Moon, Hyungsik Roger; Weidner, Martin

Working Paper

Nuclear norm regularized estimation of panel regression models

cemmap working paper, No. CWP14/19

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Moon, Hyungsik Roger; Weidner, Martin (2019) : Nuclear norm regularized estimation of panel regression models, cemmap working paper, No. CWP14/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.1419

This Version is available at:
https://hdl.handle.net/10419/211107

cemmap
centre for microdata methods and practice

Nuclear norm regularized estimation of panel regression models

Hyungsik Roger Moon
Martin Weidner

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP14/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nuclear Norm Regularized Estimation of Panel Regression Models*

Hyungsik Roger Moon[‡§] Martin Weidner[¶]

March 28, 2019

Abstract

In this paper we investigate panel regression models with interactive fixed effects. We propose two new estimation methods that are based on minimizing convex objective functions. The first method minimizes the sum of squared residuals with a nuclear (trace) norm regularization. The second method minimizes the nuclear norm of the residuals. We establish the consistency of the two resulting estimators. Those estimators have a very important computational advantage compared to the existing least squares (LS) estimator, in that they are defined as minimizers of a convex objective function. In addition, the nuclear norm penalization helps to resolve a potential identification problem for interactive fixed effect models, in particular when the regressors are low-rank and the number of the factors is unknown. We also show how to construct estimators that are asymptotically equivalent to the least squares (LS) estimator in Bai (2009) and Moon and Weidner (2017) by using our nuclear norm regularized or minimized estimators as initial values for a finite number of LS minimizing iteration steps. This iteration avoids any non-convex minimization, while the original LS estimation problem is generally non-convex, and can have multiple local minima.

Keywords: Interactive Fixed Effects, Factor Models, Nuclear Norm Regularization, Convex Optimization, Iterative Estimation

*We are grateful for comments and suggestions from the participants of the 2016 IAAE Conference, 2017 UCLA-USC Conference, 2017 Cambridge INET Panel Conference, 2018 International Conference of Econometrics in Chengdu, 2018 EMES, 2018 AMES, and seminars at Boston University, Columbia University, Northwestern University, Oxford University, University of Bath, University of Chicago, UC Riverside, University of Iowa, University of Surrey, Yale University, University Carlos III in Madrid, and CEMFI. Weidner acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grant ERC-2014-CoG-646917-ROMIA.

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

[§]School of Economics, Yonsei University, Seoul, Korea.

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

1 Introduction

In this paper we consider a linear panel regression model of the form

$$Y_{it} = \sum_{k=1}^{K} \beta_{0,k} X_{k,it} + \sum_{r=1}^{R_0} \lambda_{0,ir} f_{0,tr} + E_{it}, \tag{1}$$

where $i = 1 \dots N$ and $t = 1 \dots T$ label the cross-sectional units and the time periods, respectively, Y_{it} is an observed dependent variable, $X_{k,it}$ are observed regressors, $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,K})'$ are unknown regression coefficients, $f_{0,tr}$ and $\lambda_{0,ir}$ are unobserved factors and factor loadings, E_{it} is an unobserved idiosyncratic error term, R_0 denotes the number of factors, and K denotes the number of regressors. The factors and loadings are also called interactive fixed effects. They parsimoniously represent heterogeneity in both dimensions of the panel, and they contain the conventional additive error components as a special case. We assume that $R_0 \ll \min(N, T)$, and for our asymptotic results we will consider R_0 as fixed, as $N, T \to \infty$. We can rewrite this model in matrix notation as

$$Y = \beta_0 \cdot X + \Gamma_0 + E, \tag{2}$$

where $\beta_0 \cdot X := \sum_{k=1}^{K} X_k \beta_{0,k}$ and $\Gamma_0 := \lambda_0 f_0'$, and Y, X_k, Γ_0 and E are $N \times T$ matrices, while λ_0 and f_0 are $N \times R_0$ and $T \times R_0$ matrices, respectively. The parameters β_0 and Γ_0 are treated as non-random throughout the whole paper, that is, all stochastic statements are implicitly conditional on their realization. Without loss of generality we assume $R_0 = \text{rank}(\Gamma_0)$.

One widely used estimation technique for interactive fixed effect panel regressions is the least squares (LS) method,[1] which treats λ and f as parameters to estimate (fixed effects).[2] Let the Frobenius norm of an $N \times T$ matrix A be $\|A\|_2 := \left(\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}^2 \right)^{1/2}$. Then, the LS estimator for β reads

$$\widehat{\beta}_{\text{LS},R} := \underset{\beta \in \mathbb{R}^K}{\text{argmin}}\, L_R(\beta), \qquad L_R(\beta) := \min_{\{\lambda \in \mathbb{R}^{N \times R},\, f \in \mathbb{R}^{T \times R}\}} \frac{1}{2NT} \|Y - \beta \cdot X - \lambda f'\|_2^2, \tag{3}$$

where R is the number of factors chosen in estimation. A matrix $\Gamma \in \mathbb{R}^{N \times T}$ can be written as $\Gamma = \lambda f'$, for some $\lambda \in \mathbb{R}^{N \times R}$ and $f \in \mathbb{R}^{T \times R}$, if and only if $\text{rank}(\Gamma) \leq R$. The profiled

[1] The LS estimator in this context is also sometimes called concentrated least squares estimator, and was originally proposed by Kiefer (1980).

[2] Other estimation methods of panel regressions with interactive fixed effects include the quasi-difference approach (e.g., Holtz-Eakin, Newey, and Rosen 1988), generalized method of moments estimation (e.g. Ahn, Lee, and Schmidt 2001, 2013), the common correlated random effect method (e.g., Pesaran 2006), the decision theoretic approach (e.g., Chamberlain and Moreira 2009), and Lasso type shrinkage methods on fixed effects (e.g., Cheng, Liao, and Schorfheide 2016, Lu and Su 2016, Su, Shi, and Phillips 2016).

least square objective function $L_R(\beta)$ can therefore equivalently be expressed as

$$L_R(\beta) = \min_{\{\,\Gamma\in\mathbb{R}^{N\times T}\,|\,\mathrm{rank}(\Gamma)\leq R\}} \frac{1}{2NT}\left\|Y-\beta\cdot X-\Gamma\right\|_2^2. \tag{4}$$

It is known that under appropriate regularity conditions (including exogeneity of $X_{k,it}$ with respect to E_{it}), for $R \geq R_0$, and as $N, T \to \infty$ at the same rate, the LS estimator $\widehat{\beta}_{\mathrm{LS},R}$ is $\sqrt{NT}$-consistent and asymptotically normal, with a bias in the limiting distribution that can be corrected for (e.g., Bai 2009, Moon and Weidner 2015, 2017).

The LS estimation approach is convenient, because it does not restrict the relationship between the unobserved heterogeneity (Γ_0) and the observed explanatory variables $(X_1, ..., X_K)$. However, the calculation of $\widehat{\beta}_{\mathrm{LS},R}$ requires solving a non-convex optimization problem. While $\|Y-\beta\cdot X-\Gamma\|_2^2$ is a convex function of β and Γ, the profiled objective function $L_R(\beta)$ is in general not convex in β, and can have multiple local minima, as will be discussed in Section 2.1 in more detail. The reason for the non-convexity is that the constraint $\mathrm{rank}(\Gamma) \leq R$ is non-convex.

In this paper we use a convex relaxation of this rank constraint. Let $s(\Gamma) := [s_1(\Gamma), s_2(\Gamma), \ldots, s_{\min(N,T)}(\Gamma)]$ be the vector of singular values of Γ.[3] The rank of a matrix is equal to the number of non-zero singular values, that is, $\mathrm{rank}(\Gamma) = \|s(\Gamma)\|_0$, where $\|v\|_0$ equals the number of non-zero elements of the vector v (sometimes calles the "ℓ^0-norm" of v). The nuclear norm of Γ is defined by $\|\Gamma\|_1 := \|s(\Gamma)\|_1 = \sum_{r=1}^{\min(N,T)} s_r(\Gamma)$, that is, the nuclear norm of the matrix Γ is simply the ℓ^1-norm of the vector $s(\Gamma)$.[4] A convex relaxation of (4) can then be obtained by replacing the non-convex constraint $\mathrm{rank}(\Gamma) \leq R$ by the convex constraint $\|\Gamma\|_1 \leq c$, for some constant c. This gives

$$\begin{aligned}\min_{\left\{\Gamma\in\mathbb{R}^{N\times T}\,\middle|\,\|\Gamma\|_1\leq c_\psi\right\}} & \frac{1}{2NT}\left\|Y-\beta\cdot X-\Gamma\right\|_2^2 \\ &= \min_{\Gamma\in\mathbb{R}^{N\times T}}\left[\frac{1}{2NT}\left\|Y-\beta\cdot X-\Gamma\right\|_2^2+\frac{\psi}{\sqrt{NT}}\left\|\Gamma\right\|_1\right] =: Q_\psi(\beta),\end{aligned} \tag{5}$$

where in the second line we replaced the constraint on the nuclear norm by a nuclear-norm penalty term.[5] Choosing a particular penalization parameter $\psi > 0$ is equivalent to choosing

[3] The non-zero singular values of Γ are the square roots of non-zero eigenvalues of $\Gamma\Gamma'$. Singular values are non-negative by definition.

[4] The nuclear norm $\|\Gamma\|_1$ is the convex envelope of $\mathrm{rank}(\Gamma)$ over the set of matrices with spectral norm at most one, see e.g. Recht, Fazel, and Parrilo (2010). The nuclear norm is also sometimes called trace norm, Schatten 1-norm, or Ky Fan n-norm. Our index notation for the nuclear norm $\|\Gamma\|_1$, Frobenius norm $\|\Gamma\|_2$, and spectral norm $\|\Gamma\|_\infty = \lim_{q\to\infty}\|\Gamma\|_q$ is motivated by the unifying formula $\|\Gamma\|_q = \sum_{r=1}^{\min(N,T)}[s_r(\Gamma)]^q$.

