general error distributions.

6 Conclusions

In this paper we showed that under certain regularity conditions the limiting distribution of the QMLE of a linear panel regression with interactive fixed effects does not change when we include redundant factors in the estimation. The important empirical implication of this result is that one can use an upper bound of the number of factors in the estimation without asymptotic efficiency loss. For inference on the regression coefficients one thus need not worry about consistent estimation of the number of factors in the model. As regularity conditions we mostly impose high-level assumptions, and we verify that these hold under *iid* normal errors. Our simulation results suggest that normality of the error distribution is not necessary. Along the lines of the arguments presented in Section 4, we expect that progress in the literature on large dimensional random covariance matrices will allow verification of our high-level assumptions under more general error distributions. This is a vital and interesting topic for future research.

A Appendix

A.1 Proof of Consistency

Proof of Theorem 2.1. We first establish a lower bound on $\mathcal{L}^0_{NT}(\beta)$. Consider the last expression for $\mathcal{L}^0_{NT}(\beta)$ in equation (2.4) and plug in $Y = \sum_k \beta^0_k X_k + \lambda^0 f^{0\prime} + e$, then replace

$\lambda^0 f^{0\prime}$ by $\lambda f'$, and minimize over the $N \times R^0$ matrix λ and the $T \times R^0$ matrix f. This gives

$$
\begin{aligned}
\mathcal{L}^0_{NT}(\beta) &\geq \frac{1}{NT} \min_{\tilde{F}} \operatorname{Tr} \left[\left(\sum_k (\beta^0_k - \beta_k) X_k + e \right) M_{\tilde{F}} \left(\sum_k (\beta^0_k - \beta_k) X_k + e \right)' \right] , \\
&\geq b \, \|\beta - \beta^0\|^2 + \mathcal{O}_p \left(\frac{\|\beta - \beta^0\|}{\sqrt{\min(N,T)}} \right) + \frac{1}{NT} \operatorname{Tr} \left(ee' \right) + \mathcal{O}_p \left(\frac{1}{\min(N,T)} \right) . \quad \text{(A.1)}
\end{aligned}
$$

where in the first line we minimize over all $T \times (R + R^0)$ matrices $\tilde{F}$, and to arrive at the second line we decomposed the expression in the component quadratic in $(\beta - \beta^0)$, linear in $(\beta - \beta^0)$ and independent of $(\beta - \beta^0)$ and applied Assumption 1 and 2. Next, we establish an upper bound on $\mathcal{L}^0_{NT}(\beta^0)$. We have

$$
\begin{aligned}
\mathcal{L}^0_{NT}(\beta^0) &= \frac{1}{NT} \sum_{t=R+1}^{T} \mu_t \left[\left(\lambda^0 f^{0\prime} + e \right)' \left(\lambda^0 f^{0\prime} + e \right) \right] \\
&\leq \frac{1}{NT} \operatorname{Tr} \left(e' M_{\lambda^0} e \right) = \frac{1}{NT} \operatorname{Tr} \left(ee' \right) + \mathcal{O}_p \left(\frac{1}{\min(N,T)} \right) . \qquad \text{(A.2)}
\end{aligned}
$$

Further details regarding the derivation of the bounds (A.1) and (A.2) are presented in the supplementary material. Since we could choose $\beta = \beta^0$ in the minimization of β, the optimal $\hat{\beta}$ needs to satisfy $\mathcal{L}^0_{NT}(\hat{\beta}) \leq \mathcal{L}^0_{NT}(\beta^0)$. With the above results we thus find

$$
b \, \|\hat{\beta} - \beta^0\|^2 + \mathcal{O}_p \left(\frac{\|\hat{\beta} - \beta^0\|}{\sqrt{\min(N,T)}} \right) + \mathcal{O}_p \left(\frac{1}{\min(N,T)} \right) \leq 0 . \qquad \text{(A.3)}
$$

From this it follows that $\|\hat{\beta} - \beta^0\| = \mathcal{O}_p \left(\min(N,T)^{-1/2} \right)$, which is what we wanted to show. ■

A.2 Proof of Likelihood Expansion

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

$$
\begin{aligned}
d_{\max}(\lambda^0, f^0) &= \frac{1}{\sqrt{NT}} \, \|\lambda^0 f^{0\prime}\| = \sqrt{\frac{1}{NT} \, \mu_1(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \, , \\
d_{\min}(\lambda^0, f^0) &= \sqrt{\frac{1}{NT} \, \mu_R(\lambda^{0\prime} f^0 f^{0\prime} \lambda^0)} \, , \qquad \text{(A.4)}
\end{aligned}
$$

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

$$
r_0(\lambda^0, f^0) = \left(\frac{4 d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} + \frac{1}{2 d_{\max}(\lambda^0, f^0)} \right)^{-1} . \qquad \text{(A.5)}
$$

Lemma A.1. *If the following condition is satisfies*

$$
\sum_{k=1}^{K} \left| \beta^0_k - \beta_k \right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} < r_0(\lambda^0, f^0) \, , \qquad \text{(A.6)}
$$

then

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

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

where the expansion coefficients are given by[14]

$$\begin{aligned} L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) &= \tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{(k_1},\, X_{k_2}, \ldots, X_{k_g)}\right) \\ &= \frac{1}{g!}\left[\tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) + \textit{all permutations of } k_1, \ldots, k_g\right], \end{aligned}$$

i.e. $L^{(g)}$ is obtained by total symmetrization of the last g arguments of[15]

$$\begin{aligned} &\tilde{L}^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) \\ &\quad = \sum_{p=1}^{g} (-1)^{p+1} \sum_{\substack{\nu_1 + \ldots + \nu_p = g \\ m_1 + \ldots + m_{p+1} = p-1 \\ 2 \geq \nu_j \geq 1\ ,\ m_j \geq 0}} \mathrm{Tr}\left(S^{(m_1)}\, \mathcal{T}_{k_1\ldots}^{(\nu_1)}\, S^{(m_2)} \,\ldots\, S^{(m_p)}\, \mathcal{T}_{\ldots k_g}^{(\nu_p)}\, S^{(m_{p+1})}\right), \end{aligned}$$

with

$$\begin{aligned} S^{(0)} &= -M_{\lambda^0}\,, & S^{(m)} &= \left[\lambda^0(\lambda^{0\prime}\lambda^0)^{-1}(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\right]^l\,, \ \textit{for } m \geq 1, \\ \mathcal{T}_k^{(1)} &= \lambda^0\, f^{0\prime}\, X_k' + X_k\, f^0\, \lambda^{0\prime}\,, \quad & \mathcal{T}_{k_1 k_2}^{(2)} &= X_{k_1}\, X_{k_2}'\,, \qquad \textit{for } k, k_1, k_2 = 0 \ldots K\,, \\ X_0 &= \frac{\sqrt{NT}}{\|e\|}\, e\,, & X_k &= X_k\,, \qquad \textit{for } k = k = 1 \ldots K\,. \end{aligned}$$

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

$$M_{\hat{\lambda}}\left(\beta\right) = \sum_{g=0}^{\infty} \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \ldots \sum_{k_g=0}^{K} \epsilon_{k_1}\, \epsilon_{k_2} \,\ldots\, \epsilon_{k_g}\, M^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right),$$

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

$$\begin{aligned} M^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) &= \tilde{M}^{(g)}\left(\lambda^0,\, f^0,\, X_{(k_1},\, X_{k_2}, \ldots, X_{k_g)}\right) \\ &= \frac{1}{g!}\left[\tilde{M}^{(g)}\left(X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) + \textit{all permutations of } k_1, \ldots, k_g\right], \end{aligned}$$

[14] Here we use the round bracket notation $(k_1, k_2, \ldots, k_g)$ for total symmetrization of these indices.

[15] One finds $\tilde{L}^{(1)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) = 0$, which is why the sum in the power series of $\mathcal{L}_{NT}^0$ starts from $g = 2$ instead of $g = 1$.

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

$$\begin{aligned}
&\tilde{M}^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) \\
&\quad = \sum_{p=1}^{g} (-1)^{p+1} \sum_{\substack{\nu_1 + \ldots + \nu_p = g \\ m_1 + \ldots + m_{p+1} = p \\ 2 \geq \nu_j \geq 1 \;,\; m_j \geq 0}} S^{(m_1)}\, \mathcal{T}^{(\nu_1)}_{k_1 \ldots}\, S^{(m_2)} \ldots S^{(m_p)}\, \mathcal{T}^{(\nu_p)}_{\ldots k_g}\, S^{(m_{p+1})} \,,
\end{aligned}$$

where $S^{(m)}$, $\mathcal{T}^{(1)}_{k}$, $\mathcal{T}^{(2)}_{k_1 k_2}$, *and* X_k *are given above.*

(iii) For $g \geq 3$ *the coefficients* $L^{(g)}$ *in the series expansion of* $\mathcal{L}^0_{NT}(\beta)$ *are bounded as follows*

$$\begin{aligned}
&\frac{1}{NT}\left|L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right)\right| \\
&\qquad \leq \frac{R\, g\, d^2_{\min}(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)}\right)^g \frac{\|X_{k_1}\|}{\sqrt{NT}}\, \frac{\|X_{k_2}\|}{\sqrt{NT}} \ldots \frac{\|X_{k_g}\|}{\sqrt{NT}} \,.
\end{aligned}$$

Under the stronger condition

$$\sum_{k=1}^{K} \left|\beta^0_k - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} \;<\; \frac{d^2_{\min}(\lambda^0, f^0)}{16\, d_{\max}(\lambda^0, f^0)} \,, \tag{A.7}$$

we therefore have the following bound on the remainder when the series expansion for $\mathcal{L}^0_{NT}(\beta)$ *is truncated at order* $G \geq 2$*:*

$$\begin{aligned}
&\left|\mathcal{L}^0_{NT}\left(\beta\right) - \frac{1}{NT} \sum_{g=2}^{G} \sum_{k_1=0}^{K} \ldots \sum_{k_g=0}^{K} \epsilon_{k_1} \ldots \epsilon_{k_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right)\right| \\
&\qquad\qquad \leq \frac{R\,(G+1)\, \alpha^{G+1}\, d^2_{\min}(\lambda^0, f^0)}{2(1-\alpha)^2} \,,
\end{aligned}$$

where

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

(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* $g \geq 1$

$$\left\|M^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, \ldots, X_{k_g}\right)\right\| \leq \frac{g}{2}\left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)}\right)^g \frac{\|X_{k_1}\|}{\sqrt{NT}}\, \frac{\|X_{k_2}\|}{\sqrt{NT}} \ldots \frac{\|X_{k_g}\|}{\sqrt{NT}} \,.$$

Under the condition (A.7) *we therefore have the following bound on operator norm* *of the remainder of the series expansion of $M_{\hat{\lambda}}(\beta)$, for $G \geq 0$*

$$\left\| M_{\hat{\lambda}}(\beta) - \sum_{g=0}^{G} \sum_{k_1=0}^{K} \ldots \sum_{k_g=0}^{K} \epsilon_{k_1} \ldots \epsilon_{k_g} M^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, \ldots, X_{k_g}\right) \right\| \leq \frac{(G+1)\,\alpha^{G+1}}{2(1-\alpha)^2} .$$

The proof of the preceding lemma is presented in the supplementary material.

Proof of Theorem 3.1. The R^0 non-zero eigenvalues of the matrix $\lambda^{0\prime} f^0 f^{0\prime} \lambda^0 / NT$ are identical to the eigenvalues of the $R^0 \times R^0$ matrix $(f^0 f^{0\prime}/T)^{-1/2}(\lambda^0 \lambda^{0\prime}/N)(f^0 f^{0\prime}/T)^{-1/2}$, and Assumption 3 guarantees that these eigenvalues, including $d_{\max}(\lambda^0, f^0)$ and $d_{\min}(\lambda^0, f^0)$ converge to positive constants in probability. Therefore, also $r_0(\lambda^0, f^0)$ converges to a positive constant in probability.

Assumptions 1 and 3 furthermore imply that in the limit $N, T \to \infty$ with $N/T \to \kappa^2$, $0 < \kappa < \infty$, we have

$$\frac{\|\lambda^0\|}{\sqrt{N}} = \mathcal{O}_p(1) , \quad \frac{\|f^0\|}{\sqrt{T}} = \mathcal{O}_p(1) , \quad \left\| \left(\frac{\lambda^{0\prime}\lambda^0}{N} \right)^{-1} \right\| = \mathcal{O}_p(1) , \quad \left\| \left(\frac{f^{0\prime} f^0}{T} \right)^{-1} \right\| = \mathcal{O}_p(1) ,$$
$$\frac{\|X_k\|}{\sqrt{NT}} = \mathcal{O}_p(1) , \quad \frac{\|e\|}{\sqrt{NT}} = \mathcal{O}_p\left(N^{-1/2}\right) . \tag{A.8}$$

Thus, for $\left\|\beta - \beta^0\right\| \leq c_{NT}$, $c_{NT} = o(1)$, we have $\alpha \to 0$ as $N, T \to \infty$, i.e. the condition (A.7) in part (iii) of Lemma A.1 is asymptotically satisfied, and by applying the Lemma we find

$$\frac{1}{NT} (\epsilon_0)^{g-r} L^{(g)}\left(\lambda^0, f^0, X_{k_1}, \ldots, X_{k_r}, X_0, \ldots, X_0\right) = \mathcal{O}_p\left(\left(\frac{\|e\|}{\sqrt{NT}} \right)^{g-r} \right) = \mathcal{O}_p\left(N^{-\frac{g-r}{2}} \right) , \tag{A.9}$$

where we used $\epsilon_0 X_0 = e$ and the linearity of $L^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, \ldots, X_{k_g}\right)$ in the arguments X_k. Truncating the expansion of $\mathcal{L}^0_{NT}(\beta)$ at order $G = 3$ and applying the corresponding result in Lemma A.1(iii) we obtain

$$\begin{aligned}
\mathcal{L}^0_{NT}(\beta) =& \frac{1}{NT} \sum_{k_1,k_2=0}^{K} \epsilon_{k_1}\epsilon_{k_2} L^{(2)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}\right) \\
& + \frac{1}{NT} \sum_{k_1,k_2,k_3=0}^{K} \epsilon_{k_1}\epsilon_{k_2}\epsilon_{k_3} L^{(3)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, X_{k_3}\right) + \mathcal{O}_p\left(\alpha^4\right) \\
=& \mathcal{L}^0_{NT}(\beta^0) - \frac{2}{\sqrt{NT}} \left(\beta - \beta^0\right)' \left(C^{(1)} + C^{(2)}\right) \\
& + \left(\beta - \beta^0\right)' W \left(\beta - \beta^0\right) + \mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta) ,
\end{aligned} \tag{A.10}$$

where, using (A.9) we find

$$
\begin{aligned}
\mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta) =& \frac{3}{NT}\sum_{k_1,k_2=1}^{K} \epsilon_{k_1}\epsilon_{k_2}\epsilon_0 L^{(3)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, X_0\right) \\
&+ \frac{1}{NT}\sum_{k_1,k_2,k_3=1}^{K} \epsilon_{k_1}\epsilon_{k_2}\epsilon_{k_3} L^{(3)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, X_{k_3}\right) \\
&+ \mathcal{O}_p\left[\left(\sum_{k=1}^{K}|\beta^0_k-\beta_k|\frac{\|X_k\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)^4\right] - \mathcal{O}_p\left[\left(\frac{\|e\|}{\sqrt{NT}}\right)^4\right] \\
=& \mathcal{O}_p\left(\|\beta-\beta^0\|^2 N^{-1/2}\right) + \mathcal{O}_p\left(\|\beta-\beta^0\|^3\right) + \mathcal{O}_p\left(\|\beta-\beta^0\| N^{-3/2}\right) \\
&+ \mathcal{O}_p\left(\|\beta-\beta^0\|^2 N^{-1}\right) + \mathcal{O}_p\left(\|\beta-\beta^0\|^3 N^{-1/2}\right) + \mathcal{O}_p\left(\|\beta-\beta^0\|^4\right) \ . \qquad \text{(A.11)}
\end{aligned}
$$

Here $\mathcal{O}_p\left[\left(\frac{\|e\|}{\sqrt{NT}}\right)^4\right]$ is not just some term of that order, but exactly the term of that order contained in $\mathcal{O}_p(\alpha^4) = \mathcal{O}_p\left[\left(\sum_{k=1}^K|\beta^0_k-\beta_k|\frac{\|X_k\|}{\sqrt{NT}}+\frac{\|e\|}{\sqrt{NT}}\right)^4\right]$. This term is not present in $\mathcal{L}^{0,\mathrm{rem}}_{NT}(\beta)$ since it is already contained in $\mathcal{L}^{0}_{NT}(\beta^0)$.[16] Equation (A.11) shows that the remainder satisfies the bound stated in the theorem, which concludes the proof. ■

Proof of Corollary 3.2. Using Assumption 2(ii) we find for $R = R^0$

$$
\begin{aligned}
W \geq \mu_K(W) &= \min_{\{\alpha\in\mathbb{R}^K,\|\alpha\|=1\}} \alpha' W \alpha = \min_{\{\alpha\in\mathbb{R}^K,\|\alpha\|=1\}} \frac{1}{NT}\mathrm{Tr}\left(M_{f^0}X'_\alpha M_{\lambda^0}X_\alpha M_{f^0}\right) \\
&\geq \frac{1}{NT}\sum_{t=2R^0+1}^{T}\mu_t\left(X'_\alpha X_\alpha\right) \geq b_2 \ , \qquad \text{wpa1}, \qquad \text{(A.12)}
\end{aligned}
$$

and therefore $W^{-1} \leq 1/b_2$ wpa1. Using Assumption 1 we find

$$
|C^{(2)}_k| \leq \frac{9R^0}{2\sqrt{NT}}\|e\|^2\|X_k\|\left\|\lambda^0\left(\lambda^{0\prime}\lambda^0\right)^{-1}\left(f^{0\prime}f^0\right)^{-1}f^{0\prime}\right\| = \mathcal{O}_p(1) \ , \qquad \text{(A.13)}
$$

and therefore $\gamma \equiv W^{-1}\left(C^{(1)}+C^{(2)}\right)/\sqrt{NT} = \mathcal{O}_p(1/\sqrt{NT})$. Applying Theorem 3.1 to the inequality $\mathcal{L}^0_{NT}(\hat\beta_{R^0}) \leq \mathcal{L}^0_{NT}\left(\beta^0+\gamma\right)$ then gives

$$
\begin{aligned}
\left(\hat\beta_{R^0}-\beta^0-\gamma\right)' W \left(\hat\beta_{R^0}-\beta^0-\gamma\right) &\leq \mathcal{L}^{0,\mathrm{rem}}_{NT}(\gamma) - \mathcal{L}^{0,\mathrm{rem}}_{NT}(\hat\beta_{R^0}) \\
&= o_p\left(\frac{1}{NT}\right) - \mathcal{L}^{0,\mathrm{rem}}_{NT}(\hat\beta_{R^0}) \ . \qquad \text{(A.14)}
\end{aligned}
$$

From this and consistency of $\hat\beta_{R^0}$ it follows that $\sqrt{NT}(\hat\beta_{R^0}-\beta^0) = \mathcal{O}_p(1)$, since otherwise the inequality is violated asymptotically due to the bound on $\mathcal{L}^{0,\mathrm{rem}}_{NT}(\hat\beta_{R^0})$. From $\sqrt{NT}$ consistency of $\hat\beta_{R^0}$ it now follows that $\mathcal{L}^{0,\mathrm{rem}}_{NT}(\hat\beta_{R^0}) = o_p(1/NT)$, and using this the above

[16] Alternatively, we could have truncated the expansion at order $G = 4$. Then, the term $\mathcal{O}_p\left[\left(\frac{\|e\|}{\sqrt{NT}}\right)^4\right]$ would be more explicit, namely it would equal $\frac{1}{NT}\epsilon_0^4 L^{(4)}\left(\lambda^0, f^0, X_0, X_0, X_0, X_0\right)$, which is clearly contained in $\mathcal{L}^0_{NT}(\beta^0)$.

inequality yields $\sqrt{NT}(\hat{\beta}_{R^0} - \beta^0 - \gamma) = o_p(1)$, which proves the corollary. ■

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

$$\hat{e}(\beta) = M_{\lambda^0}\, e\, M_{f^0} + \hat{e}_e^{(1)} + \hat{e}_e^{(2)} - \sum_{k=1}^{K} \left(\beta_k - \beta_k^0\right) \left(\hat{e}_{X,k}^{(1)} + \hat{e}_{X,k}^{(2)}\right) + \hat{e}^{(\mathrm{rem})}(\beta) \ ,$$

where the $N \times T$ matrix valued expansion coefficients read

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

and the remainder term satisfies for any sequence $c_{NT} \to 0$

$$\sup_{\{\beta : \|\beta - \beta^0\| \leq c_{NT}\}} \frac{\left\|\hat{e}^{(\mathrm{rem})}(\beta)\right\|}{N\|\beta - \beta^0\|^2 + \|\beta - \beta^0\| \ + N^{-1}} = \mathcal{O}_p\left(1\right) \ .$$

Proof. The general expansion of $M_{\hat{\lambda}}(\beta)$ is given in Lemma A.1, and the analogous expansion for $M_{\hat{f}}(\beta)$ is obtained by applying the symmetry $N \leftrightarrow T$, $\lambda \leftrightarrow f$, $e \leftrightarrow e'$, $X_k \leftrightarrow X_k'$. Lemma S.1 in the supplementary material provides a more explicit version of these projector expansions. For the residuals $\hat{e}(\beta)$ we have

$$\hat{e}(\beta) = M_{\hat{\lambda}}(\beta)\ (Y - \beta \cdot X)\ M_{\hat{f}}(\beta) = M_{\hat{\lambda}}(\beta)\ \left[e - \left(\beta - \beta^0\right) \cdot X + \lambda^0 f^{0\prime}\right]\ M_{\hat{f}}(\beta) \ , \quad \text{(A.15)}$$

and plugging in the expansions of $M_{\hat{\lambda}}(\beta)$ and $M_{\hat{f}}(\beta)$ it is straightforward to derive the expansion of $\hat{e}(\beta)$ from this, including the bound on the remainder. ■

Proof of Theorem 3.3. The terms in $B(\beta)+B'(\beta)$ in addition to $A(\beta)$ all have a spectral norm of order $\mathcal{O}_p(\sqrt{N})$ for $\sqrt{N}\|\beta - \beta^0\| \leq c$. Thus, the first part of the Theorem directly follows from the second part by applying Weyl's inequality. What is left to show is that the second part holds. Applying the expansion $\hat{e}(\beta)$ in Lemma A.2 together with $\|M_{\lambda^0} e M_{f^0}\| = \mathcal{O}_p(\sqrt{N})$, $\|\hat{e}^{(1)}_e\| = \mathcal{O}_p(1)$, $\|\hat{e}^{(2)}_e\| = \mathcal{O}_p(N^{-1/2})$, $\|\hat{e}^{(1)}_k\| = \mathcal{O}_p(N)$ $\|\hat{e}^{(2)}_k\| = \mathcal{O}_p(\sqrt{N})$ and the bound on $\|\hat{e}^{(\mathrm{rem})}\|$ given in the Lemma we obtain

$$\hat{e}'(\beta)\hat{e}(\beta) = B(\beta) + B'(\beta) + T^{(\mathrm{rem})}(\beta) \ , \tag{A.16}$$

where the terms $B^{(\mathrm{rem},1)}(\beta)$ and $B^{(\mathrm{rem},2)}$ in $B(\beta)$ are given by

$$\begin{aligned}
B^{(\mathrm{rem},1)}(\beta) &= M_{f^0}[(\beta-\beta^0\cdot X)]' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \\
&\quad + M_{f^0} e' M_{\lambda^0} [(\beta-\beta^0\cdot X)] M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \\
&\quad + M_{f^0} e' M_{\lambda^0} e M_{f^0} [(\beta-\beta^0\cdot X)]' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \\
&\quad + M_{f^0} \left(M_{f^0} e' M_{\lambda^0} \hat{e}^{(2)}_e + \hat{e}^{(1)\prime}_e \hat{e}^{(2)}_e + \hat{e}^{(2)\prime}_e M_{\lambda^0} e' M_{f^0} \right) P_{f^0} \ , \\
B^{(\mathrm{rem},2)} &= \tfrac{1}{2} P_{f^0} \left(M_{f^0} e' M_{\lambda^0} \hat{e}^{(2)}_e + \hat{e}^{(1)\prime}_e \hat{e}^{(2)}_e + \hat{e}^{(2)\prime}_e M_{\lambda^0} e' M_{f^0} \right) P_{f^0} \\
&= f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime},
\end{aligned} \tag{A.17}$$

and for $\sqrt{N}\|\beta-\beta^0\| \le c$ (which implies $\|\hat{e}(\beta)\| = \mathcal{O}_p(\sqrt{N})$) we have

$$\|T^{(\mathrm{rem})}(\beta)\| = \mathcal{O}_p(N^{-1/2}) + \|\beta-\beta^0\| \mathcal{O}_p(N^{1/2}) + \|\beta-\beta^0\|^2 \mathcal{O}_p(N^{3/2}) \ . \tag{A.18}$$

which holds uniformly over β. Note also that

$$B^{(eeee)} + B^{(eeee)\prime} = M_{f^0} \left(M_{f^0} e' M_{\lambda^0} \hat{e}^{(2)}_e + \hat{e}^{(1)\prime}_e \hat{e}^{(2)}_e + \hat{e}^{(2)\prime}_e M_{\lambda^0} e' M_{f^0} \right) M_{f^0}. \tag{A.19}$$