[5] The normalizations with $1/(2NT)$ and $1/\sqrt{NT}$ in (5) are somewhat arbitrary, but turn out to be convenient for our purposes.

a particular value for $c = c_\psi$, and we find it more convenient to parameterize the convex relaxation $Q_\psi(\beta)$ of $L_R(\beta)$ by ψ instead of c. For a given $\psi > 0$ the nuclear-norm regularized estimator reads

$$\widehat{\beta}_\psi := \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \, Q_\psi(\beta).$$

We also define $\widehat{\beta}_* := \lim_{\psi \to 0} \widehat{\beta}_\psi$ for fixed N and T.[6] We will show in Section 2.2 that $\widehat{\beta}_* = \operatorname{argmin}_\beta \|Y - \beta \cdot X\|_1$, that is, $\widehat{\beta}_*$ can alternatively be obtained by minimizing the nuclear norm of $Y - \beta \cdot X$. The main goal of this paper is to explore the properties of $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$, that is, we want to understand how these estimators can be used to help identify and estimate β_0.

Those estimators have a very important computational advantage compared to the LS estimator, in that they are defined as minimizers of convex objective functions. The LS objective function in (4) is in general non-convex and can have multiple local minima. This can become a serious computational obstacle if the dimension of the regression coefficients is large. If the underlying panel regression model is nonlinear (e.g., Chen 2014, Chen, Fernandez-Val, and Weidner 2014), then optimizing a non-convex objective function with respect to the high-dimensional parameters λ and f becomes even more challenging. By contrast, under appropriate non-collinearity conditions on the regressors, the nuclear norm penalized objective function in (5) is strictly convex and therefore has a unique local minimum that is also the global minimum.

In addition to this computational advantage the nuclear norm penalization also helps to resolve a potential identification problem for interactive fixed effect models. Namely, without restrictions on the parameter matrix Γ_0 in (2), we cannot separate $\beta_0 \cdot X$ and Γ_0 uniquely, because for any other parameter β we can write

$$Y = \beta_0 \cdot X + \Gamma_0 + E = \beta \cdot X + \Gamma(\beta, X) + E, \quad \text{where} \quad \Gamma(\beta, X) := \Gamma_0 - (\beta - \beta_0)X,$$

implying that (β_0, Γ_0) and $(\beta, \Gamma(\beta, X))$ are observationally equivalent. If any non-trivial linear combination of the regressors X_k is a high-rank matrix, then the assumption that $R_0 = \operatorname{rank}(\Gamma_0) \ll \min(N, T)$ is sufficient to identify β_0, because $\operatorname{rank}[\Gamma(\beta, X)]$ will be large for any other value of β. However, if some of the regressors X_k have low rank, and the true number of factors R_0 is unknown, then there is an identification problem, and some regularization device is needed to resolve this. In Section 2 we show that the nuclear norm

[6] Here, the limit $\psi \to 0$ is for fixed N and T, and has nothing to do with our large N, T asymptotic considerations.

penalization indeed provides such a regularization device to uniquely identify β_0.

After that identification discussion, we establish asymptotic results for $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ when both panel dimensions become large. Under appropriate regularity conditions we show $\sqrt{\min(N,T)}$-consistency of these estimators. We also show how to use $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ as initial values for a finite iteration procedure that gives improved estimates that are asymptotically equivalent to the LS estimator.

Nuclear norm penalized estimation has been widely studied in machine learning and statistical learning literature. There, the parameter of interest is usually the matrix that we call Γ in our model, in particular, there are many papers that use this penalization method in matrix completion (e.g., Recht, Fazel, and Parrilo 2010 and Hastie, Tibshirani, and Wainwright 2015 for recent surveys), and for reduced rank regression estimation (e.g., Rohde and Tsybakov 2011). More recently, nuclear norm penalization has also been used in the econometrics literature: Bai and Ng (2017) use it to improve estimation in a pure factor models. Athey, Bayati, Doudchenko, Imbens, and Khosravi (2017) apply nuclear norm penalization to treatment effect estimation with unbalanced panel data due to missing observations together with a regularization on the high dimensional regression coefficients – their primary interest is to predict the left-hand side variable using the regularization. Chernozhukov, Hansen, Liao, and Zhu (2018) consider panel regression models with heterogeneous coefficients, while in this paper we focus on panel regression with homogenous coefficients. To the best of our knowledge, our results here on the estimates of the common regression coefficients β_0 are new in this literature, and the nuclear norm minimizing estimator $\widehat{\beta}_*$ has also not been proposed previously.

The paper is organized as follows. Section 2 provides theoretical motivations of nuclear regularization over the conventional rank restriction. In Section 3 we derive consistency results on $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$. Section 4 shows how to use these two estimators as a preliminary estimator to construct an estimator through iterations that achieves asymptotic equivalence to the fixed effect estimator. Section 5 investigates finite sample properties of the estimators. In Section 6 we briefly discuss extensions to nonlinear panel models with interactive fixed effects, and Section 7 concludes the paper. All technical derivations and proofs are presented in the appendix or supplementary appendix.

2 Motivation of Nuclear Norm Regularization

In this section we provide further motivation and explanation of the nuclear norm regularized estimation method. This estimation approach comes with the computational advantage of having a convex objective function, and it also provides a solution to the identification

problem of interactive fixed effect models with low-rank regressors.

2.1 Convex Relaxation

We have already introduced the profile LS objective function $L_R(\beta)$ and its convex relaxation $Q_\psi(\beta)$ in the introduction. Here, we explain those objective functions further. Firstly, we want to briefly explain why $Q_\psi(\beta)$ is indeed convex. We have introduced the nuclear norm as $\|\Gamma\|_1 := \sum_{r=1}^{\min(N,T)} s_r(\Gamma)$, but it is not obvious from this definition that $\|\Gamma\|_1$ is convex in Γ, because the singular values $s_r(\Gamma)$ themselves are generally not convex functions of Γ, except for $r = 1$. A useful alternative definition of the nuclear norm is

$$\|\Gamma\|_1 = \max_{\left\{A \in \mathbb{R}^{N\times T} \middle| \|A\|_\infty \leq 1\right\}} \mathrm{Tr}(A'\,\Gamma), \tag{6}$$

that is, the nuclear norm is dual to the spectral norm $\|\cdot\|_\infty$. From this it is easy to see that $\|\cdot\|_1$ is indeed a matrix norm, and thus convex in Γ.[7] Therefore, the nuclear norm regularized objective function

$$\frac{1}{2NT}\left\|Y - \beta\cdot X - \Gamma\right\|_2^2 + \frac{\psi}{\sqrt{NT}}\left\|\Gamma\right\|_1$$

as a function of (β, Γ) is convex. Profiling with respect to Γ preserves convexity, that is, $Q_\psi(\beta)$ is also convex.

By contrast, the least squares objective $\frac{1}{2NT}\left\|Y - \beta\cdot X - \lambda f'\right\|_2^2$ is generally non-convex in the parameters β, λ and f. However, the non-convexity of the LS minimization over λ and f is actually not a serious problem in computing the profile objective function $L_R(\beta)$, as long as the regression model is linear and one of the dimensions N or T is not too large.[8] Recall that $s_r(Y - \beta\cdot X)$ is the r^{th} largest singular value of the matrix $(Y - \beta\cdot X)$, for $r = 1, ..., \min(N, T)$. One can show that (see Moon and Weidner (2017)) the profile least

[7] Let B and C be matrices of the same size. Then, by (6) there exists a matrix A of the same size with $\|A\|_\infty \leq 1$ such that $\|B+C\|_1 = \mathrm{Tr}[A'\,(B+C)] = \mathrm{Tr}(A'B) + \mathrm{Tr}(A'C) \leq \|B\|_1 + \|C\|_1$, which is the triangle inequality for the nuclear norm. Together with absolute homogeneity of $\|\cdot\|_1$ this implies convexity.

[8] The optimal $\widehat{\lambda}$ and $\widehat{f}$ are simply given by the leading R principal components of $Y - \beta\cdot X$. Calculating them requires to find the eigenvalues and eigenvectors of either the $N \times N$ matrix $(Y-\beta\cdot X)(Y-\beta\cdot X)'$ or the $T \times T$ matrix $(Y-\beta\cdot X)'(Y-\beta\cdot X)$, which takes at most a few seconds on modern computers, as long as $\min(N,T) \lessapprox 5.000$, or so. The non-zero eigenvalues of $(Y-\beta\cdot X)(Y-\beta\cdot X)'$ and $(Y-\beta\cdot X)'(Y-\beta\cdot X)$ are identical, and are equal to the square of the non-zero singular values of $Y - \beta\cdot X$.

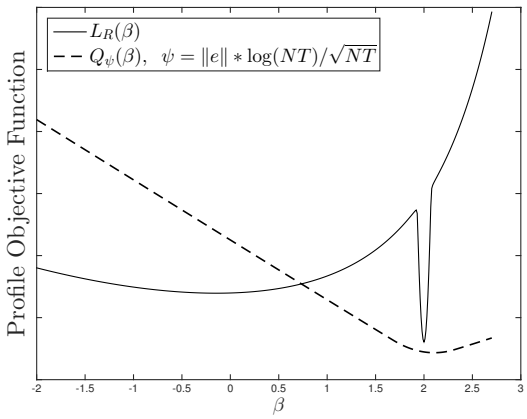

Figure 1: Plot of $L_R(\beta)$ and $Q_\psi(\beta)$ for the example detailed in Appendix A.1. The true parameter is $\beta_0 = 2$.

squares objective function is

$$L_R(\beta) = \frac{1}{2\,NT} \sum_{r=R+1}^{\min(N,T)} [s_r(Y - \beta \cdot X)]^2, \tag{7}$$

where the largest R singular values are omitted in the sum, because they were absorbed by the principal component estimates $\widehat{\lambda}$ and $\widehat{f}$. The remaining problem in calculating $\widehat{\beta}_{\mathrm{LS},R}$ is the generally non-convex minimization of $L_R(\beta)$ over β.[9] To illustrate the potential difficulty caused by this non-convexity, in Figure 1 we plot $L_R(\beta)$ for the simple example described in Appendix A.1. In this example $L_R(\beta)$ is non-convex and has two local minima, one of which (the global one) is close to the true parameter $\beta_0 = 2$. The figure also shows that $Q_\psi(\beta)$ is convex and only has a single local minimum.

For any $\psi > 0$ define the functions $\ell_\psi : [0,\infty) \mapsto [0,\infty)$ and $q_\psi : [0,\infty) \mapsto [0,\infty)$ by

$$\ell_\psi(s) := \begin{cases} \frac{1}{2}\, s^2, & \text{for } s < \psi, \\ 0, & \text{for } s \geq \psi, \end{cases} \qquad q_\psi(s) := \begin{cases} \frac{1}{2}\, s^2, & \text{for } s < \psi, \\ \psi s - \frac{\psi^2}{2}, & \text{for } s \geq \psi. \end{cases} \tag{8}$$

For an $N \times T$ matrix A let $\ell_\psi(A) := \sum_{r=1}^{\min(N,T)} \ell_\psi(s_r(A))$ and $q_\psi(A) := \sum_{r=1}^{\min(N,T)} q_\psi(s_r(A))$.