Thus, we have $\|B^{(\mathrm{rem},2)}\| = \mathcal{O}_p(1)$, and for $\sqrt{N}\|\beta-\beta^0\| \le c$ we have $\|B^{(\mathrm{rem},1)}(\beta)\| = \mathcal{O}_p(1) + \|\beta-\beta^0\|\mathcal{O}_p(N)$, and by Weyl's inequality

$$\mu_t\left[\hat{e}'(\beta)\hat{e}(\beta)\right] = \mu_t\left[B(\beta) + B'(\beta)\right] + o_p\left[\left(1 + \|\beta-\beta^0\|\right)^2\right] \ , \tag{A.20}$$

again uniformly over β. This proves the Theorem. ■

Proof of Corollary 3.4. From Theorem 2.1 we know that $\sqrt{N}(\hat{\beta}_R - \beta_0) = \mathcal{O}_p(1)$, so that the bounds in Theorem 3.3 and Assumption 4 are applicable. Since $\hat{\beta}_R$ minimizes $\mathcal{L}^R_{NT}(\beta)$ it must in particular satisfy $\mathcal{L}^R_{NT}(\hat{\beta}_R) \le \mathcal{L}^R_{NT}(\beta^0)$. Applying Theorem 3.3($i$), Theorem 3.1, and Assumption 4 to this inequality gives

$$\begin{aligned}
&\left(\hat{\beta}_R - \beta^0\right)' W \left(\hat{\beta}_R - \beta^0\right) - \frac{2}{\sqrt{NT}} \left(\hat{\beta}_R - \beta^0\right)' \left(C^{(1)} + C^{(2)}\right) \\
&\quad \le \frac{1}{NT} \left\{ \sum_{t=1}^{R-R^0} \mu_r \left[\tilde{A}\left(\hat{\beta}_R\right)\right] + \mathcal{O}_p\left[\sqrt{N} + N^{5/4}\|\hat{\beta}_R - \beta^0\| + N^2\|\hat{\beta}_R - \beta^0\|/\log(N)\right] \right\}.
\end{aligned} \tag{A.21}$$

Our assumptions guarantee $C^{(2)} = \mathcal{O}_p(1)$, and we explicitly assume $C^{(1)} = \mathcal{O}_p(1)$. Furthermore, Assumption 2 guarantees that

$$\left(\hat{\beta}_R - \beta^0\right)' W \left(\hat{\beta}_R - \beta^0\right) - \frac{1}{NT} \sum_{t=1}^{R-R^0} \mu_r \left[\tilde{A}\left(\hat{\beta}_R\right)\right] \geq b_2 \|\hat{\beta}_R - \beta^0\|^2. \tag{A.22}$$

Thus we obtain

$$b_2 \left(N^{3/4}\|\hat{\beta}_R - \beta^0\|\right)^2 \leq \mathcal{O}_p(1) + \mathcal{O}_p\left(N^{3/4}\|\hat{\beta}_R - \beta^0\|\right) + o_p\left[\left(N^{3/4}\|\hat{\beta}_R - \beta^0\|\right)^2\right], \tag{A.23}$$

from which we can conclude that $N^{3/4}\|\hat{\beta}_R - \beta^0\| = \mathcal{O}_p(1)$, which proves the first part of the Theorem. ■

Proof of Corollary 3.5. Having $N^{3/4}\|\hat{\beta}_R - \beta^0\| = \mathcal{O}_p(1)$ the bound in Assumption 5 becomes applicable. We already introduced $\gamma \equiv W^{-1}\left(C^{(1)} + C^{(2)}\right)/\sqrt{NT} = \mathcal{O}_p(1/\sqrt{NT})$. Since $\hat{\beta}_R$ minimizes $\mathcal{L}^R_{NT}(\beta)$ it must in particular satisfy $\mathcal{L}^R_{NT}(\hat{\beta}_R) \leq \mathcal{L}^R_{NT}\left(\beta^0 + \gamma\right)$. Using Theorem 3.3($ii$) and Assumption 5 it follows that

$$\mathcal{L}^0_{NT}(\hat{\beta}_R) \leq \mathcal{L}^0_{NT}\left(\beta^0 + \gamma\right) + \frac{1}{NT}\, o_p\left[\left(1 + \sqrt{NT}\|\hat{\beta}_R - \beta^0\|^2\right)^2\right]. \tag{A.24}$$

The rest of the proof is analogous to the proof of corrollary 3.2. ■

The proofs for the results of Section 4 can be found in the supplementary material.

References

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

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

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

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

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

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

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

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

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

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

Chudik, A., Pesaran, M., and Tosetti, E. Weak and Strong Cross Section Dependence and Estimation of Large Panels. Manuscript.

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

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

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

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

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

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

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

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

Moon, H. and Weidner, M. (2009). Likelihood Expansion for Panel Regression Models with Factors. Manuscript.

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

Muirhead, R. (1982). *Aspects of multivariate statistical theory*. Wiley-Interscience.

Onatski, A. (2006). Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak. Manuscript.

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

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

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

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

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

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

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

S Supplementary Material

S.1 Detailed Proof of Consistency

Detailed Proof of Theorem 2.1. We first establish a lower bound on $\mathcal{L}^0_{NT}(\beta)$. Consider the the last expression for $\mathcal{L}^0_{NT}(\beta)$ in equation (2.4) and plug in $Y = \sum_k \beta^0_k X_k + \lambda^0 f^{0\prime} + e$, then replace $\lambda^0 f^{0\prime}$ by $\lambda f'$, and minimize over the $N \times R^0$ matrix λ and the $T \times R^0$ matrix f. This gives

$$
\begin{aligned}
\mathcal{L}^0_{NT}(\beta) &\geq \frac{1}{NT} \min_{\lambda,f} \sum_{t=R+1}^{T} \mu_t \left[\left(\sum_k (\beta^0_k - \beta_k) X_k + e + \lambda f' \right)' \left(\sum_k (\beta^0_k - \beta_k) X_k + e + \lambda f' \right) \right] \\
&= \frac{1}{NT} \sum_{t=R+R^0+1}^{T} \mu_t \left[\left(\sum_k (\beta^0_k - \beta_k) X_k + e \right)' \left(\sum_k (\beta^0_k - \beta_k) X_k + e \right) \right] \\
&= \frac{1}{NT} \min_{\tilde{F}} \mathrm{Tr} \left[\left(\sum_k (\beta^0_k - \beta_k) X_k + e \right) M_{\tilde{F}} \left(\sum_k (\beta^0_k - \beta_k) X_k + e \right)' \right] ,
\end{aligned}
\tag{S.1}
$$

where in the last line we minimize over all $T \times (R+R^0)$ matrices $\tilde{F}$. We now decompose this expression into a the component quadratic in $(\beta - \beta^0)$, linear in $(\beta - \beta^0)$ and independent of $(\beta - \beta^0)$. For the quadratic component we use Assumption 2(ii) to obtain

$$
\begin{aligned}
&\frac{1}{NT} \min_{\tilde{F}} \mathrm{Tr} \left[\left(\sum_k (\beta_k - \beta^0_k) X_k \right) M_{\tilde{F}} \left(\sum_k (\beta_k - \beta^0_k) X_k \right)' \right] \\
&\quad = \frac{1}{NT} \sum_{t=R+R^0+1}^{T} \mu_t \left[\left(\sum_k (\beta_k - \beta^0_k) X_k \right)' \left(\sum_k (\beta_k - \beta^0_k) X_k \right) \right] \;\geq\; b \, \|\beta - \beta^0\|^2 \, .
\end{aligned}
\tag{S.2}
$$

For the coefficient of the linear component we use assumption 1 and 2(i) to find

$$
\begin{aligned}
\left| \frac{1}{NT} \mathrm{Tr} \left(X_k \, M_{\tilde{F}} \, e' \right) \right| &\leq \left| \frac{1}{NT} \mathrm{Tr} \left(X_k \, e' \right) \right| + \left| \frac{1}{NT} \mathrm{Tr} \left(X_k \, P_{\tilde{F}} \, e' \right) \right| \\
&\leq \mathcal{O}_p \left(\frac{1}{\sqrt{NT}} \right) + \frac{R+R^0}{NT} \, \|e\| \, \|X_k\| = \mathcal{O}_p \left(\frac{1}{\sqrt{\min(N,T)}} \right) \, .
\end{aligned}
\tag{S.3}
$$

For the constant term we use Assumption 1 to obtain

$$
\begin{aligned}
\frac{1}{NT} \mathrm{Tr} \left(e \, M_{\tilde{F}} \, e' \right) &= \frac{1}{NT} \mathrm{Tr} \left(e e' \right) - \frac{1}{NT} \mathrm{Tr} \left(e \, P_{\tilde{F}} \, e' \right) \\
&= \frac{1}{NT} \mathrm{Tr} \left(e e' \right) + \mathcal{O}_p \left(\frac{1}{\min(N,T)} \right) \, ,
\end{aligned}
\tag{S.4}
$$

because $\left|\mathrm{Tr}\left(e\, P_{\tilde F}\, e'\right)\right| \leq (R+R^0)\,\|e\|^2 = \mathcal{O}_p(\max(N,T))$. Combining these results we have

$$\mathcal{L}^0_{NT}(\beta) \;\geq\; b\,\|\beta-\beta^0\|^2 + \mathcal{O}_p\left(\frac{\|\beta-\beta^0\|}{\sqrt{\min(N,T)}}\right) + \frac{1}{NT}\,\mathrm{Tr}\left(ee'\right) + \mathcal{O}_p\left(\frac{1}{\min(N,T)}\right) \;. \tag{S.5}$$

Next, we establish an upper bound on $\mathcal{L}^0_{NT}(\beta^0)$. We have

$$\begin{aligned}
\mathcal{L}^0_{NT}(\beta^0) &= \frac{1}{NT}\sum_{t=R+1}^{T}\mu_t\left[\left(\lambda^0 f^{0\prime}+e\right)'\left(\lambda^0 f^{0\prime}+e\right)\right] \\
&= \frac{1}{NT}\min_{\lambda}\sum_{t=R-R^0+1}^{T}\mu_t\left[\left(\lambda^0 f^{0\prime}+e\right)' M_\lambda\left(\lambda^0 f^{0\prime}+e\right)\right] \\
&\leq \frac{1}{NT}\sum_{t=R-R^0+1}^{T}\mu_t\left[\left(\lambda^0 f^{0\prime}+e\right)' M_{\lambda^0}\left(\lambda^0 f^{0\prime}+e\right)\right] \\
&\leq \frac{1}{NT}\,\mathrm{Tr}\left(e' M_{\lambda^0} e\right) \\
&= \frac{1}{NT}\,\mathrm{Tr}\left(ee'\right) + \mathcal{O}_p\left(\frac{1}{\min(N,T)}\right) \;.
\end{aligned} \tag{S.6}$$

To arrive at the last line we use $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$ and the same argument as in equation (S.4). Since we could choose $\beta = \beta^0$ in the minimization of β, the optimal $\hat\beta$ needs to satisfy $\mathcal{L}^0_{NT}(\hat\beta) \leq \mathcal{L}^0_{NT}(\beta^0)$. With the above results we thus find

$$b\,\|\hat\beta-\beta^0\|^2 + \mathcal{O}_p\left(\frac{\|\hat\beta-\beta^0\|}{\sqrt{\min(N,T)}}\right) + \mathcal{O}_p\left(\frac{1}{\min(N,T)}\right) \;\leq\; 0\;. \tag{S.7}$$

From this it follows that $\|\hat\beta-\beta^0\| = \mathcal{O}_p\left(\min(N,T)^{-1/2}\right)$, which is what we wanted to show.

■

S.2 Details of Likelihood Expansion

Proof of Lemma A.1.

(i,ii) We apply perturbation theory in Kato (1980). The unperturbed operator is $\mathcal{T}^{(0)} = \lambda^0 f^{0\prime} f^0 \lambda^{0\prime}$, 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 $\mathcal{T}^{(1)} = \sum_{k=0}^{K}\epsilon_k X_k f^0 \lambda^{0\prime} + \lambda^0 f^{0\prime}\sum_{k=0}^{K}\epsilon_k X_k'$, and $\mathcal{T}^{(2)} = \sum_{k_1=0}^{K}\sum_{k_2=0}^{K}\epsilon_{k_1}\epsilon_{k_2}X_{k_1}X_{k_2}'$. 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 $d = NT d^2_{\min}(\lambda^0, f^0)$. The bound (A.6) guarantees that

$$\|\mathcal{T}^{(1)} + \mathcal{T}^{(2)}\| \leq \frac{NT}{2}\, d^2_{\min}(\lambda^0, f^0)\ . \tag{S.8}$$

By Weyl's inequality we therefore find that the $N - R$ smallest eigenvalues of $\mathcal{T}$ (also counting multiplicity) are all smaller than $\frac{NT}{2} d^2_{\min}(\lambda^0, f^0)$, and they "originate" from the zero-eigenvalue of $\mathcal{T}^{(0)}$, with the power series expansion for $\mathcal{L}^0_{NT}(\beta)$ 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 $\|\mathcal{T}^{(n)}\| \leq a c^{n-1}$, $n = 1, 2, \ldots$, and it is satisfied for $a = 2\sqrt{NT} d_{\max}(\lambda^0, f^0) \sum_{k=0}^K |\epsilon_k| \|X_k\|$ and $c = \sum_{k=0}^K |\epsilon_k| \|X_k\| / \sqrt{NT} / 2 / d_{\max}(\lambda^0, f^0)$. According to equation (3.51) in Kato p.95, we therefore find that the power series for $\mathcal{L}^0_{NT}(\beta)$ and $M_{\hat{\lambda}}$ are convergent ($r_0 \geq 1$ in his notation) if $1 \leq \left(\frac{2a}{d} + c\right)^{-1}$, and this becomes exactly our condition (A.6).

When $\mathcal{L}^0_{NT}(\beta)$ is approximated up to order $G \in \mathbb{N}$, Kato's equation (3.6) at p.89 gives the following bound on the remainder

$$\begin{aligned} &\left| \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{g=2}^{G} \sum_{k_1=0}^{K} \cdots \sum_{k_g=0}^{K} \epsilon_{k_1} \ldots \epsilon_{k_g} L^{(g)}\left(\lambda^0, f^0, X_{k_1}, X_{k_2}, \ldots, X_{k_g}\right) \right| \\ &\qquad\qquad \leq \frac{(N-R)\gamma^{G+1}\, d^2_{\min}(\lambda^0, f^0)}{4(1-\gamma)}\ , \end{aligned} \tag{S.9}$$

where

$$\gamma \;=\; \frac{\sum_{k=1}^K \left|\beta^0_k - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}}}{r_0(\lambda^0, f^0)} \;<\; 1\ . \tag{S.10}$$

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

(iii,iv) We have $\|S^{(m)}\| = \left(NT d^2_{\min}(\lambda^0, f^0)\right)^{-m}$, $\|\mathcal{T}^{(1)}_k\| \leq 2\sqrt{NT} d_{\max}(\lambda^0, f^0) \|X_k\|$, and

$\|\mathcal{T}^{(2)}_{k_1 k_2}\| \leq \|X_{k_1}\| \|X_{k_2}\|$. Therefore

$$
\begin{aligned}
&\left\| S^{(m_1)}\, \mathcal{T}^{(\nu_1)}_{k_1 \ldots}\, S^{(m_2)} \,\ldots\, S^{(m_p)}\, \mathcal{T}^{(\nu_p)}_{\ldots k_g}\, S^{(m_{p+1})} \right\| \\
&\qquad \leq \left(NT d^2_{\min}(\lambda^0, f^0)\right)^{-\sum m_j} \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0)\right)^{2p-\sum \nu_j} \|X_{k_1}\| \|X_{k_2}\| \ldots \|X_{k_g}\| \, .
\end{aligned}
\tag{S.11}
$$

We have

$$
\begin{aligned}
&\sum_{\substack{\nu_1 + \ldots + \nu_p = g \\ 2 \geq \nu_j \geq 1}} 1 \leq 2^p \, , \\
&\sum_{\substack{m_1 + \ldots + m_{p+1} = p-1 \\ m_j \geq 0}} 1 \leq \sum_{\substack{m_1 + \ldots + m_{p+1} = p \\ m_j \geq 0}} 1 = \frac{(2p)!}{(p!)^2} \leq 4^p \, .
\end{aligned}
\tag{S.12}
$$

Using this we find[17]

$$
\begin{aligned}
&\left\| M^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) \right\| \\
&\qquad \leq \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0)\right)^{-g} \|X_{k_1}\| \|X_{k_2}\| \ldots \|X_{k_g}\| \sum_{p=\lceil g/2 \rceil}^{g} \left(\frac{32\, d^2_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^p \\
&\qquad \leq \frac{g}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^g \frac{\|X_{k_1}\|}{\sqrt{NT}} \frac{\|X_{k_2}\|}{\sqrt{NT}} \ldots \frac{\|X_{k_g}\|}{\sqrt{NT}} \, .
\end{aligned}
\tag{S.13}
$$

For $g \geq 3$ there always appears at least one factor $S^{(m)}$, $m \geq 1$, inside the trace of the terms that contribute to $L^{(g)}$, and we have $\operatorname{rank}(S^{(m)}) = R$ for $m \geq 1$. Using $\operatorname{Tr}(A) \leq \operatorname{rank}(A)\|A\|$, and the equations (S.11) and (S.12), we therefore find[18] for $g \geq 3$

$$
\begin{aligned}
&\frac{1}{NT} \left| L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \ldots, X_{k_g}\right) \right| \\
&\qquad \leq R\, d^2_{\min}(\lambda^0, f^0) \left(2\sqrt{NT} d_{\max}(\lambda^0, f^0)\right)^{-g} \\
&\qquad\qquad \|X_{k_1}\| \|X_{k_2}\| \ldots \|X_{k_g}\| \sum_{p=\lceil g/2 \rceil}^{g} \left(\frac{32\, d^2_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^p \\
&\qquad \leq \frac{R\, g\, d^2_{\min}(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^g \frac{\|X_{k_1}\|}{\sqrt{NT}} \frac{\|X_{k_2}\|}{\sqrt{NT}} \ldots \frac{\|X_{k_g}\|}{\sqrt{NT}} \, .
\end{aligned}
\tag{S.14}
$$

[17] 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 $\nu_j \leq 2$.

[18] 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 $\sum m_j = p - 1$ (not p as before), there is also an additional factor $NT d^2_{\min}(\lambda^0, f^0)$.

This implies for $g \geq 3$

$$\frac{1}{NT} \left| \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \dots \sum_{k_g=0}^{K} \epsilon_{k_1}\, \epsilon_{k_2}\, \dots\, \epsilon_{k_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \dots, X_{k_g}\right) \right|$$
$$\leq \frac{R\, g\, d^2_{\min}(\lambda^0, f^0)}{2} \left(\frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \right)^g \left(\sum_{k=0}^{K} \frac{\|\epsilon_k\, X_k\|}{\sqrt{NT}} \right)^g . \qquad \text{(S.15)}$$

Therefore for $G \geq 2$ we have

$$\left| \mathcal{L}^0_{NT}(\beta) - \frac{1}{NT} \sum_{g=2}^{G} \sum_{k_1=0}^{K} \dots \sum_{k_g=0}^{K} \epsilon_{k_1} \dots \epsilon_{k_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \dots, X_{k_g}\right) \right|$$
$$= \frac{1}{NT} \sum_{g=G+1}^{\infty} \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} \dots \sum_{k_g=0}^{K} \epsilon_{k_1}\, \epsilon_{k_2}\, \dots\, \epsilon_{k_g}\, L^{(g)}\left(\lambda^0,\, f^0,\, X_{k_1},\, X_{k_2}, \dots, X_{k_g}\right)$$
$$\leq \sum_{g=G+1}^{\infty} \frac{R\, g\, \alpha^g\, d^2_{\min}(\lambda^0, f^0)}{2}$$
$$\leq \frac{R\, (G+1)\, \alpha^{G+1}\, d^2_{\min}(\lambda^0, f^0)}{2(1-\alpha)^2} , \qquad \text{(S.16)}$$

where

$$\alpha = \frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \sum_{k=0}^{K} \frac{\|\epsilon_k\, X_k\|}{\sqrt{NT}}$$
$$= \frac{16\, d_{\max}(\lambda^0, f^0)}{d^2_{\min}(\lambda^0, f^0)} \left(\sum_{k=1}^{K} \left|\beta^0_k - \beta_k\right| \frac{\|X_k\|}{\sqrt{NT}} + \frac{\|e\|}{\sqrt{NT}} \right) < 1 . \qquad \text{(S.17)}$$

Using the same argument we can start from equation (S.13) to obtain the boundfor the remainder of the series expansion for $M_{\hat{\lambda}}(\beta)$.

Note that compared to the bound (S.9) on the remainder, the new bound (S.16) 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.

■

Lemma S.1. *Under the assumptions of Theorem 3.1 we have*

$$M_{\hat\lambda}(\beta) = M_{\lambda^0} + M^{(1)}_{\hat\lambda,e} + M^{(2)}_{\hat\lambda,e} - \sum_{k=1}^{K} \left(\beta_k - \beta^0_k\right) M^{(1)}_{\hat\lambda,X,k} + M^{(\mathrm{rem})}_{\hat\lambda}(\beta) \; ,$$

$$M_{\hat f}(\beta) = M_{f^0} + M^{(1)}_{\hat f,e} + M^{(2)}_{\hat f,e} - \sum_{k=1}^{K} \left(\beta_k - \beta^0_k\right) M^{(1)}_{\hat f,X,k} + M^{(\mathrm{rem})}_{\hat f}(\beta) \; ,$$

where the expansion coefficients in the expansion of $M_{\hat\lambda}(\beta)$ are $N \times N$ matrices given by

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

and analogously we have $T \times T$ matrices

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

Finally, the remainder terms of the expansions satisfy for any sequence $c_{NT} \to 0$

$$\sup_{\{\beta: \|\beta-\beta^0\| \le c_{NT}\}} \frac{\left\| M_{\hat{\lambda}}^{(\mathrm{rem})}(\beta) \right\|}{\|\beta-\beta^0\|^2 + N^{-1/2}\|\beta-\beta^0\| \ + N^{-3/2}} = \mathcal{O}_p(1) \ ,$$

$$\sup_{\{\beta: \|\beta-\beta^0\| \le c_{NT}\}} \frac{\left\| M_{\hat{f}}^{(\mathrm{rem})}(\beta) \right\|}{\|\beta-\beta^0\|^2 + N^{-1/2}\|\beta-\beta^0\| \ + N^{-3/2}} = \mathcal{O}_p(1) \ .$$

Proof. The general expansion of $M_{\hat{\lambda}}(\beta)$ is given in Lemma A.1. The present Lemma just makes this expansion explicit for the first few orders. The bound on the remainder $M_{\hat{\lambda}}^{(\mathrm{rem})}(\beta)$ is obtained from the bound (S.13) by the same logic as in the proof of Theorem 3.1. The analogous result for $M_{\hat{f}}(\beta)$ is obtained by applying the symmetry $N \leftrightarrow T$, $\lambda \leftrightarrow f$, $e \leftrightarrow e'$, $X_k \leftrightarrow X'_k$. ■