[9] In our discussion here we focus on the calculation of $\widehat{\beta}_{\mathrm{LS},R}$ via minimization of the profile objective function $L_R(\beta)$. More generally, $\widehat{\beta}_{\mathrm{LS},R}$ can be obtained by any method that minimizes $\|Y - \beta \cdot X - \lambda f'\|_2^2$ over β, λ, f, see e.g. Bai (2009) or the supplementary appendix in Moon and Weidner (2015). For any such method the non-convexity of the objective function is a potential problem, because the algorithm may converge to a local minimum, or potentially even to a critical point that is not a local minimum.

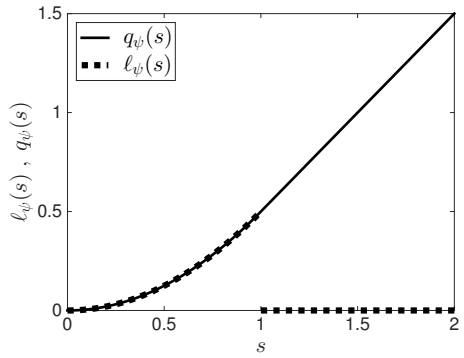

Figure 2: Plot of the functions $q_\psi(s)$ and $\ell_\psi(s)$ for $\psi = 1$.

We can then rewrite (7) as

$$L_R(\beta) = \ell_{\psi(\beta,R)}\left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right), \tag{9}$$

where $\psi(\beta, R)$ satisfies

$$s_{R+1}\left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right) < \psi(\beta, R) \leq s_R\left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right). \tag{10}$$

Here, the normalization with $1/\sqrt{NT}$ is natural, because under standard assumptions the largest singular value of $Y - \beta \cdot X$ is of order $\sqrt{NT}$, as N and T grow. The formulation (9) is interesting for us, because the following lemma shows that we have a very similar representation for $Q_\psi(\beta)$.

Lemma 1. *For any $\beta \in \mathbb{R}^K$ and any $\psi > 0$ we have*

$$Q_\psi(\beta) = q_\psi\left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right).$$

The proof is given in the appendix. Figure 2 shows the functions $q_\psi(s)$ and $\ell_\psi(s)$ for real valued arguments s and $\psi = 1$. For values $s < \psi$ the functions are identical, but at $s = \psi$ the function $\ell_\psi(s)$ has a non-continuous jump, implying that $\ell_\psi(s)$ is non-convex, while $q_\psi(s)$ continues linearly for $s \geq \psi$, thus remaining convex.

Comparing $L_R(\beta)$ and $Q_\psi(\beta)$ we see that the parameter R that counts the number of factors is replaced by the parameter ψ that characterizes the magnitude at which the singular values of $(Y - \beta \cdot X)/\sqrt{NT}$ are considered to be factors, and for a given β the relationship between R and ψ is given by (10). Large R corresponds to small ψ, and vice versa.

Furthermore, $\widehat{\Gamma}_\psi(\beta) := \text{argmin}_\Gamma \frac{1}{2NT} \|Y - \beta \cdot X - \Gamma\|_2^2 + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1$ has singular values[10]

$$s_r\left(\widehat{\Gamma}_\psi(\beta)\right) = \max\left(s_r\left((Y - \beta \cdot X)/\sqrt{NT}\right) - \psi, 0\right) \quad r = 1, ..., \min(N, T),$$

that is, the nuclear norm penalization shrinks the singular of $Y - \beta \cdot X$ towards zero by a fixed amount.

Fixing ψ as opposed to fixing R already changes the functional form of the profile objective function, because according to (10) their relationship depends on β. In addition, the objective function is convexified by replacing the function $\ell_\psi(s)$ that is applied to the singular values of $(Y - \beta \cdot X)/\sqrt{NT}$ with the function $q_\psi(s)$, as defined in (8). The function $q_\psi(s)$ provides a convex continuation of $\ell_\psi(s)$ for $s \geq \psi$.

Using the closed-form expression for $Q_\psi(\beta)$ in Lemma 1, and noticing that it is convex in β, one can compute the minimizer $\widehat{\beta}_\psi$ of $Q_\psi(\beta)$ using various optimizing algorithms for a convex function (see chapter 5 of Hastie, Tibshirani, and Wainwright 2015). If the dimension of β is small, then one may even use a simple grid search method to find $\widehat{\beta}_\psi$. We will discuss a data dependent choice of the penalty parameter ψ in Section 5.

2.2 Unique Matrix Separation

When estimating the interactive fixed effect model (1) in practice both β_0 and R_0 are unknown. Showing that β_0 and R_0 can be consistently estimated jointly is a difficult problem in general.[11] Within the interactive fixed effects estimation framework this joint inference problem has only been successfully addressed when both of the following assumptions are satisfies:[12]

(C1) There is a known upper bound $R_{\max}$ such that $R_0 \leq R_{\max}$.

(C2) All the regressors X_k are "high-rank regressors", that is, $\text{rank}(X_k)$ is large for all k.

Under those assumptions (and other regularity conditions) the consistency proofs of Bai (2009) and Moon and Weidner (2015) are applicable to the LS estimator for β that uses $R = R_{\max} \geq R_0$ factors in the estimation, and one can also show the convergence rate result

[10] See Lemma S.1 in the supplementary appendix for details.

[11] The problem of joint identification of β_0 and R_0 is often avoided in the literature. Some papers (e.g. Bai 2009, Li, Qian, and Su 2016, Moon and Weidner 2017) assume that the number of factors R_0 is known when showing consistency for an estimator of β_0. Alternatively, Lu and Su (2016) allow for unknown R_0, but assume consistency of their estimator for β_0.

[12] Some existing estimation methods avoid specifying R when estimating β_0, but always at the cost of some additional assumptions on the data generating process. For example, the common correlated effects estimator of Pesaran (2006) avoids choosing R, but requires assumptions on how the factors f_0 enter into the observed regressors X_k, and requires all regressors of interest to be high-rank.

$\left\|\widehat{\beta}_{\mathrm{LS},R_{\max}} - \beta_0\right\| = O_P\left(\min(N,T)^{-1/2}\right)$, as $N,T \to \infty$. To obtain a consistent estimator for R_0 one can then apply inference methods from pure factor models without regressors (e.g. Bai and Ng 2002, Onatski 2010, Ahn and Horenstein 2013) to the matrix $Y - \widehat{\beta}_{\mathrm{LS},R_{\max}} \cdot X$.

The condition (C2) above is particularly strong, because "low-rank regressors" are quite common in practice. If we can write $X_{k,it} = w_{k,i}v_{k,t}$, then we have $\mathrm{rank}(X_k) = 1$, and the condition (C2) is violated. For example, Gobillon and Magnac (2016) estimate an interactive fixed effects model in a panel treatment effect setting, where the main regressor of interest indeed can be multiplicatively decomposed in this way, with $w_{k,i}$ being the treatment indicator of unit i, and $v_{k,t}$ being the time indicator of treatment. Those interactive fixed effects models for panel treatment effect applications have grown very popular recently.[13] However, when R_0 is unknown, then the presence of such low-rank regressors creates an identification problem, as illustrated by the following example.

Example 1. *Consider a single ($K = 1$) low-rank regressor $X_1 = vw'$, with vectors $v \in \mathbb{R}^N$ and $w \in \mathbb{R}^T$. Let $R_\bigstar = R_0 + 1$, $\lambda_\bigstar = [\lambda_0, v]$, and $f_\bigstar = [f_0, (\beta_{0,1} - \beta_{\bigstar,1})w]$. Then, model (1) with parameters β_0, R_0, λ_0, f_0 is observationally equivalent to the same model with parameters $\beta_\bigstar$, $R_\bigstar$, $\lambda_\bigstar$, $f_\bigstar$, because we have $\beta_{0,1}X_1 + \lambda_0 f_0' = \beta_{\bigstar,1}X_1 + \lambda_\bigstar f_\bigstar'$. Thus, β_0 is observationally equivalent to any other value $\beta_\bigstar$ if the true number of factors is unknown.*

The example shows that regression coefficients of low-rank regressors are not identified if R_0 is unknown, because $\beta \cdot X$ could simply be absorbed into the factor structure $\lambda f'$, which is also a low-rank matrix. Therefore, without some additional assumption or regularization device, the two low-rank matrices $\beta_0 \cdot X$ and $\Gamma_0 = \lambda_0 f_0'$ cannot be uniquely disentangled, which is what we mean by "unique matrix separation" in the title of this section.

Nuclear Norm Minimizing Estimation

In the following we explain how the nuclear norm minimization approach overcomes the restrictions (C1), that is, how to estimate regression coefficients when R_0 is unknown. We already introduced $\widehat{\beta}_* = \lim_{\psi\to 0}\widehat{\beta}_\psi$ in Section 1. Using Lemma 1 we can now characterize $\widehat{\beta}_*$ differently. It is easy to see that $\lim_{\psi\to 0}\psi^{-1}q_\psi(s) = s$, for $s \in [0,\infty)$, and therefore $\lim_{\psi\to 0}\psi^{-1}q_\psi(A) = \|A\|_1$, for $A \in \mathbb{R}^{N\times T}$. Lemma 1 thus implies that $\lim_{\psi\to 0}\psi^{-1}Q_\psi(\beta) = \|(Y-\beta\cdot X)/\sqrt{NT}\|_1$. Another way to see this is as follows. According to (10), the limit $\psi \to 0$ corresponds to choosing R very large, i.e., $R = \min(N,T)$. In this case, $\widehat{\Gamma}_\psi(\beta) = Y - \beta \cdot X$,

[13] Other recent applications in the same vein as Gobillon and Magnac (2016) are Chan and Kwok (2016), Powell (2017), Gobillon and Wolff (2017), Adams (2017), Piracha, Tani, and Tchuente (2017), Li (2018), to list just a few. This literature is also related to the synthetic control method (Abadie and Gardeazabal 2003, Abadie, Diamond, and Hainmueller 2010, Abadie, Diamond, and Hainmueller 2015; see also Hsiao, Ching, and Wan 2012).

and the profile objective function is $Q_\psi(\beta) = \frac{\psi}{\sqrt{NT}}\|\widehat{\Gamma}_\psi(\beta)\|_1 = \frac{\psi}{\sqrt{NT}}\|Y - \beta \cdot X\|_1$. From this we deduce $\lim_{\psi\to 0}\psi^{-1}Q_\psi(\beta) = \frac{\|Y-\beta\cdot X\|_1}{\sqrt{NT}}$.

Notice that for $\psi = 0$ we trivially have $Q_0(\beta) = 0$, but the rescaled objective function $\psi^{-1}Q_\psi(\beta)$ nevertheless has a non-trivial limit as $\psi \to 0$. Since rescaling the objective function by a constant does not change the minimizer we thus find that

$$\widehat{\beta}_* = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \|Y - \beta \cdot X\|_1 , \tag{11}$$

that is, the small ψ limit of the nuclear norm regularized estimator $\widehat{\beta}_\psi$ is the nuclear norm minimizing estimator $\widehat{\beta}_*$. The objective function $\|Y - \beta \cdot X\|_1$ is convex in β.

We cannot expect the LS estimator $\widehat{\beta}_{\mathrm{LS},R}$ to have good properties (in particular consistency) if we choose the number of factors equal to, or close to, its maximum possible value $R = \min(N, T)$. It is therefore somewhat surprising that $\widehat{\beta}_\psi$ has a well-defined limit as $\psi \to 0$, and that we are able to show consistency of the limiting estimator $\widehat{\beta}_*$ under appropriate regularity conditions in the following sections, because the resulting estimator for Γ is certainly not consistent for Γ_0 in that limit.[14]

The main significance of $\widehat{\beta}_*$ is that it provides an estimator for β that does not require any choice of "bandwidth parameter", because neither R nor ψ needs to be specified. It thus provides a method to estimate β_0 consistently without requiring knowledge of an upper bound on R_0 as in the condition (C1) above. In a second step we can then estimate R_0 consistently by applying, for example, the Bai and Ng (2002) method for pure factor models without regressors to the matrix $Y - \widehat{\beta}_* \cdot X$.

Notice that the pooled OLS estimator β_0 minimizes $\|Y-\beta\cdot X\|_2^2 = \sum_{r=1}^{\min(N,T)} s_r(Y-\beta\cdot X)^2$, the ℓ^2-norm of the singular values of the residual matrix, $Y - \beta \cdot X$, while the nuclear norm minimizing estimator $\widehat{\beta}_*$ minimizes the ℓ^1-norm, $\|Y - \beta \cdot X\|_1 = \sum_{r=1}^{\min(N,T)} s_r(Y - \beta \cdot X)$, of the residual matrix. The relationship between these two estimators is therefore analogous to that of the OLS estimator and the LAD (least absolute deviation) estimator for cross-sectional samples. $\widehat{\beta}_*$ is robust with respect to the unobserved factors, which are "outliers" in the singular value spectrum, while the pooled OLS estimator is not robust towards the presence of those unobserved factors (because they may be correlated with the regressors).

Nuclear Norm Penalization Approach for Matrix Separation

Next, we explain how the nuclear norm regularization approach helps to overcome the restrictions (C2) above, that is, how to estimate regression coefficients for low-rank regressors

[14] The $\psi \to 0$ limit (for fixed N, T) of the optimal Γ in (5) is $Y - \widehat{\beta}_* \cdot X$, which as N and T grow converges to $\lambda_0 f_0' + E$ for consistent $\widehat{\beta}_*$, that is, the estimator for Γ that corresponds to $\widehat{\beta}_*$ is not consistent for $\lambda_0 f_0'$.

when R_0 is unknown. The goal is to provide conditions on the regressors X_k under which the nuclear norm penalization approach indeed solves the matrix separation problem for low-rank regressors and interactive fixed effects.

We first want to answer this in a simplified setting, where the objective function is replaced by the expected objective function, that is, we consider

$$\bar{\beta}_\psi := \operatorname*{argmin}_\beta \min_\Gamma \left\{ \frac{1}{2NT} \mathbb{E}\left[\|Y - \beta \cdot X - \Gamma\|_2^2 \,\middle|\, X \right] + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1 \right\}. \tag{12}$$

Here, the expectation is conditional on all the regressors $(X_1, \ldots, X_K)$, and also implicitly on all the parameters β_0 and Γ_0, because those are treated as non-random.[15]

For a matrix A, let $\mathbf{P}_A := A(A'A)^\dagger A'$ and $\mathbf{M}_A := \mathbf{I} - \mathbf{P}_A$ be the projectors onto and orthogonal to the column span of A, where $\mathbf{I}$ is the identity matrix of appropriate dimensions, and $\dagger$ refers to the Moore-Penrose generalized inverse. Remember also our notation $\alpha \cdot X := \sum_{k=1}^K \alpha_k X_k$ for $\alpha \in \mathbb{R}^K$. For vectors v we write $\|v\|$ for the Euclidian norm.

Proposition 1. *Suppose that N, T, R_0 and K are fixed. Let $\mathbb{E}(E_{it} \mid X) = 0$, and $\mathbb{E}(E_{it}^2 \mid X) < \infty$, for all i, t. For all $\alpha \in \mathbb{R}^K \setminus \{0\}$ assume that*

$$\|\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\|_1 > \|\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}\|_1 . \tag{13}$$

Then, $\left\|\bar{\beta}_\psi - \beta_0\right\| = O(\psi)$, as $\psi \to 0$.

The proof is given in the appendix. The proposition considers fixed N, T, with only $\psi \to 0$.[16] The statement of the proposition implies that $\lim_{\psi \to 0} \bar{\beta}_\psi = \beta_0$. Thus, the proposition provides conditions under which the nuclear norm regularization approach identifies the true parameter β_0. The proposition does not restrict the rank of the regressors, so the result is applicable to both low-rank and high-rank regressors. The assumption $\mathbb{E}(E_{it} \mid X) = 0$ requires strict exogeneity of all regressors, but we will allow for pre-determined regressors in consistency results of Section 3.2 below.

The beauty of Proposition 1 is that it provides a very easy to interpret non-collinearity condition on the regressors X_k. It requires that for any linear combination of the regressors the part $\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}$, which cannot be explained by neither λ_0 nor f_0, is larger in terms of nuclear norm than the part $\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}$, which can be explained by both λ_0 and f_0. For a single ($K = 1$) regressor with $X_{1,it} = v_i w_t$, as in Example 1, the condition simply becomes

[15] $\bar{\beta}_\psi$ can be viewed as a population version of $\widehat{\beta}_\psi$ for an appropriately defined population distribution of Y conditional on X. But independent of this interpretation, $\bar{\beta}_\psi$ is a convenient tool of discussing the necessary non-collinearity condition on the regressors without requiring asymptotic analysis, yet.

[16] Display (S.6) in the appendix provides a bound on $\|\bar{\beta}_\psi - \beta_0\|$ for finite ψ, but the limit $\psi \to 0$ is what matters most to us, because that limit allows to identify β_0.

$\|\mathbf{M}_{\lambda_0} v\| \|\mathbf{M}_{f_0} w\| > \|\mathbf{P}_{\lambda_0} v\| \|\mathbf{P}_{f_0} w\|$. Here, $\|\mathbf{M}_{\lambda_0} v\|^2$ and $\|\mathbf{P}_{\lambda_0} v\|^2$ are the residual sum of squares, and the explained sum of squares of a regression of v_i on the $\lambda_{0,i}$, and analogously for $\|\mathbf{M}_{f_0} w\|^2$ and $\|\mathbf{P}_{f_0} w\|^2$. In Example 1 we obviously have $\|\mathbf{M}_{\lambda_\star} v\| = 0$ and $\|\mathbf{M}_{f_\star} w\| = 0$, that is, the parameters $R_\star$, $\beta_\star$, $\lambda_\star$, $f_\star$ are ruled out by the condition on the regressors in Proposition 1.

Related to the regularity condition (13) of Proposition 1, it is possible to show (see Bai 2009, Moon and Weidner 2017) that the weaker condition $\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0} \neq 0$ for any linear combination $\alpha \neq 0$ is sufficient for local identification of β in a sufficiently small neighborhood around β_0. However, that weaker condition is not sufficient for global identification of β_0, as illustrated by the examples in the supplementary appendix S.3 of Moon and Weidner (2017). The stronger condition (13) in Proposition 1 guarantees global identification of β_0 when using the nuclear norm penalization approach as a regularization device.

Providing such global identification conditions for models with low-rank regressors and unknown R_0 is a new contribution to the interactive fixed effects literature.[17] Our approach here is similar to the "Identification via a Strict Convex Penalty" proposed in Chen and Pouzo (2012).

3 Consistency of $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$

Proposition 1 above provides an identification result for β_0 for fixed N and T, based on the expected objective function. We now turn to the actual estimators $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ and investigates their sampling properties as $N, T \to \infty$.

All our consistency results for $\widehat{\beta}_\psi$ are for asymptotic sequences where $\psi = \psi_{NT} \to 0$, as $N, T \to \infty$, but we do not usually make the dependence of ψ on the sample size explicit. In addition, we assume that the number of the regressors K, and the true number of factors $R_0 = \text{rank}(\Gamma_0)$ are both fixed. However, do not restrict whether the factors are strong or weak, nor do we restrict the magnitude of Γ_0 in any matrix norm.

3.1 Consistency Results for Low-Rank Regressors

Here, we consider a special case where the regressors $X_1, ..., X_K$ are of low rank. This section is short, because the results here are relatively straightforward extensions of Section 2.2. The

[17] If the model would not have any idiosyncratic errors (i.e. $E = 0$), then $Y - \beta \cdot X = (\beta_0 - \beta) \cdot X + \Gamma_0$, and a natural solution to this identification problem would be to choose β as the solution to the rank minimization problem $\min_{\beta \in \mathbb{R}^K} \text{rank}\,(Y - \beta \cdot X)$, where at the true parameters we have $\text{rank}\,(Y - \beta_0 \cdot X) = \text{rank}(\Gamma_0) = R_0$, that is, we are minimizing the number of factors required to describe the data. However, once idiosyncratic errors E are present, then this rank minimization does not work, because $Y - \beta \cdot X$ is of large rank for all β.

more general case that allows both high-rank and low-rank regressors will be discussed in the following subsection.