S.3 Proofs for Section 4

Lemma S.2. *Let A and B be symmetric $n \times n$ matrices, and let A be positive semi-definite. Let $\mu_1(A) \ge \mu_2(A) \ge \ldots \ge \mu_n(A) \ge 0$ be the sorted eigenvalues of A, and let $\nu_1, \nu_2, \ldots, \nu_n$ be the corresponding eigenvectors that are orthogonal and normalized such that $\|\nu_i\| = 1$ for $i = 1, \ldots, n$. Let $b = \max_{i,j=1,\ldots,n} |\nu_i' B \nu_j|$. Let r and q be positive integers with $r < q \le n$, and let $\sum_{i=q}^n b\,(\mu_r(A) - \mu_i(A))^{-1} \le 1$ be satisfied. Then we have*

$$|\mu_r(A+B) - \mu_r(A)| \ \le \ \frac{(q-1)\,b}{1 - \sum_{i=q}^n \frac{b}{\mu_r(A)-\mu_i(A)}} \tag{S.18}$$

Proof. For the eigenvalues of $A+B$ we have

$$\mu_r(A+B) = \min_{\Gamma} \max_{\{\gamma:\, \|\gamma\|=1,\, P_\Gamma \gamma = 0\}} \gamma'(A+B)\gamma \ , \tag{S.19}$$

where Γ is a $n \times (r-1)$ matrix with full rank $r-1$, and γ is a $n \times 1$ vector. In the following we only consider those γ that lie in the span of the first r eigenvectors A, i.e. $\gamma = \sum_{i=1}^r c_i \nu_i$. The condition $\|\gamma\| = 1$ implies $\sum_{i=1}^r c_i^2 = 1$. The column space of Γ is $(r-1)$-dimensional. Therefore, for a given $\gamma = \sum_{i=1}^r c_i \nu_i$ there always exists a Γ such that the conditions $\|\gamma\| = 1$ and $P_\Gamma \gamma = 0$ *uniquely* determine γ up to the sign. We thus have

$$
\begin{aligned}
\mu_r(A+B) &\geq \min_{\Gamma} \max_{\left\{\gamma:\, \gamma=\sum_{i=1}^r c_i \nu_i,\, \|\gamma\|=1,\, P_\Gamma\gamma=0\right\}} \gamma'(A+B)\gamma \\
&= \min_{\left\{\gamma:\, \gamma=\sum_{i=1}^r c_i \nu_i,\, \|\gamma\|=1\right\}} \gamma'(A+B)\gamma \\
&\geq \min_{\left\{(c_1,\ldots,c_r):\, \sum_{i=1}^r c_i^2=1\right\}} \left[\sum_{i=1}^r c_i^2\, \mu_i(A) - b \left(\sum_{i=1}^r |c_i| \right)^2 \right] \\
&\geq \mu_r(A) - r\, b \\
&\geq \mu_r(A) - \frac{(q-1)\, b}{1 - \sum_{i=q}^n \frac{b}{\mu_r(A)-\mu_i(A)}} \,,
\end{aligned}
\tag{S.20}
$$

where we used that $q-1 \geq r$ and that the additional fraction we multiplied with is larger than one. This is the lower bound for $\mu_r(A+B)$ that we wanted to show. We now want to derive the upper bound. Let $\widetilde{A}$, $\widetilde{B}$ and $\bar{B}$ be $(n-r+1)\times(n-r+1)$ matrices defined by $\widetilde{A}_{ij} = \nu'_{i+r-1} A \nu_{j+r-1}$, $\widetilde{B}_{ij} = \nu'_{i+r-1} B \nu_{j+r-1}$, and $\bar{B}_{ij} = b$, where $i,j = 1,\ldots,n-r+1$. We can choose $\Gamma = (\nu_1, \nu_2, \ldots, \nu_{r-1})$ in the above minimization problem, in which case γ is restricted to the span of ν_r, $\nu_{r+1}, \ldots, \nu_n$. Therefore

$$
\begin{aligned}
\mu_r(A+B) &\leq \max_{\{\widetilde{\gamma}:\, \|\widetilde{\gamma}\|=1\}} \widetilde{\gamma}'(\widetilde{A}+\widetilde{B})\widetilde{\gamma} \\
&= \mu_1(\widetilde{A}+\widetilde{B}) \,,
\end{aligned}
\tag{S.21}
$$

where $\widetilde{\gamma}$ is a $(n-r+1)$-dimensional vector, whose components are denoted $\widetilde{\gamma}_i$, $i = 1,\ldots,n-r+1$, in the following. Note that $\widetilde{A}$ is a diagonal matrix with entries $\mu_{i+r-1}(A)$, $i = 1,\ldots,n-r+1$. Therefore

$$
\begin{aligned}
\mu_r(A+B) &\leq \max_{\{\widetilde{\gamma}:\, \|\widetilde{\gamma}\|=1\}} \left[\sum_{i=1}^{n+r-1} (\widetilde{\gamma}_i)^2 \mu_{i+r-1}(A) + \sum_{i,j=1}^{n+r-1} \widetilde{\gamma}_i\, \widetilde{\gamma}_j\, \widetilde{B}_{ij} \right] \\
&\leq \max_{\{\widetilde{\gamma}:\, \|\widetilde{\gamma}\|=1\}} \left[\sum_{i=1}^{n+r-1} (\widetilde{\gamma}_i)^2 \mu_{i+r-1}(A) + b \sum_{i,j=1}^{n+r-1} |\widetilde{\gamma}_i|\, |\widetilde{\gamma}_j| \right] \\
&= \max_{\{\widetilde{\gamma}:\, \|\widetilde{\gamma}\|=1\}} \left[\sum_{i=1}^{n+r-1} (\widetilde{\gamma}_i)^2 \mu_{i+r-1}(A) + \sum_{i,j=1}^{n+r-1} \widetilde{\gamma}_i\, \widetilde{\gamma}_j\, \bar{B}_{ij} \right] \\
&= \mu_1(\tilde{A}+\bar{B}) \,.
\end{aligned}
\tag{S.22}
$$

In the last maximization problem the maximum is always attained at a point with $\widetilde{\gamma}_i \geq 0$, which is why we could omit the absolute values around $\widetilde{\gamma}_i$.

The eigenvalue $\widetilde{\mu} \equiv \mu_1(\widetilde{A}+\bar{B})$ is a solution of the characteristic polynomial of $\widetilde{A}+\bar{B}$

which can be written as

$$1 = \sum_{i=r}^{n} \frac{b}{\widetilde{\mu} - \mu_i(A)} \ , \tag{S.23}$$

where $\mu_i(A) = \mu_{i-r+1}(\widetilde{A})$ are the eigenvalues of $\widetilde{A}$. In addition we have $\widetilde{\mu} = \mu_1(\widetilde{A} + \bar{B}) > \mu_1(\widetilde{A}) = \mu_r(A)$, because $\bar{B}$ is positive semi-definite (which gives $\geq$) and the eigenvectors of $\widetilde{A}$ do not agree with those of $\bar{B}$ (which gives $\neq$). From the characteristic polynomial we therefore find

$$\begin{aligned} 1 &= \sum_{i=r}^{q-1} \frac{b}{\widetilde{\mu} - \mu_i(A)} + \sum_{i=q}^{n} \frac{b}{\widetilde{\mu} - \mu_i(A)} \\ &\leq \frac{b(q-1)}{\widetilde{\mu} - \mu_r(A)} + \sum_{i=q}^{n} \frac{b}{\mu_r(A) - \mu_i(A)} \end{aligned} \tag{S.24}$$

Since we assume $1 \geq \sum_{i=q}^{n} \frac{b}{\mu_r(A)-\mu_i(A)}$, this gives an upper bound on $\widetilde{\mu}$, and since $\mu_r(A + B) \leq \widetilde{\mu}$ the same bound holds for $\mu_r(A+B)$, namely

$$\mu_r(A+B) \leq \mu_r(A) + \frac{(q-1)\, b}{1 - \sum_{i=q}^{n} \frac{b}{\mu_r(A)-\mu_i(A)}} \ . \tag{S.25}$$

This is what we wanted to show. ■

Proof of Theorem 4.1. Define

$$\begin{aligned} C^{\pm}(\beta) = B(\beta) + B'(\beta) &\pm \left(\sqrt{\frac{4}{aN}} M_{f^0} B^{(\text{rem},1)}(\beta) P_{f^0} \mp \sqrt{\frac{aN}{4}} P_{f^0} \right) \\ &\times \left(\sqrt{\frac{4}{aN}} M_{f^0} B^{(\text{rem},1)}(\beta) P_{f^0} \mp \sqrt{\frac{aN}{4}} P_{f^0} \right)' \\ &\pm \left(\sqrt{\frac{4}{aN}} M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \pm \sqrt{\frac{aN}{4}} P_{f^0} \right) \\ &\times \left(\sqrt{\frac{4}{aN}} M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} f^{0\prime} \pm \sqrt{\frac{aN}{4}} P_{f^0} \right)' . \end{aligned} \tag{S.26}$$

Since $C^{+}(\beta)$ (or $C^{-}(\beta)$) is obtained be adding (or subtracting) a positive definite matrix from $B(\beta) + B'(\beta)$, we have

$$\mu_t\left(C^{-}(\beta)\right) \leq \mu_t\left(B(\beta) + B'(\beta)\right) \leq \mu_t\left(C^{+}(\beta)\right) . \tag{S.27}$$

The advantage of considering $C^{\pm}(\beta)$ instead of $B(\beta) + B'(\beta)$ directly is that there are no "mixed terms" in $C^{\pm}(\beta)$, which start with M_{f^0} and end with P_{f^0}, or vice versa, i.e. we can write $C^{\pm}(\beta) = C_1^{\pm}(\beta) + C_2^{\pm}$, where $C_1^{\pm}(\beta) = M_{f^0} C_1^{\pm}(\beta) M_{f^0}$ and $C_2^{\pm} = P_{f^0} C_2^{\pm} P_{f^0}$. Concretely, we have

$$\begin{aligned}
C_1^{\pm}(\beta) = A(\beta) &\pm \frac{4}{aN} M_{f^0} B^{(\mathrm{rem},1)}(\beta) P_{f^0} B^{(\mathrm{rem},1)\prime}(\beta) M_{f^0} \\
&\pm \frac{4}{aN} M_{f^0} e' M_{\lambda^0} e M_{f^0} e' \lambda^0 (\lambda^{0\prime}\lambda^0)^{-1} (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} e' M_{\lambda^0} e M_{f^0} \\
&+ M_{f^0} \left[(\beta - \beta^0) \cdot X - e \right]' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
&+ M_{f^0} e' M_{\lambda^0} \left[(\beta - \beta^0) \cdot X \right] f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} e M_{f^0} \\
&+ M_{f^0} e' M_{\lambda^0} e f^0 (f^{0\prime} f^0)^{-1} (\lambda^{0\prime}\lambda^0)^{-1} \lambda^{0\prime} \left[(\beta - \beta^0) \cdot X \right] M_{f^0} \\
&+ \text{the last three lines transposed} + B^{(eeee)} + B^{(eeee)\prime}, \\
C_2^{\pm} = P_{f^0} & B^{(\mathrm{rem},2)} P_{f^0} + P_{f^0} B^{(\mathrm{rem},2)\prime} P_{f^0} \pm \frac{aN}{2} P_{f^0}.
\end{aligned} \tag{S.28}$$

In the rest of the proof we always assume that $N^{3/4} \left\| \beta - \beta^0 \right\| \leq c$. We apply Lemma S.2 to $C_1^{\pm}(\beta)$, with the A in the lemma equal to the leading term $M_{f^0} e' M_{\lambda^0} e M_{f^0}$, the B in the lemma equal to the remainder of $C_1^{\pm}(\beta)$, and $q = q_{NT}$. If we can show that

$$\sum_{\tau = q_{NT}}^{T-R^0} \frac{b_{NT}}{\rho_{R-R^0} - \rho_\tau} = o_p(1), \tag{S.29}$$

then the lemma becomes applicable asymptotically, and for $t = 1, \ldots, R - R^0$ we have wpa1

$$\left| \mu_t \left(C_1^{\pm}(\beta) \right) - \rho_t \right| \leq \frac{(q_{NT} - 1)\, b_{NT}}{1 - \sum_{\tau=q_{NT}}^{T-R^0} \frac{b_{NT}}{\rho_t - \rho_\tau}} \leq \frac{q_{NT}\, b_{NT}}{1 - \sum_{\tau=q_{NT}}^{T-R^0} \frac{b_{NT}}{\rho_{R-R^0} - \rho_\tau}}\,, \tag{S.30}$$

where

$$b_{NT} = \max_{t,\tau = 1, \ldots, T-R^0} w_t' \left(C_1^{\pm}(\beta) - M_{f^0} e' M_{\lambda^0} e M_{f^0} \right) w_\tau. \tag{S.31}$$

We now check how the different terms in $C_1^{\pm}(\beta) - M_{f^0} e' M_{\lambda^0} e M_{f^0}$ contribute to b_{NT}. We have

$$\begin{aligned}
\max_{t,\tau}\left|w_t' M_{f^0} e' M_{\lambda^0}[(\beta-\beta^0)\cdot X] M_{f^0} w_\tau\right| &\le K\|e\|\|\beta-\beta^0\| \max_{k,i,\tau}\|v_i' X_k w_\tau\| \\
&\le d_{NT}\mathcal{O}_p(N^{-1/4}), \\
\max_{t,\tau}\left|w_t' M_{f^0}[(\beta-\beta^0)\cdot X]' M_{\lambda^0}[(\beta-\beta^0)\cdot X] M_{f^0} w_\tau\right| &\le K^2\left\|\beta-\beta^0\right\|^2 \max_{k,t}\|M_{\lambda^0} X_k w_t\|^2 \\
&\le K^2 N\left\|\beta-\beta^0\right\|^2 \max_{k,i,t}\|v_i' X_k w_t\|^2 \\
&\le d_{NT}^2\mathcal{O}_p(N^{-1/2}), \\
\left|w_t' \frac{4}{aN} M_{f^0} B^{(\mathrm{rem},1)}(\beta) P_{f^0} B^{(\mathrm{rem},1)\prime}(\beta) M_{f^0} w_\tau\right| &\le \frac{4}{aN}\|B^{(\mathrm{rem},1)}(\beta)\|^2 = \mathcal{O}_p(N^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
&\max_{t,\tau}\left|w_t' \frac{4}{aN} M_{f^0} e' 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} e M_{f^0} e' M_{\lambda^0} e M_{f^0} w_\tau\right| \\
&\quad\le \frac{4}{aN}\|e\|^4\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\| \max_t\|w_t' e' P_{\lambda^0}\|^2 \le d_{NT}\mathcal{O}_p(N^{-1}), \\
&\max_{t,\tau}\left|w_t' M_{f^0} e' 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} w_\tau\right| \\
&\quad\le \|e\|\left\|f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\right\| \max_i\|v_i' e P_{f^0}\| \max_t\|w_t' e' P_{\lambda^0}\| \le d_{NT}^2\mathcal{O}_p(N^{-1/2}), \\
&\max_{t,\tau}\left|w_t' M_{f^0}\left[(\beta-\beta^0)\cdot X\right]' 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} w_\tau\right| \\
&\quad= \max_{t,\tau}\left|w_t'\left[(\beta-\beta^0)\cdot X\right]'\left(\sum_i v_i' v_i\right) e f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime} e w_\tau\right| \\
&\quad\le K\|\beta-\beta^0\| N \max_{i,t,k}|v_i' X_k w_t| \max_i\|v_i' e P_{f^0}\| \max_t\|w_t' e' P_{\lambda^0}\|\left\|f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\right\| \\
&\quad\le d_{NT}^3\mathcal{O}_p(N^{-3/4}), \\
&\max_{t,\tau}\left|w_t' M_{f^0} e' M_{\lambda^0}\left[(\beta-\beta^0)\cdot X\right] f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime} e M_{f^0} w_\tau\right| \\
&\quad\le K\|e\|\|\beta-\beta^0\|\left\|f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\right\| \max_{i,k}\|v_i' X_k P_{f^0}\| \max_t\|w_t' e' P_{\lambda^0}\| \\
&\quad\le d_{NT}^2\mathcal{O}_p(N^{-1/2}), \\
&\max_{t,\tau}\left|w_t' M_{f^0} e' M_{\lambda^0} e f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\left[(\beta-\beta^0)\cdot X\right] M_{f^0} w_\tau\right| \\
&\quad\le K\|e\|\|\beta-\beta^0\|\left\|f^0(f^{0\prime}f^0)^{-1}(\lambda^{0\prime}\lambda^0)^{-1}\lambda^{0\prime}\right\| \max_{i,k}\|v_i' e P_{f^0}\| \max_t\|w_t' X_k' P_{\lambda^0}\| \\
&\quad\le d_{NT}^2\mathcal{O}_p(N^{-1/2}).
\end{aligned}$$

and analogously one can check that

$$\max_{t,\tau}\left|w_t' B^{(eeee)} w_\tau\right| \le d_{NT}^2\mathcal{O}_p(N^{-1}) + d_{NT}^3\mathcal{O}_p(N^{-3/2}). \tag{S.32}$$

All in all, we thus have

$$\begin{aligned} b_{NT} &\leq \mathcal{O}_p(N^{-1/2}) + d_{NT}\mathcal{O}_p(N^{-1/4}) + d_{NT}^2\mathcal{O}_p(N^{-1/2}) + d_{NT}^3\mathcal{O}_p(N^{-3/4}) \\ &\leq d_{NT}\mathcal{O}_p(N^{-1/4}) \ , \end{aligned} \tag{S.33}$$

where in the last we used that by assumption $d_{NT} \geq 1$ and $d_{NT} = o_p(N^{1/4})$. Therefore

$$\sum_{\tau=q_{NT}}^{T-R^0} \frac{b_{NT}}{\rho_{R-R^0} - \rho_\tau} = q_{NT} d_{NT} \mathcal{O}_p(N^{-1/4}) \frac{1}{q_{NT}} \leq \sum_{\tau=q_{NT}}^{T-R^0} \frac{1}{\rho_{R-R^0} - \rho_\tau} = o_p(1), \tag{S.34}$$

so that Lemma S.2 is indeed applicable asymptotically, and we find

$$\left|\mu_t\left(C_1^{\pm}(\beta)\right) - \rho_t\right| \leq \frac{q_{NT}\, b_{NT}}{1 - o_p(1)} \leq q_{NT}\, d_{NT}\, \mathcal{O}_p(N^{-1/4}) = o_p(1) \ . \tag{S.35}$$

For $t = 1, \ldots, R - R^0$ we thus have

$$\mu_t\left(C_1^{\pm}(\beta)\right) = \rho_t + o_p(1) \geq \rho_{R-R^0} + o_p(1) \geq \|C_2^{\pm}\|, \quad \text{wpa1}, \tag{S.36}$$

where the last step follows because $\|C_2^{\pm}\| = aN/2 + \mathcal{O}_p(1)$ and we assumed $\rho_{R-R^0} > aN$, wpa1. Since $C^{\pm}(\beta)$ is block-diagonal with blocks $C_1^{\pm}(\beta)$ and $C_2^{\pm}$ (in the basis defined by f^0), and $\mu_t\left(C_1^{\pm}(\beta)\right) \geq \|C_2^{\pm}\|$, it must be the case that wpa1 the largest $R - R^0$ eigenvalues of $C^{\pm}(\beta)$ are those of $C_1^{\pm}(\beta)$. Thus,

$$\left|\mu_t\left(C^{\pm}(\beta)\right) - \rho_t\right| = o_p(1) \ , \tag{S.37}$$

and also

$$\left|\mu_t\left(B(\beta) + B'(\beta)\right) - \rho_t\right| = o_p(1) \ , \tag{S.38}$$

which holds uniformly over all β with $N^{3/4}\left\|\beta - \beta^0\right\| \leq c$. This concludes the proof. ■

Lemma S.3. *Let g be an $N \times Q$ matrix and h be a $T \times Q$ matrix such that $g'g = h'h = \mathbb{1}_Q$. Let U be an $N \times T$ matrix whose entries U_{it} are distributed independently of g and h, and are iid $\mathcal{N}(0, \sigma^2)$. Let R be a positive integer. For asymptotics where Q and R are constant and $N, T \to \infty$ such that $N/T \to \kappa^2$, $0 < \kappa < \infty$, we then have*

$$\sup_{C \in \mathbb{R}^{Q\times Q}} \frac{\sum_{t=1}^{R} \mu_t\left[(U + gCh')'\,(U + gCh')\right] - \sum_{t=1}^{\min(Q,R)} \mu_t\left(C'C\right) - T\sigma^2(1+\kappa)^2}{\sqrt{N} + \|C\|} \leq \mathcal{O}_p\left(1\right).$$

Proof. Let u be a $T \times Q$ matrix with iid normal entries of mean zero and variance σ^2, independent of U. We decompose

$$\left(U + gCh'\right)' \left(U + gCh'\right) = A_1 + A_2 + A_3(C) \,, \tag{S.39}$$

where

$$\begin{aligned} A_1 &= U'U + hu'uh' \,, \\ A_2 &= T\sigma^2 P_{(h,\,U'g)} - U'gg'U - hu'uh' \,. \\ A_3(C) &= \left(U + gCh'\right)' gg' \left(U + gCh'\right) - T\sigma^2 P_{(h,\,U'g)} \,, \end{aligned} \tag{S.40}$$

We then have

$$\sum_{t=1}^{R} \mu_t \left[\left(U + gCh'\right)' \left(U + gCh'\right)\right] \leq \sum_{t=1}^{R} \mu_t(A_1) + \sum_{t=1}^{R} \mu_t(A_2) + \sum_{t=1}^{R} \mu_t \left[A_3(C)\right] . \tag{S.41}$$

Thus, the theorem is proven if we can show that

$$\sum_{t=1}^{R} \mu_t(A_1) = T\sigma^2(1+\kappa)^2 + \mathcal{O}_p(N^{1/3}) \,, \qquad \sum_{t=1}^{R} \mu_t(A_2) = \mathcal{O}_p(\sqrt{N}) \,, \tag{S.42}$$

and

$$\sup_{C \in \mathbb{R}^{Q\times Q}} \frac{\sum_{t=1}^{R} \mu_t \left[A_3(C)\right] - \sum_{t=1}^{\min(Q,R)} \mu_t \left(C'C\right)}{\sqrt{N} + \|C\|} \leq \mathcal{O}_p\left(1\right). \tag{S.43}$$

If h would be distributed according to the Haar measure on the Stiefel manifold defined by $h'h = \mathbb{1}_Q$, then A_1 would have Wishart distribution $W_T(N+Q, \mathbb{1}_T)$, see e.g. Muirhead (1982). Since the distribution of U is rotationally invariant, the choice of h does not matter for the probability distribution of the eigenvalues of A_1. Thus, the eigenvalues of A_1 have the same joint distribution as the eigenvalues of the Wishart distribution $W_T(N+Q, \mathbb{1}_T)$. Using Theorem 1 in Soshnikov (2002) we can thus conclude that $\mu_t(A_1) = T\sigma^2(1+\kappa)^2 + \mathcal{O}_p(N^{1/3})$, which proves the first part of (S.42).

We have $h'h = \mathbb{1}_Q$ and thus $\|h\| = \mathcal{O}_p(T^{-1/2}) = \mathcal{O}_p(N^{-1/2})$, and it is straightforward to show that $g'UU'g = T\sigma^2 \mathbb{1}_Q + \mathcal{O}_p(\sqrt{N})$, $gUh' = \mathcal{O}_p(1)$ and $u'u = T\sigma^2 \mathbb{1}_Q + \mathcal{O}_p(\sqrt{N})$.

Therefore

$$
\begin{aligned}
P_{(h,\,U'g)} &= (h,\,U'g)\left[(h,\,U'g)'(h,\,U'g)\right]^{-1}(h,\,U'g)' \\
&= (h,\,U'g)\begin{pmatrix} \mathbb{1}_Q & \mathcal{O}_p(1) \\ \mathcal{O}_p(1) & T\sigma^2\mathbb{1}_Q + \mathcal{O}_p(\sqrt{N}) \end{pmatrix}^{-1}(h,\,U'g)' \\
&= (h,\,U'g)\begin{pmatrix} \mathbb{1}_Q & \mathcal{O}_p(N^{-1}) \\ \mathcal{O}_p(N^{-1}) & T^{-1}\sigma^{-2}\mathbb{1}_Q + \mathcal{O}_p(N^{-1/2})) \end{pmatrix}(h,\,U'g)' \\
&= hh' + T^{-1}\sigma^{-2}U'gg'U + r_1 \\
&= T^{-1}\sigma^{-2}hu'uh' + T^{-1}\sigma^{-2}U'gg'U + r_2\,, \qquad \text{(S.44)}
\end{aligned}
$$

where $\|r_1\| = \mathcal{O}_p(N^{-1/2})$ and $\|r_2\| = \mathcal{O}_p(N^{-1/2})$. Since $A_2 = T\sigma^2 r_2$ we have $\|A_2\| = \mathcal{O}_p(\sqrt{N})$, from which the second part of (S.42) follows.