Theorem 1. *Consider $N, T \to \infty$ with $\psi \to 0$, and assume that*

(i) There exists a constant c such that

$$\min_{\{\alpha \in \mathbb{R}^K : \|\alpha\|=1\}} \left\| \frac{\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}} \right\|_1 - \left\| \frac{\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}} \right\|_1 \geq c > 0, \tag{14}$$

for all sample sizes N, T.

(ii) $\|E\|_\infty = O_P(\sqrt{\max(N,T)})$, and $\sum_{k=1}^K \text{rank}(X_k) = O_P(1)$.

Then we have

$$\left\| \widehat{\beta}_\psi - \beta_0 \right\| = O_P(\psi) + O_P\left(\frac{1}{\sqrt{\min(N,T)}} \right), \qquad \left\| \widehat{\beta}_* - \beta_0 \right\| = O_P\left(\frac{1}{\sqrt{\min(N,T)}} \right).$$

Various examples of DGP's for E that satisfy the assumption $\|E\|_\infty = O_P(\sqrt{\max(N,T)})$ can be found in the supplementary appendix S.2 of Moon and Weidner (2017). Loosely speaking, that condition is satisfied as along as the entries E_{it} have zero mean, some appropriately bounded moments, and are not too strongly correlated across i and over t. The condition $\sum_{k=1}^K \text{rank}(X_k) = O_P(1)$ requires all regressors to be low-rank. The interpretation of condition (14) is the same as for condition (13) in Proposition 1, and Theorem 1 is indeed a sample version of that proposition, except that low-rank regressors are required here.

The theorem shows that both $\widehat{\beta}_*$ and $\widehat{\beta}_\psi$, for $\psi = \psi_{NT} = O\left(1/\sqrt{\min(N,T)}\right)$, converge to β_0 at a rate of at least $\sqrt{\min(N,T)}$. The proof of the theorem is provided in the appendix, and is a relatively easy generalization of the proof of Proposition 1. This is because the assumption that all the regressors X_k are low-rank allows to easily decouple the contribution of the high-rank matrix E and the low-rank matrix $\beta \cdot X + \Gamma$ to the penalized objective function $Q_\psi(\beta)$. However, dealing with the contribution of the idiosyncratic errors E becomes more complicated once high-rank regressors are present, as will be explained in the following.

3.2 Consistency Results for General Regressors

The previous subsection considered the case where all regressor matrices X_k are low-rank. We now study situation where all or some of the regressor matrices X_k are high-rank.

3.2.1 Consistency of $\widehat{\beta}_\psi$ and $\widehat{\Gamma}_\psi$

Applying Lemma 1 and the model for Y we have

$$Q_\psi(\beta) = q_\psi\left(\frac{E+\Gamma-(\beta-\beta_0)\cdot X}{\sqrt{NT}}\right) = \sum_{r=1}^{\min(N,T)} q_\psi\left(s_r\left(\frac{E+\Gamma-(\beta-\beta_0)\cdot X}{\sqrt{NT}}\right)\right).$$

The proof strategy for Theorem 1 requires that both Γ and X_k are low-rank, which allows to (approximately) separate off E in this expression for $Q_\psi(\beta)$. But if one of the regressors X_k is a high-rank matrix that proof strategy turns out not to work anymore, because the singular value spectrum of the sum of two high-rank matrices E and X_k does not decompose (or approximately decompose) into a contribution from E and from X_k, but instead all singular values depend on both of those high-rank matrices in a complicated non-linear way.

We therefore now follow a different strategy, where instead of studying the objective function after profiling out Γ, we now explicitly study the properties of the estimator for Γ. Let

$$(\widehat{\beta}_\psi, \widehat{\Gamma}_\psi) = \left[\operatorname*{argmin}_{\beta,\Gamma} \underbrace{\frac{1}{2NT}\|Y-\beta\cdot X-\Gamma\|_2^2}_{=:L(\beta,\Gamma)} + \frac{\psi}{\sqrt{NT}}\|\Gamma\|_1\right].$$

For the results in this subsection we are going to first show consistency of $\widehat{\Gamma}_\psi$, and afterwards use that to obtain consistency of $\widehat{\beta}_\psi$. This is a very different logic than in the preceding section, where consistency of $\widehat{\Gamma}_\psi$ is usually not achieved, because we do not impose any lower bound on ψ. In order to achieve consistency of $\widehat{\Gamma}_\psi$ one requires ψ not be too small. The approach here is much more similar to the machine learning literature (e.g., Negahban, Ravikumar, Wainwright, and Yu 2012), where the matrix that we call Γ is usually the object of interest, and correspondingly a lower bound on the penalization parameter is required. We also follow that literature here by imposing a so-called "restricted strong convexity" condition below, which is critical to show consistency of $\widehat{\Gamma}_\psi$ and consequently of $\widehat{\beta}_\psi$ is the following.

It is convenient to introduce some additional notation: Let vec(A) be the vector that vectorizes the columns of A. Denote mat$(\cdot)$ as the inverse operator of vec$(\cdot)$, so for $a = \text{vec(A)}$ we have mat(a) = A. We use small letters to denote vectorized variables and parameters. Let $y = \text{vec}(Y), x_k = \text{vec}(X_k), \gamma_0 = \text{vec}(\Gamma_0)$, and $e = \text{vec}(E)$. Define $x = (x_1, ..., x_k)$. Using this, we express the model (2) as $y = x\beta_0 + \gamma_0 + e$, where all the summands are NT-vectors, and the least-squares objective function reads $L(\beta, \Gamma) = \frac{1}{2NT}(y - x\beta - \gamma)'(y - x\beta - \gamma)$.

Assumption 1 (Restricted Strong Convexity).

Let $\mathbb{C} = \left\{\Theta \in \mathbb{R}^{N\times T} \mid \|\mathbf{M}_{\lambda_0}\Theta\mathbf{M}_{f_0}\|_1 \leq 3\|\Theta - \mathbf{M}_{\lambda_0}\Theta\mathbf{M}_{f_0}\|_1\right\}$. *We assume that there exists* $\mu > 0$, *independent from* N *and* T, *such that for any* $\theta \in \mathbb{R}^{NT}$ *with* $\text{mat}(\theta) \in \mathbb{C}$ *we have* $\theta'\mathbf{M}_x\theta \geq \mu\,\theta'\theta$, *for all* N, T.

The intuitive interpretation of Assumption 1 is very similar to condition (13) in Proposition 1: The cone $\mathbb{C}$ contains matrices Θ that are close to $\Gamma_0 = \lambda_0 f_0'$, in the sense that the part $\mathbf{M}_{\lambda_0}\Theta\mathbf{M}_{f_0}$ of Θ that cannot be explained by λ_0 and f_0 is small compared to remaining part of Θ, in terms of nuclear norm. The assumption then imposes that all those matrices $\Theta \in \mathbb{C}$ in the cone are sufficiently different from the regressors, in the sense that $\theta = \text{vec}(\Theta)$ cannot be perfectly explained by $x_k = \text{vec}(X_k)$.

Specifically, the condition assumes that the quadratic term, $\frac{1}{2NT}(\gamma-\gamma_0)'\mathbf{M}_x(\gamma-\gamma_0)$, of the profile likelihood function, $\min_\beta L(\beta,\Gamma)$, is bounded below by a strictly convex function, $\frac{\mu}{2NT}(\gamma-\gamma_0)'(\gamma-\gamma_0)$, if $\Theta = \Gamma - \Gamma_0$ belongs in the cone $\mathbb{C}$. Notice that without any restriction on the parameter $\theta = \gamma - \gamma_0$, we cannot find a strictly positive constant $\mu > 0$ such that $\min_\Gamma(\gamma-\gamma_0)'\mathbf{M}_x(\gamma-\gamma_0) \geq \mu(\gamma-\gamma_0)'(\gamma-\gamma_0)$. Assumption 1 imposes that if we restrict the parameter set to be the cone $\mathbb{C}$, then we can find a strictly convex lower bound of the quadratic term of the profile likelihood. Assumption 1 corresponds to the restricted strong convexity condition in Negahban, Ravikumar, Wainwright, and Yu (2012), and it plays the same role as the restricted eigenvalue condition in recent LASSO literature (e.g., see Candes and Tao (2007) and Bickel, Ritov, and Tsybakov (2009)).

Notice that for $R_0 = 0$ we have $\mathbf{M}_{\lambda_0} = \mathbb{I}_N$ and $\mathbf{M}_{f_0} = \mathbb{I}_N$, and therefor $\mathbb{C} = \{0_{N\times T}\}$, implying that Assumption 1 is trivially satisfied for any $\mu > 0$.

The requirement in Assumption 1 is to take a lower bound of $\theta'\mathbf{M}_x\theta$ with strictly convex function. To have some intuition, suppose that the regressor is scalar and assume that $\|X\|_2 = (x'x)^{1/2} = 1$ without loss of generality because the projection operator $\mathbf{M}_x$ is invariant to the scale change. Also assume that $\theta \neq 0$. Then,

$$\begin{aligned}\theta'\mathbf{M}_x\theta = \theta'\theta - (\theta'x)^2 &= (\theta'\theta)\left(1 - \frac{(\theta'x)^2}{\theta'\theta}\right) = (\theta'\theta)\left(x'x - x'\theta(\theta'\theta)^{-1}\theta'x\right) \\ &\geq (\theta'\theta)\min_{\theta\in\mathbb{C}}\|x-\theta\|^2.\end{aligned}$$

In this case, if the limit of the distance between the regressor and the restricted parameter set is positive, Assumption 1 is satisfied if $\mu := \liminf_{N,T}\min_{\theta\in\mathbb{C}}\|x-\theta\|^2$, the distance of the normalized regressor x and convex cone $\mathbb{C}$ is positive. An obvious necessary condition for this is that the normalized regressor does not belong in the cone $\mathbb{C}$, that is,

$$\|\mathbf{M}_{\lambda_0}X\mathbf{M}_{f_0}\|_1 > 3\|X - \mathbf{M}_{\lambda_0}X\mathbf{M}_{f_0}\|_1.$$

For example, if X has an approximate factor structure

$$X = \lambda_x f_x' + E_x,$$

with $E_{x,it} \sim i.i.d.\mathcal{N}(0, \sigma^2)$, then we can use random matrix theory results to show that Assumption 1 is satisfied.