Finally, we consider $A_3(C)$. We have

$$
A_3(C) = P_{(h,\,U'g)}\left[\left(U+gCh'\right)'gg'\left(U+gCh'\right) - T\sigma^2\mathbb{1}_T\right]P_{(h,\,U'g)} \qquad \text{(S.45)}
$$

This shows that $A_3(C)$ has $T-2Q$ zero-eigenvalues, Q eigenvalues $-T\sigma^2$ and the remaining eigenvalues of $A_3(C)$ are equal to the Q non-zero eigenvalues of $(U+gCh')'\,gg'\,(U+gCh')$ minus $T\sigma^2$. The non-zero eigenvalues of $(U+gCh')'\,gg'\,(U+gCh')$ are identical to the eigenvalues of $g'\,(U+gCh')\,(U+gCh')'\,g$. Therefore

$$
\begin{aligned}
&\mu_t\left[A_3(C))\right] \le \max\left\{0,\ \mu_t\left[g'\left(U+gCh'\right)\left(U+gCh'\right)'g\right] - T\sigma^2\right\}\,, \quad \text{for } t \le Q, \\
&\mu_t\left[A_3(C))\right] \le 0\,, \quad \text{for } t > Q. \qquad \text{(S.46)}
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
&\sum_{t=1}^{\min(Q,R)} \mu_t\left[g'\left(U+gCh'\right)\left(U+gCh'\right)'g\right] \\
&\quad \le \sum_{t=1}^{\min(Q,R)} \mu_t\left(g'UU'g\right) + \sum_{t=1}^{\min(Q,R)} \mu_t\left(CC'\right) + \sum_{t=1}^{\min(Q,R)} \mu_t\left(g'UhC' + Ch'U'g\right) \\
&\quad \le \min(Q,R)\,T\,\sigma^2 + \mathcal{O}_p(\sqrt{N}) + \sum_{t=1}^{\min(Q,R)} \mu_t\left(C'C\right) + \min(Q,R)\,\|g'Uh\|\,\|C\|. \qquad \text{(S.47)}
\end{aligned}
$$

Since $\mu_t\,(C'C) \geq 0$ we can thus conclude that

$$\sum_{t=1}^{R} \mu_t\left[A_3(C))\right] \leq \sum_{t=1}^{\min(Q,R)} \mu_t\left(C'C\right) + \mathcal{O}_p(\sqrt{N}) + \min(Q,R)\,\|g'Uh\|\,\|C\|\,. \tag{S.48}$$

Since $g'Uh = \mathcal{O}_p(1)$ it follows that (S.43) holds. This concludes the proof. ■

Proof of Theorem 4.2. We have to show that the Assumption 1, 2, 4 and 5 as well as $C^{(1)} = \mathcal{O}_p(1)$ are satisfied.

• *Proof for Assumption 1 and 2, and for $C^{(1)} = \mathcal{O}_p(1)$.* Theorem 2.1 does not require $N/T \to \kappa^2$. For consistency reasons, we therefore want to show that Assumption 1 and 2 are satisfied in an arbitrary limit $N, T \to \infty$, although this is not explicitly stated in Theorem 4.2. By assumption, the errors e_{it} are iid $\mathcal{N}(0,\sigma^2)$. Since an arbitrary limit $N, T \to \infty$ is not considered very often in Random Matrix Theory, we define the $\max(N,T) \times \max(N,T)$ matrix e^{big}, which contains e as a submatrix, and whose remaining elements are also iid $\mathcal{N}(0,\sigma^2)$ and independent of e. We then have $\|e\| \leq \|e^{\mathrm{big}}\| = \mathcal{O}_p(\sqrt{\max(N,T)})$, where the last step is due to Geman (1980). As discussed in the main text, neither normality nor homoscedasticity nor independence of e_{it} are actually required to conclude $\|e\| = \mathcal{O}_p(\sqrt{\max(N,T)})$. For the spectral norm of X_k we have $\|X_k\| \leq \|\overline{X}_k\| + \|\tilde{X}_k\|$ and therefore $\|X_k\| = \mathcal{O}_p(\sqrt{NT})$, which concludes the proof of Assumption 1. Using the matrix norm inequality $\|A\|_{HS} \leq \sqrt{\mathrm{rank}(A)}\|A\|$ we find

$$\|X_k\|_{HS} \leq \sqrt{Q_k}\|\overline{X}_k\| + \|\tilde{X}_k\|_{HS} = \mathcal{O}_p(\sqrt{NT})\,. \tag{S.49}$$

Therefore, $\mathbb{E}\left[\mathrm{Tr}(X_k e')^2 | X_k\right] = \sigma^2 \|X_k\|_{HS}^2 = \mathcal{O}(NT)$, and thus $(NT)^{-1/2}\mathrm{Tr}(X_k e') = \mathcal{O}_p(1)$, which is part (i) of Assumption 2. Analogously we obtain $C^{(1)} = \mathcal{O}_p(1)$ from $\|M_{\lambda^0} X_k M_{f^0}\|_{HS}^2 \leq \|X_k\|_{HS}^2 = \mathcal{O}(NT)$.

To prove the second part of Assumption 2, define the $N \times KT$ matrix $\overline{X} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k)$. Note that $\mathrm{rank}(\overline{X}) \leq \sum_k Q_k$, i.e. the projector $M_{\overline{X}}$ projects out at most $\sum_k Q_k$ dimensions of the N dimensional Euclidian space. For $\|\alpha\| = 1$ we have

$$
\begin{aligned}
&(NT)^{-1} \sum_{t=R+R^0+1}^{T} \mu_t \left(X'_\alpha X_\alpha\right) \geq (NT)^{-1} \sum_{t=R+R^0+1}^{T} \mu_t \left(X'_\alpha M_{\overline{X}} X_\alpha\right) \\
&= (NT)^{-1} \sum_{t=R+R^0+1}^{T} \mu_t \left(\sum_{k_1,k_2=1}^{K} \alpha_{k_1}\alpha_{k_2} \tilde{X}'_{k_1} M_{\overline{X}} \tilde{X}_{k_2} \right) \\
&\geq (NT)^{-1} \mathrm{Tr} \left(\sum_{k_1,k_2=1}^{K} \alpha_{k_1}\alpha_{k_2} \tilde{X}'_{k_1} M_{\overline{X}} \tilde{X}_{k_2} \right) - (NT)^{-1}(R+R^0)K^2 \|\tilde{X}_k\|^2 \\
&= (NT)^{-1} \sum_{k_1,k_2=1}^{K} \alpha_{k_1}\alpha_{k_2} \mathrm{Tr} \left(\tilde{X}'_{k_1} M_{\overline{X}} \tilde{X}_{k_2} \right) - (NT)^{-1}(R+R^0)K^2 \|\tilde{X}_k\|^2 \\
&\geq (NT)^{-1} \sum_{k_1,k_2=1}^{K} \alpha_{k_1}\alpha_{k_2} \mathrm{Tr} \left(\tilde{X}'_{k_1} M_{\overline{X}} \tilde{X}_{k_2} \right) - (NT)^{-1} \left(R+R^0+\sum_k Q_k \right) K^2 \|\tilde{X}_k\|^2 \\
&= \alpha' \tilde{W} \alpha + \mathcal{O}_p\left(N^{3/2}/NT\right) \geq b_1 + o_p(1) \; . \qquad \text{(S.50)}
\end{aligned}
$$

Thus, Assumption $2(ii)$ is satisfied for any b with $b_1 > b > 0$.

- *Proof for Assumption 4.* Using Weyl's inequality we find

$$
\begin{aligned}
\sum_{t=1}^{R-R^0} \mu_t \left[A(\beta)\right] &= \sum_{t=1}^{R-R^0} \mu_t \left\{ M_{f^0} \left[e-(\beta-\beta^0)\cdot X\right]' M_{\lambda^0} \left[e-(\beta-\beta^0)\cdot X\right] M_{f^0} \right\} \\
&= \sum_{t=1}^{R-R^0} \mu_t \Big\{ M_{f^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right]' M_{\lambda^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right] M_{f^0} \\
&\qquad\qquad - M_{f^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right]' M_{\lambda^0} \left[(\beta-\beta^0)\cdot \tilde{X}\right] M_{f^0} \\
&\qquad\qquad - M_{f^0} \left[(\beta-\beta^0)\cdot \tilde{X}\right]' M_{\lambda^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right] M_{f^0} \\
&\qquad\qquad + M_{f^0} \left[(\beta-\beta^0)\cdot \tilde{X}\right]' M_{\lambda^0} \left[(\beta-\beta^0)\cdot \tilde{X}\right] M_{f^0} \Big\} \\
&\leq \sum_{t=1}^{R-R^0} \mu_t \left\{ M_{f^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right]' M_{\lambda^0} \left[e-(\beta-\beta^0)\cdot \overline{X}\right] M_{f^0} \right\} \\
&\quad + 2(R-R^0) \left(\|e\| + K\|\beta-\beta^0\| \max_k \|\overline{X}_k\| \right) \|\beta-\beta^0\| \max_k \|\tilde{X}_k\| \\
&\quad + (R-R^0)K^2 \|\beta-\beta^0\|^2 \max_k \|\tilde{X}_k\|^2 \\
= \sum_{t=1}^{R-R^0} \mu_t &\left\{ \left[M_{\lambda^0} e M_{f^0} - (\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right]' \left[M_{\lambda^0} e M_{f^0} - (\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right] \right\} \\
&\quad + \|\beta-\beta^0\| \, \mathcal{O}_p(N^{5/4}) + \|\beta-\beta^0\|^2 \, \mathcal{O}_p(N^{7/4}). \qquad \text{(S.51)}
\end{aligned}
$$

In a basis orthogonal to λ^0 and f^0 the matrix $M_{\lambda^0} e M_{f^0}$ is a $(N-R^0)\times(T-R^0)$ matrix with iid normal entries, so that Lemma S.3 becomes applicable, and we have

$$
\begin{aligned}
\sum_{t=1}^{R-R^0} \mu_t \left[A(\beta)\right] \leq (T-R^0)\sigma^2(1+\kappa)^2 + & \\
& + \sum_{t=1}^{R-R^0} \left\{ \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right]' \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right] \right\} \\
& + \mathcal{O}_p(\sqrt{N}) + \mathcal{O}_p\left(\|(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\|\right) \\
& + \|\beta-\beta^0\|\,\mathcal{O}_p(N^{5/4}) + \|\beta-\beta^0\|^2\,\mathcal{O}_p(N^{7/4}) \\
= T\sigma^2(1+\kappa)^2 & \\
& + \sum_{t=1}^{R-R^0} \left\{ \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right]' \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right] \right\} \\
& + \mathcal{O}_p(\sqrt{N}) + \|\beta-\beta^0\|\,\mathcal{O}_p(N^{5/4}) + \|\beta-\beta^0\|^2\,\mathcal{O}_p(N^{7/4}). \qquad \text{(S.52)}
\end{aligned}
$$

uniformly over β. Setting $C=0$ we can also conclude from Lemma S.3 that

$$
\sum_{t=1}^{R-R^0} \mu_t \left[A(\beta^0)\right] = T\sigma^2(1+\kappa)^2 + \mathcal{O}_p(\sqrt{N}). \qquad \text{(S.53)}
$$

Finally, similarly to (S.51) we find that

$$
\begin{aligned}
\sum_{t=1}^{R-R^0} \mu_t \left[\tilde{A}(\beta)\right] = & \sum_{t=1}^{R-R^0} \left\{ \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right]' \left[(\beta-\beta^0)\cdot M_{\lambda^0}\overline{X}M_{f^0}\right] \right\} \\
& + \|\beta-\beta^0\|\,\mathcal{O}_p(N^{5/4}) + \|\beta-\beta^0\|^2\,\mathcal{O}_p(N^{7/4}). \qquad \text{(S.54)}
\end{aligned}
$$

Combining the last three results we find that

$$
\begin{aligned}
& \sum_{t=1}^{R-R^0} \left\{ \mu_t\left[A(\beta)\right] - \mu_t\left[A(\beta^0)\right] - \mu_t\left[\tilde{A}(\beta)\right] \right\} \\
& \qquad \leq \mathcal{O}_p(\sqrt{N}) + \|\beta-\beta^0\|\,\mathcal{O}_p(N^{5/4}) + \|\beta-\beta^0\|^2\,\mathcal{O}_p(N^{7/4}), \qquad \text{(S.55)}
\end{aligned}
$$

which shows that Assumption 4 holds.

• *Proof for Assumption 5.* We want to show that the assumptions of Theorem 4.1 are satisfied with $q_{NT} = \log(N) N^{1/6}$. First, we want to show that $d_{NT} q_{NT} = o_p(N^{1/4})$, i.e. $d_{NT} = o_p(N^{1/12}/\log(N))$. Let $\tilde{v}$ be an N-vector with $iid\mathcal{N}(0,1)$ entries, and let $\tilde{w}$ be an T-vector, independent of $\tilde{v}$, also $iid\mathcal{N}(0,1)$. For all $i=1,\ldots,N-R^0$ and $t=1,\ldots,T-R^0$

we then have

$$v_i \underset{d}{=} \frac{M_{\lambda^0}\tilde{v}}{\|M_{\lambda^0}\tilde{v}\|} , \qquad\qquad w_t \underset{d}{=} \frac{M_{f^0}\tilde{w}}{\|M_{f^0}\tilde{w}\|} . \tag{S.56}$$

This is true, because e has $iid\mathcal{N}(0,\sigma^2)$ entries, i.e. rotational invariance dictates that the distribution of v_i and w_i is given by the Haar measure on the unit sphere of dimension $N-R^0$ and $T-R^0$, respectively. Define

$$\tilde{d}_k^{(1)} \equiv \frac{\tilde{v}' M_{\lambda^0} X_k M_{f^0}\tilde{w}}{\|M_{\lambda^0}\tilde{v}\|\|M_{f^0}\tilde{w}\|} \underset{d}{=} v_i' X_k w_t . \tag{S.57}$$

We want to show that $\tilde{d}_k^{(1)}$ has sufficiently high bounded moments. The squares $\|M_{\lambda^0}\tilde{v}\|^{-1}$ and $\|M_{f^0}\tilde{w}\|^{-1}$ have inverse chi-square distributions with $N-R^0$ and $T-R^0$ degrees of freedom, respectively. The inverse chi-square distribution with dof ν possesses all moments smaller than $\nu/2$, and for every $\xi>0$ there exists a constant $a_1>0$ such that we have

$$\mathbb{E}\left\|\frac{M_{\lambda^0}\tilde{v}}{\sqrt{N}}\right\|^{-\xi} < a_1, \qquad\qquad \mathbb{E}\left\|\frac{M_{f^0}\tilde{w}}{\sqrt{T}}\right\|^{-\xi} < a_1,$$

for all $N-R^0>4\xi$ and $T-R^0>4\xi$. Since we assume that the $(24+\epsilon)$'th moment of $M_{\lambda^0}X_kM_{f^0}$ is uniformly bounded, there exists a constant $a_2>0$ such that

$$\mathbb{E}\left|\frac{\tilde{v}' M_{\lambda^0} X_k M_{f^0}\tilde{w}}{\sqrt{NT}}\right|^{24+\epsilon} < a_2 ,$$

Applying the Cauchy-Schwarz-inequality we thus obtain

$$\mathbb{E}\left|\tilde{d}_k^{(1)}\right|^{\frac{1}{1/(24+\epsilon)+2/\xi}} = \mathbb{E}\left|v_i' X_k w_j\right|^{\frac{1}{1/(24+\epsilon)+2/\xi}} < \max(a_1,a_2) , \tag{S.58}$$

and thus

$$\max_{i,t,k}\left|v_i' X_k w_t\right| = \mathcal{O}_p\left((NT)^{1/(24+\epsilon)+2/\xi}\right) = o_p(N^{1/12}/\log(N)) , \tag{S.59}$$

where in the last step we chose ξ sufficiently large such that $2/(24+\epsilon)+4/\xi<1/12$. We have thus shown that $d_{NT}^{(1)} = o_p(N^{1/12}/\log(N))$.

Let $\tilde{f}$ be a $N\times R^0$ matrix such that $P_{f^0}=P_{\tilde{f}}$, i.e. the column spaces of f^0 and $\tilde{f}$ are identical, and $\tilde{f}'\tilde{f}=\mathbb{1}_{R^0}$. Then we have $\|v_i' e P_{f^0}\| = \|v_i' e \tilde{f}'\|$. Note that $e\tilde{f}'$ is a $N\times R^0$ matrix with iid normal entries, independently distributed of v_i for all $i=1,\ldots,n$. Together with the fact that the v_i have iid normal entries, as well, it is straightforward to show that $\max_i \|v_i' e P_{f^0}\| = \mathcal{O}_p(N^\delta)$ for any $\delta>0$, and the same is true for $\max_i \|w_i' e' P_{\lambda^0}\|$, i.e. we have $d_{NT}^{(2)} = o_p(N^{1/12}/\log(N))$ and $d_{NT}^{(3)} = o_p(N^{1/12}/\log(N))$.

We have

$$
\begin{aligned}
d_{NT}^{(4)} = N^{-3/4} \max_{i,k} \|v_i' X_k P_{f^0}\| \leq N^{-3/4} \max_{i,k} \|v_i' X_k\| \\
\leq N^{-3/4} \sqrt{T} \max_{i,t,k} |v_i' X_{k,\cdot t}| + o_p(1) \\
\leq N^{-3/4} \sqrt{T} \max_{i,t,k} \sum_{j=1}^{N} v_{i,j} X_{k,jt} + o_p(1),
\end{aligned}
\tag{S.60}
$$

where $t = 1, \ldots, T$, and we applied the inequality $\|z\| \leq \sqrt{T} \max_t z_t$, which holds for all T-vectors z. The remaining treatment of $d_{NT}^{(4)}$ is analogous to that of $d_{NT}^{(1)}$. Using the assumption that $(M_{\lambda^0} X_k)_{it}$ has uniformly bounded $(6+\epsilon)$'th moment and (S.56) one can show that $\max_{i,t,k} \sum_{j=1}^{N} v_{i,j} X_{k,jt}$ grows at a rate of $\mathcal{O}_p(N^{2/(6+\epsilon)+1/\xi})$, and therefore $d_{NT}^{(4)} = o_p(N^{1/12}/\log(N))$. Analogously one can show that $d_{NT}^{(5)} = o_p(N^{1/12}/\log(N))$, so that we can indeed conclude $d_{NT} = o_p(N^{1/12}/\log(N))$.

Without loss of generality, we set $\sigma = 1$ in the rest of the proof. We want to show that $q_{NT} = \log(N) N^{1/6}$ also satisfies

$$
\frac{1}{q_{NT}(T-R^0)} \sum_{i=q_{NT}}^{N-R^0} \frac{1}{\mu_{R-R^0} - \mu_i} = \mathcal{O}_p(1),
$$

where $\mu_i = \rho_i/(T-R^0)$. Note that it is not important whether the sum runs to $N-R^0$ or $T-R^0$, since those contributions of small eigenvalues are of order $\mathcal{O}_p(1)$ anyways. Without loss of generality let $\lim_{N,T\to\infty} N/T = \kappa^2 \leq 1$ in the rest of this proof (the proof for $\kappa \geq 1$ is analogous, since all arguments are symmetric under interchange of N and T). Let $\mu_{NT} = \left[(N-R)^{1/2} + (T-R)^{1/2}\right]^2$, $\sigma_{NT} = \left[(N-R)^{1/2} + (T-R)^{1/2}\right] \left[(N-R)^{-1/2} + (T-R)^{-1/2}\right]^{1/3}$, $\overline{x} = \lim_{N,T\to\infty} \mu_{NT}/(T-R) = (1+\kappa)^2$, and $\underline{x} = (1-\kappa)^2$. From Theorem 1 in Soshnikov (2002) we know that that the joint distribution of $\sigma_{NT}^{-1}(\rho_1 - \mu_{NT}, \rho_2 - \mu_{NT}, \ldots, \rho_{R+1} - \mu_{NT})$ converges to the Tracy-Widom law, i.e. to the limiting distribution of the first R+1 eigenvalues of the Gaussian Orthogonal Ensemble. Note that σ_{NT} is of order $N^{1/3}$, and that the Tracy-Widom law is a continuous distribution, so that the result of Soshnikov implies that

$$
\overline{x} - \mu_R = \mathcal{O}_p(N^{-2/3}), \qquad (\mu_R - \mu_{R+1})^{-1} = \mathcal{O}_p\left(N^{2/3}\right). \tag{S.61}
$$

The empirical distribution of the μ_i is defined as $F_{NT}(x) = N^{-1} \sum_{i=1}^{N} 1(\mu_i \leq x)$, where $1(.)$ is the indicator function. This empirical distribution converges to the Marchenko-Pastur limiting spectral distribution $F_{\text{LSD}}(x)$, which has domain $[\underline{x}, \overline{x}]$, and whose density

$f_{\mathrm{LSD}}(x) = dF_{\mathrm{LSD}}(x)/dx$ is given by

$$f_{\mathrm{LSD}}(x) = \frac{1}{2\pi\kappa^2 x}\sqrt{(\overline{x}-x)(x-\underline{x})}\ . \tag{S.62}$$

An upper bound for $f_{\mathrm{LSD}}(x)$ is given by $\frac{1}{2\pi\kappa^2\underline{x}}\sqrt{(\overline{x}-x)(\overline{x}-\underline{x})}$, and by integrating that upper bound we obtain

$$1 - F_{\mathrm{LSD}}(x)\ \le\ a\,(\overline{x}-x)^{3/2}\ , \qquad\qquad a = \frac{2}{3\pi\kappa^{3/2}\underline{x}}\ . \tag{S.63}$$

From Theorem 1.2 in Götze and Tikhomirov (2007) we know that

$$\sup_x |F_{NT}(x) - F_{LSD}(x)| = \mathcal{O}_p(N^{-1/2})\ . \tag{S.64}$$

Let $c_{1,NT} = \lceil 2N^{1/2+\epsilon}\rceil$ and $c_{2,NT} = \lceil 2N^{3/4}\rceil$, where $\lceil a\rceil$ is the smallest integer larger or equal to a. Plugging in $x = \mu_{c_{1,NT}}$ into the result of Götze and Tikhomirov, and using $F_{NT}(\mu_i) = 1-(i-1)/N$, we find

$$\begin{aligned} a\left(\overline{x}-\mu_{c_{1,NT}}\right)^{3/2} &\ge 1 - F_{\mathrm{LSD}}(\mu_{c_{1,NT}}) = \frac{c_{1,NT}-1}{N} + \mathcal{O}_p(N^{-1/2}) \\ &\ge N^{-1/2+\epsilon}, \qquad \text{wpa1}. \end{aligned} \tag{S.65}$$

Using this and (S.61) we obtain $(\mu_R - \mu_{c_1})^{-1} = \mathcal{O}_p\left(N^{1/3-2/3\epsilon}\right)$. Analogously one can show that $(\mu_R-\mu_{c_2})^{-1} = \mathcal{O}_p\left(N^{1/6}\right)$. In the following we just write q, c_1 and c_2 for q_{NT}, $c_{1,NT}$ and $c_{2,NT}$, and we set $n = T - R^0$. Combining the above results we find

$$\begin{aligned} \frac{1}{q\,n}\sum_{i=q}^{N-R^0}\frac{1}{\mu_R-\mu_i} &= \frac{1}{q\,n}\sum_{i=q}^{c_1-1}\frac{1}{\mu_R-\mu_i} + \frac{1}{q\,n}\sum_{i=c_1}^{c_2-1}\frac{1}{\mu_R-\mu_i} + \frac{1}{q\,n}\sum_{i=c_2}^{N-R^0}\frac{1}{\mu_R-\mu_i} \\ &\le \frac{c_1}{qn(\mu_R-\mu_{R+1})} + \frac{c_2}{qn(\mu_R-\mu_{c_1})} + \frac{N-R^0}{qn(\mu_R-\mu_{c_2})} \\ &= \mathcal{O}_p(1) + \mathcal{O}_p(N^{-1/12-5/3\epsilon}) + \mathcal{O}_p(N^{-\epsilon}) = \mathcal{O}_p(1)\ . \end{aligned}$$

This is what we wanted to show. ■