Lemma 2 (Convergence Rate of $\widehat{\Gamma}_\psi$). *Let Assumption 1 holds and assume that*

$$\psi \geq \frac{2}{\sqrt{NT}} \|\text{mat}(\mathbf{M}_x e)\|_\infty. \tag{15}$$

Then we have

$$\frac{1}{\sqrt{NT}} \left\| \widehat{\Gamma}_\psi - \Gamma_0 \right\|_2 \leq \frac{3\sqrt{2R_0}}{\mu} \psi.$$

The lemma shows that once we impose restricted strong convexity and a lower bound on ψ, then we can indeed bound the difference between $\widehat{\Gamma}_\psi$ and Γ_0. This lemma is obviously key to obtain a consistency results for $\widehat{\Gamma}_\psi$. Notice furthermore that

$$\widehat{\beta}_\psi - \beta_0 = (x'x)^{-1} x' \left(y - \widehat{\gamma}_\psi \right) = (x'x)^{-1} \left[x'e - x'(\widehat{\gamma}_\psi - \gamma_0) \right],$$

that is, once we have a consistency result for $\widehat{\Gamma}_\psi$ (or equivalently $\widehat{\gamma}_\psi$), then we can also show consistency of $\widehat{\beta}_\psi$. Using that derivation strategy we obtain the following theorem, which provides a consistency result for both $\widehat{\Gamma}_\psi$ and $\widehat{\beta}_\psi$.

Theorem 2. *Let Assumption 1 hold, and as $N, T \to \infty$ assume that*

(i) $\|E\|_\infty = O_P\left(\max(N,T)^{1/2}\right)$,

(ii) $\frac{1}{\sqrt{NT}} e'x = O_P(1)$,

(iii) $\frac{1}{NT} x'x \to_p \Sigma_x > 0$,

(iv) $\psi = \psi_{NT} \to 0$ *such that* $\sqrt{\min(N,T)}\, \psi_{NT} \to \infty$.

Then we have

$$\frac{1}{\sqrt{NT}} \left\| \widehat{\Gamma}_\psi - \Gamma_0 \right\|_2 \leq O_P(\psi). \qquad \left\| \widehat{\beta}_\psi - \beta_0 \right\| \leq O_P(\psi).$$

The additional regularity conditions imposed in Theorem 2 are weak and quite general. As mentioned before, various examples of E that satisfy (i) can be found in the supplementary

appendix S.2 of Moon and Weidner (2017); these include weakly dependent errors, and nonidentical but independent sequences of errors. Condition (ii) is satisfied if the regressors are exogenous with respect to the error, $\mathbb{E}(x_{it}e_{it}) = 0$, and $x_{it}e_{it}$ are weakly correlated over t and across i so that $\frac{1}{NT}\sum_{i,j=1}^{N}\sum_{t,s=1}^{T}\mathbb{E}(x_{k,it}x_{l,js}e_{it}e_{js})$ is bounded asymptotically. Condition (iii) is the standard non-collinearity condition for the regressors. Condition (iv) restricts the choice of the regularization parameter ψ, which has to converge to zero (as discussed before for identification and consistency of β_0), but not to quickly (if ψ is too small, then $\widehat{\Gamma}_\psi$ picks up all the noise E and cannot be consistent). The conditions (i) and (iv) are sufficient regularity conditions for (15). To see this in more detail, since $\mathrm{mat}(\mathbf{M}_x e) = E - \sum_{k=1}^{K}\widehat{E}_k$ with $\widehat{E}_k = X_k(x_k'x_k)^{-1}(x_k'e)$, we have

$$\begin{aligned}\|\mathrm{mat}(\mathbf{M}_x e)\|_\infty &= \left\|E - \sum_{k=1}^{K}\widehat{E}_k\right\|_\infty \leq \|E\|_\infty + \sum_{k=1}^{K}\|\widehat{E}_k\|_\infty \\ &= \|E\|_\infty + \sum_{k=1}^{K}\left\|\frac{X_k}{\sqrt{NT}}\right\|_\infty \left(\frac{x_k'x_k}{NT}\right)^{-1}\left|\frac{x_k'e}{\sqrt{NT}}\right| \leq \|E\|_\infty\left(1 + \frac{O_P(1)}{\|E\|_\infty}\right).\end{aligned}$$

Then, choosing $\psi \geq \frac{2}{\sqrt{NT}}\|E\|_\infty\left(1 + \frac{O_P(1)}{\|E\|_\infty}\right)$ makes ψ satisfy (15) with probability approaching one, and the rate condition in condition (iv) guarantees this.

Theorem 2 requires $\psi = \psi_{NT}$ to grow faster than $1/\sqrt{\min(N,T)}$. By choosing ψ appropriately we can therefore obtain a convergence rate of $\widehat{\beta}_\psi$ that is just below $\sqrt{\min(N,T)}$, which is essentially the same convergence rate that we found in Section 3.1 for the case of only low-rank regressors.

For the special case $R_0 = 0$ we have $\Gamma_0 = \mathbf{0}_{N\times T}$, and if ψ_{NT} then satisfies (15), one can show that

$$\|\widehat{\Gamma}_\psi - \Gamma_0\|_1 = 0, \tag{16}$$

wpa1, see the appendix for a proof of this. In this case, the regularized estimator of β becomes the pooled OLS estimator, $\widehat{\beta}_\psi = (x'x)^{-1}x'y$, wpa1.

3.2.2 Consistency of $\widehat{\beta}_*$

Here, we establish consistency of the nuclear norm minimization estimator $\widehat{\beta}_*$ for high-rank regressors. For simplicity we only discuss the case of a single regressor ($K = 1$) in the main text, and we simply write X for the $N \times T$ regressor matrix X_1 in this subsection. The general case of multiple regressors ($K > 1$) is discussed in Appendix B.6.

Remember that $\widehat{\beta}_*$ is the minimizer of the objective function $\|Y - \beta\cdot X\|_1 = \|E + (\beta_0 - \beta)X + \Gamma_0\|_1 = \sum_r s_r\left(E + (\beta_0 - \beta)X + \Gamma_0\right)$. Asymptotically separating the contribution of

the low-rank matrix Γ_0 to the singular values of the sum $E + (\beta_0 - \beta)X + \Gamma_0$ is possible under a strong factor assumption.[18] However, characterizing the singular values of the sum of two high-rank matrices $E + (\beta_0 - \beta)X$ requires results from random matrix theory that are usually only shown under relatively strong assumptions on the distribution of the matrix entries. We therefore first provide a theorem under high-level assumptions, and afterwards discuss how to verify those assumptions using results from random matrix theory. We write SVD for "singular value decomposition" in the following.

Theorem 3. *Suppose that $K = 1$, and assume that as $N, T \to \infty$, with $N > T$, we have*

(i) $\|E\|_\infty = O_P(\sqrt{N})$, *and* $\|X\|_\infty = O_P(\sqrt{NT})$.

(ii) There exists a finite positive constant c_{up} *such that* $\frac{1}{T\sqrt{N}}\|E\|_1 \leq \frac{1}{2}c_{\text{up}}$, *wpa1.*

(iii) Let $U_E S_E V_E'$ *be the SVD of* $\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}$.[19] *We assume* $\text{Tr}\left(X' U_E V_E'\right) = O_P(\sqrt{NT})$.

(iv) There exists a constant $c_{\text{low}} > 0$ *such that* $T^{-1}N^{-1/2}\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1 \geq c_{\text{low}}$, *wpa1.*

(v) Let $U_x S_x V_x' = \mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$ *be the SVD of the matrix* $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$. *We assume that there exists* $c_x \in (0,1)$ *such that* $\text{Tr}\left(U_E' U_x S_x U_x' U_E\right) \leq (1 - c_x)\text{Tr}(S_x)$, *wpa1.*

We then have $\sqrt{T}\left(\widehat{\beta}_* - \beta_0\right) = O_P(1)$.

The theorem considers the case $N > T$, because the two panel dimensions are not treated symmetrically in the assumptions and proof of this theorem. Alternatively, we could consider $T < N$, but then we also need to swap N and T, and replace X by X' and E by E' in all the assumptions (the case $T = N$ is ruled out here for technical reasons). For both $N > T$ and $T < N$ the statement of theorem can be written as $\sqrt{\min(N,T)}\left(\widehat{\beta}_* - \beta_0\right) = O_P(1)$, that is, we have the same convergence rate result here for $\widehat{\beta}_*$ as in Theorem 1 above.

Condition (i) in the theorem is quite weak, we already discussed the rate restriction on $\|E\|_\infty$ above, and we have $\|X\|_\infty \leq \|X\|_2 = \sqrt{\sum_i \sum_t X_{it}^2} = O_P(\sqrt{NT})$ as long as $\sup_{it} \mathbb{E}(X_{it}^2)$ is finite. Condition (ii) almost follows from $\|E\|_\infty = O_P(\sqrt{N})$, because we have $\|E\|_1 \leq \text{rank}(E)\,\|E\|_\infty \leq T\|E\|_\infty = O_P(T\sqrt{N})$, and the assumption is only slightly stronger than this in assuming a fixed upper bound with probability approaching one, which can also be verified for many error distributions. Condition (iii) is a high level condition and will be satisfied if

$$\sup_r \mathbb{E}|V_{E,r}' X' U_{E,r}| \leq M, \tag{17}$$

[18] In Moon and Weidner (2015, 2017) we use the perturbation theory of linear operator to do exactly that.

[19] That is, $U_E S_E V_E' = \mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}$ and U_E is an $N \times \text{rank}(\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0})$ matrix of singular vectors, S_E is a $\text{rank}(\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}) \times \text{rank}(\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0})$ diagonal matrix, and V_E is an $T \times \text{rank}(\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0})$ matrix of singular vectors.

for some finite constant M, where $U_{E,r}$ and $V_{E,r}$ are the r^{th} columns of $U_{E,r}$ and V_E, respectively. An example of DGP's of X and E that satisfies condition (17) is given by Assumption LL (i) and (ii) in Moon and Weidner (2015). Condition (iv) rules out "low-rank regressors", for which we typically have $\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1 = O_P(\sqrt{NT})$, but is satisfied generically for "high-rank regressors", for which $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$ has T singular values of order $\sqrt{N}$, so that $\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_1$ is of order $T\sqrt{N}$. Condition (v) requires that the singular vectors of $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}$ are sufficiently different from the singular vectors $\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}$. If X and E are independent, then we expect that assumption to hold quite generally.

4 Post Nuclear Norm Regularized Estimation

In Section 3 we have shown that $\widehat{\beta}_{\psi}$ and $\widehat{\beta}_{*}$ are consistent for β_0 at a $\sqrt{\min(N,T)}$-rate, which is a slower convergence rate than the $\sqrt{NT}$-rate at which the LS estimator $\widehat{\beta}_{\mathrm{LS},R}$ converges to β_0 under appropriate regularity conditions. Our Monte Carlo results in Section 5 confirm this relatively slow rate of convergence of $\widehat{\beta}_{\psi}$ and $\widehat{\beta}_{*}$, that is, those rates are not an artifact of our proof strategy, but are a genuine property of those estimators. In this section we investigate how to establish an estimator that is asymptotically equivalent to the LS estimator, and yet avoids minimizing any non-convex objective function. Our suggestion is to use either $\widehat{\beta}_{\psi}$ or $\widehat{\beta}_{*}$ as a preliminary estimator and iterate estimating $\Gamma_0 = \lambda_0 f_0'$ and β_0 a finite number of times.

The conditions that are usually needed to show that the global minimizer $\widehat{\beta}_{\mathrm{LS},R}$ of the objective function $L_R(\beta)$ is consistent for β_0 (i.e. Assumption A in Bai (2009), or Assumption 4 in Moon and Weidner (2017)) are not required here, because we have already shown consistency of $\widehat{\beta}_{\psi}$ or $\widehat{\beta}_{*}$ under different conditions (our discussion in Section 2.2 highlights those differences). It is therefore convenient to introduce a local version of the LS estimator in (3) as

$$\widehat{\beta}^{\mathrm{local}}_{\mathrm{LS},R} := \underset{\beta \in \mathcal{B}(\beta_0, r_{NT})}{\operatorname{argmin}} L_R(\beta), \qquad \mathcal{B}(\beta_0, r_{NT}) := \left\{ \beta \in \mathbb{R}^K \,:\, \|\beta - \beta_0\| \leq r_{NT} \right\}, \tag{18}$$

where r_{NT} is a sequence of positive numbers such that $r_{NT} \to 0$ and $\sqrt{NT}\, r_{NT} \to \infty$. Those rate conditions guarantee that $\widehat{\beta}^{\mathrm{local}}_{\mathrm{LS},R}$ is an interior point of $\mathcal{B}(\beta_0, r_{NT})$, wpa1, under the assumptions of Theorem 4 below. If the global minimizer $\widehat{\beta}_{\mathrm{LS},R}$ is consistent, then we expect $\widehat{\beta}_{\mathrm{LS},R} = \widehat{\beta}^{\mathrm{local}}_{\mathrm{LS},R}$ wpa1, but $\widehat{\beta}^{\mathrm{local}}_{\mathrm{LS},R}$ is consistent by definition even if $\widehat{\beta}_{\mathrm{LS},R}$ is not. Our goal in the following is to obtain an estimator that is asymptotically equivalent to $\widehat{\beta}^{\mathrm{local}}_{\mathrm{LS},R}$.

For simplicity, we first discuss the case where the number of factors R_0 is known. For unknown R_0 we recommend to use a consistent estimate instead, and we discuss estimation

of R_0 in Section 5 below. Starting from our initial nuclear norm regularized or minimized estimators we consider the following iteration procedure to obtain improved estimates of β:

Step 1: For $s = 0$ set $\widehat{\beta}^{(s)} = \widehat{\beta}_{\psi}$ (or $= \widehat{\beta}_{*}$), the preliminary consistent estimate for β_0.

Step 2: Estimate the factor loadings and the factors of the $s-$step residuals $Y - \widehat{\beta}^{(s)} \cdot X$ by the principle component method:

$$(\widehat{\lambda}^{(s+1)}, \widehat{f}^{(s+1)}) \in \underset{\lambda \in \mathbb{R}^{N \times R_0}, f \in \mathbb{R}^{T \times R_0}}{\operatorname{argmin}} \left\| Y - \widehat{\beta}^{(s)} \cdot X - \lambda f' \right\|_2^2.$$

Step 3: Update the s-stage estimate $\widehat{\beta}^{(s)}$ by

$$\begin{aligned} \widehat{\beta}^{(s+1)} &= \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \min_{g \in \mathbb{R}^{T \times R_0}, h \in \mathbb{R}^{N \times R_0}} \left\| Y - X \cdot \beta - \widehat{\lambda}^{(s+1)} g' + h \, \widehat{f}^{(s+1)\prime} \right\|_2^2 \\ &= \left(x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) x \right)^{-1} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) y. \end{aligned} \tag{19}$$

Step 4: Iterate step 2 and 3 a finite number of times.

The following theorem shows that if the initial estimator $\widehat{\beta}^{(0)}$ is consistent, then $\widehat{\beta}^{(s)}$ gets close to $\widehat{\beta}^{\text{local}}_{\text{LS},R_0}$ as the number of iteration s increases. This result is very similar to the quadratic convergence result of a Newton-Raphson algorithm for minimizing a smooth objective function, and the above iteration step is indeed very similar to performing a Newton-Raphson step to minimize $L_{R_0}(\beta)$.

Theorem 4. *Assume that N and T grow to infinity at the same rate, and that*

(i) $\operatorname{plim}_{N,T\to\infty} (\lambda_0'\lambda_0/N) > 0$, *and* $\operatorname{plim}_{N,T\to\infty} (f_0'f_0/T) > 0$.

(ii) $\|E\|_\infty = O_P\left(\max(N,T)^{1/2}\right)$, *and* $\|X_k\|_\infty = O_P\left((NT)^{1/2}\right)$ *for all* $k \in \{1,\ldots,K\}$.

(iii) $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} x' \left(\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}\right) x > 0$.

(iv) $\frac{1}{\sqrt{NT}} x' \left(\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}\right) e = O_P(1)$.

Then, if the sequence $r_{NT} > 0$ in (18) satisfies $r_{NT} \to 0$ and $\sqrt{NT}\, r_{NT} \to \infty$ we have

$$\sqrt{NT}\left(\widehat{\beta}^{\text{local}}_{\text{LS},R_0} - \beta_0\right) = O_P(1).$$

Assume furthermore that

(iv) $\|\widehat{\beta}^{(0)} - \beta_0\| = O_P(c_{NT})$, *for a sequence $c_{NT} > 0$ such that $c_{NT} \to 0$.*

For $s \in \{1, 2, 3, \ldots\}$ we then have

$$\left\| \widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0} \right\| = O_P\left\{ c_{NT} \left(c_{NT} + \frac{1}{\sqrt{\min(N,T)}} \right)^s \right\}.$$

Here, assumption (i) is a strong factor condition, and is often used in the literature on interactive fixed effects. The conditions in assumption (ii) of the theorem have been discussed in previous sections and are quite weak (remember that $\|X_k\|_\infty \leq \|X_k\|_2 = \sqrt{x_k' x_k}$). Assumption (iii) guarantees that $L_R(\beta)$ is locally convex around β_0 – that condition can equivalently be written as $\text{plim}_{N,T\to\infty} \|\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\|_2 > 0$ for any $\alpha \in \mathbb{R}^K \setminus \{0\}$, which connects more closely to our discussion in Section 2.2. This is a non-collinearity condition on the regressors after profiling out both λ_0 and f_0. Only the true values λ_0 and f_0 appear in that non-collinearity condition, and it is therefore much weaker than the corresponding assumptions required for consistency of $\widehat{\beta}_{\text{LS},R_0}$ in Bai (2009) and Moon and Weidner (2017). Our results from the previous sections show that $\|\widehat{\beta}^{(0)} - \beta_0\| = O_P(c_{NT})$ for both $\widehat{\beta}^{(0)} = \widehat{\beta}_\psi$ and $\widehat{\beta}^{(0)} = \widehat{\beta}_*$, under appropriate assumptions, where c_{NT} is typically either $c_{NT} = 1/\sqrt{\min(N,T)}$ or slightly slower than this, if $\psi = \psi_{NT}$ is chosen appropriately.

The following corollary is an immediate consequence of Theorem 4.

Corollary 1. *Let the assumptions of Theorem 4 hold, and assume that $c_{NT} = o((NT)^{-1/6})$. For $s \in \{2, 3, 4, \ldots\}$ we then have*

$$\sqrt{NT}\left(\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right) = o_P(1), \qquad \sqrt{NT}\left(\widehat{\beta}^{(s)} - \beta_0\right) = O_P(1).$$

The first statement of the corollary shows that if the initial estimators $\widehat{\beta}_\psi$ and $\widehat{\beta}_*$ satisfy typical convergence rates results derived in the previous sections, then the iterated estimator $\widehat{\beta}^{(s)}$ is asymptotically equivalent to $\widehat{\beta}^{\text{local}}_{\text{LS},R_0}$ after $s = 2$ iterations or more. Remember that if $\widehat{\beta}_{\text{LS},R_0}$ is consistent, then we have $\widehat{\beta}^{\text{local}}_{\text{LS},R_0} = \widehat{\beta}_{\text{LS},R_0}$ wpa1, but by showing asymptotic equivalence with $\widehat{\beta}^{\text{local}}_{\text{LS},R_0}$ here we avoid imposing conditions that require consistency of $\widehat{\beta}_{\text{LS},R_0}$.

From the results in Bai (2009) and Moon and Weidner (2017) we also know that $\widehat{\beta}^{\text{local}}_{\text{LS},R_0}$ is asymptotically normally distributed, but potentially with a bias in the limiting distribution. According to the corollary the same is therefore true for $\widehat{\beta}^{(s)}$ for $s \geq 2$. Asymptotic bias corrections could then also be applied to $\widehat{\beta}^{(s)}$, $s \geq 2$, to eliminate the bias in the limiting distribution and allow for inference on β_0. See Bai (2009) and Moon and Weidner (2017) for details.

5 Implementation and Monte Carlo Simulations

To implement the nuclear norm regularized estimator we need to choose the regularization parameter ψ, and for the post estimator $\widehat{\beta}^{(s)}$ we need to determine the number of factors R_0. In this section we suggest a data depenence choice of ψ as well as an estimate of R_0. We assume that an upper bound $R_{\max} \geq R_0$ is known.

Data Dependent Choice of ψ.

We suggest the following procedure to choose ψ.

Step 1: Calculate the nuclear norm minimizing estimator $\widehat{\beta}_*$, and the corresponding residuals

$$\widehat{E}_* = Y - \widehat{\beta}_* \cdot X.$$

Step 2: Choose $R_{\max}$, calculate $R_{\max}$ principal components of $\widehat{E}_*$,

$$\left\{\widehat{\lambda}_{\max}, \widehat{f}_{\max}\right\} \in \underset{\lambda \in \mathbb{R}^{N \times R_{\max}},\, f \in \mathbb{R}^{T \times R_{\max}}}{\operatorname{argmin}} \left\| \widehat{E}_* - \lambda f' \right\|_2^2,$$

and use those to eliminate all the factors in $\widehat{E}_*$, the new residuals are

$$\widetilde{E}_* = \widehat{E} - \widehat{\lambda}_{\max} \widehat{f}'_{\max}.$$

Step 3: Choose

$$\widehat{\psi} = \frac{2\|\widetilde{E}_*\|_\infty}{\sqrt{NT}}.$$

This choice of $\widehat{\psi}$ is motivated by the condition (15) in Lemma 2, which guarantees that ψ is sufficiently large to obtain estimates $\widehat{\Gamma}_\psi$ that are close to Γ_0. Notice also that the nuclear norm minimizing estimator $\widehat{\beta}_*$ in step 1 does not require any regularization parameter to be specified.

Estimation of R_0.

The post nuclear norm regularized estimator introduced above assumes that the number of factors R_0 is known. In practice R_0 needs to be estimated, for example, by applying a consistent estimation method for the number of the factors in a pure factor model to the residuals $\widehat{E}_*$, see e.g. Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013).

For our Monte Carlo simulations below we use an alternative estimation that thresholds the singular values of $\widehat{E}_*$ using the estimate $\widehat{\psi}$ introduced above. Namely, we estimate R_0 by

$$\widehat{R} = \sum_{r=1}^{\min(N,T)} \mathbb{1}\left\{ s_r\left(\widehat{E}_*\right) \geq 2\sqrt{NT}\,\widehat{\psi} \right\}.$$

The motivation behind this estimator is that those singular values of $\widehat{E}_*$ that are significantly larger than $\sqrt{NT}\,\widehat{\psi}$ should correspond to factors, while singular values close to $\sqrt{NT}\,\widehat{\psi}$ and smaller should originate from idiosyncratic noise. The choice of the factor 2 in the formula for $\widehat{R}$ is somewhat arbitrary, any alternative factor larger than one would also be plausible here.

Monte Carlo Results

We generate data from the following linear panel model regression model with two regressor (including the intercept) and two factors:

$$\begin{aligned} Y_{it} &= \beta_{0,1} + \beta_{0,2}\, X_{it} + \sum_{r=1}^{2} \lambda_{0,ir} f_{0,tr} + E_{it}, \\ X_{it} &= 1 + E_{x,it} + \sum_{r=1}^{2} (\lambda_{0,ir} + \lambda_{x,ir})(f_{0,tr} + f_{0,t-1,r}), \end{aligned} \tag{20}$$

where $f_{0,tr} \sim i.i.d.\mathcal{N}(0,1)$; $\lambda_{0,ir}, \lambda_{x,ir} \sim i.i.d.\mathcal{N}(1,1)$; $E_{x,it}, E_{it} \sim i.i.d.\mathcal{N}(0,1)$; and all mutually independent. Table 1 reports the bias and standard deviation for the various estimators for different combinations of N and T.

As shown in Table 1, the nuclear norm regularized estimator $\widehat{\beta}_\psi$ and the nuclear norm minimization estimator $\widehat{\beta}_*$ have biases due to the regularization which vanish slowly as the sample size increases. This confirms that those estimator are indeed not $\sqrt{NT}$ consistent, but only have a $\sqrt{\min(N,T)}$ convergence rate to β_0. The table also shows that the post nuclear norm regularized estimators $\widehat{\beta}^{(s)}$ quickly reduces the bias, and essentially agrees with the LS estimator (which is a consistent estimator in this MC design) after two iterations, as the theory predicts. The columns ATL(1) - ALT(5) in that table contain the results for an alternative bias corrected estimator that is presented in the appendix. It turns out that the alternative bias correction method is less effective in reducing the bias, and we therefore do not discuss it in the main text. Our recommendation in practice for inference on β_0 is the iteration procedure for $\widehat{\beta}^{(s)}$ explained in the previous section.

(N/T)	POLS	LS	NNmin	NNpen	POST(1)	POST(2)	POST(3)	ALT(1)	ALT(2)	ALT(3)	ALT(4)	ALT(5)
(25/25)												
BIAS	0.2379	0.0508	0.1447	0.1712	0.0695	0.0527	0.0510	0.1005	0.0677	0.0520	0.0442	0.0403
STD	(0.0241)	(0.0613)	(0.0259)	(0.0237)	(0.0479)	(0.0598)	(0.0612)	(0.0287)	(0.0341)	(0.0387)	(0.0418)	(0.0438)
(100/25)												
BIAS	0.2382	0.0603	0.1349	0.1706	0.0750	0.0614	0.0603	0.0998	0.0684	0.0542	0.0476	0.0446
STD	(0.0150)	(0.0612)	(0.0159)	(0.0143)	(0.0479)	(0.0601)	(0.0611)	(0.0218)	(0.0301)	(0.0358)	(0.0393)	(0.0413)
(100/100)												
BIAS	0.2395	0.0000	0.1024	0.1504	0.0209	0.0008	0.0000	0.0656	0.0285	0.0123	0.0053	0.0022
STD	(0.0105)	(0.0061)	(0.0102)	(0.0095)	(0.0061)	(0.0061)	(0.0061)	(0.0085)	(0.0067)	(0.0058)	(0.0056)	(0.0057)
(400/25)												
BIAS	0.2388	0.0546	0.1339	0.1695	0.0704	0.0558	0.0547	0.0967	0.0644	0.0499	0.0432	0.0401
STD	(0.0111)	(0.0589)	(0.0139)	(0.0117)	(0.0456)	(0.0579)	(0.0588)	(0.0190)	(0.0274)	(0.0332)	(0.0365)	(0.0384)
(400/100)												
BIAS	0.2397	0.0000	0.0941	0.1348	0.0175	0.0006	0.0000	0.0515	0.0195	0.0073	0.0028	0.0010
STD	(0.0058)	(0.0026)	(0.0076)	(0.0065)	(0.0037)	(0.0026)	(0.0026)	(0.0050)	(0.0034)	(0.0028)	(0.0026)	(0.0026)
(400/400)												
BIAS	0.2399	0.0000	0.0672	0.1091	0.0114	0.0002	0.0000	0.0326	0.0095	0.0028	0.0008	0.0002
STD	(0.0050)	(0.0013)	(0.0042)	(0.0042)	(0.0017)	(0.0013)	(0.0013)	(0.0027)	(0.0016)	(0.0014)	(0.0013)	(0.0013)

Table 1: Monte Carlo results based on 1000 repetitions for the design specified in display (20). Reported are the bias and standard deviation for the pooled OLS estimator (POLS), the least squares estimator with $R_0 = 2$ factors (LS), the nuclear norm minimizing estimator $\widehat{\beta}_*$ (NNpen), the nuclear norm penalized estimator with $\psi = \widehat{\psi}$ (NNpen), the post estimator $\widehat{\beta}^{(s)}$ for $s = 1, 2, 3$ iterations and using $R = \widehat{R}$ factors (POST(s)), and the alternative bias correction method (see the appendix) using $R = \widehat{R}$ factors and $s = 1, 2, 3, 4, 5$ iterations.

6 Extension to Single Index Models

We now consider the following generalization of the penalized LS estimator,

$$\left(\widehat{\beta}_\psi, \widehat{\Gamma}_\psi\right) \in \underset{\beta\in\mathbb{R}^K,\, \Gamma\in\mathbb{R}^{N\times T}}{\operatorname{argmin}} Q_\psi(\beta,\Gamma), \quad Q_\psi(\beta,\Gamma) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m_{it}\left(X_{it}'\beta + \Gamma_{it}\right) + \frac{\psi}{\sqrt{NT}}\left\|\Gamma\right\|_1,$$

where $m_{it}(z) := m(W_{it}, z)$ is a known convex function of the single index $z \in \mathbb{R}$, which also depends on the observed variables W_{it}. The single index $X_{it}'\beta + \Gamma_{it}$ has the same structure as the conditional mean of the linear model (2), and for $W_{it} = Y_{it}$ and $m_{it}(z) = \frac{1}{2}(Y_{it} - z)^2$ we obtain the penalized LS estimator that was studied in previous sections. The nuclear norm penalty term is unchanged.

Let $\overline{m}_{it}(z) = \mathbb{E}(m_{it}(z)|X)$ be the expected objective function, conditional on $X = \{X_{it} : i = 1, \ldots, N; t = 1, \ldots, T\}$,[20] and denote derivatives of $m_{it}(z)$ and $\overline{m}_{it}(z)$ with respect to z by $\partial_z m_{it}(z)$, $\partial_z \overline{m}_{it}(z)$, $\partial_{z^2}\overline{m}_{it}(z)$, etc. Let $z_{it}^0 = X_{it}'\beta_0 + \Gamma_{0,it}$ be the index evaluated at the true parameters. Let $\mathcal{W}$ denote the domain of W_{it}. We make the following assumptions on the objective function.

[20]Remember that we consider Γ_0 as non-random, that is, all expectations are implicitly conditional on Γ_0 as well. Also, we condition on all the observed X here, implying that we only consider strictly exogenous regressors in this section, but in principle the results could be extended to dynamic models.

Assumption 2. *Let $\mathcal{Z} \subset \mathbb{R}$ be such that $\cup_{i,t}[z_{it}^0 - \epsilon, z_{it}^0 + \epsilon] \subset \mathcal{Z}$, for some $\epsilon > 0$. Assume:*

(i) W_{it} is independently distributed across i and over t, conditional on X.

(ii) The objective function $m(w, z)$ is convex in z, and once continuously differentiable in z almost everywhere in $\mathcal{W} \times \mathcal{Z}$. For any function $z_{it} = z_{it}(X) \in \mathcal{Z}$ the first derivative $\partial_z m_{it}(z_{it})$ exists almost surely, and satisfies $\max_{i,t,N,T} \mathbb{E}\left\{ [\partial_z m_{it}(z_{it})]^4 \,\middle|\, X \right\} < \infty$.

(iii) $\overline{m}_{it}(z)$ is four times continuously differentiable in $\mathcal{Z}$, with derivatives bounded uniformly over i, t, N, T, $\mathcal{Z}$. There exists $b > 0$ such that $\min_{i,t,N,T} \min_{z \in \mathcal{Z}} \partial_{z^2} \overline{m}_{it}(z) \geq b$.

(iv) $\partial_z \overline{m}_{it}(z_{it}^0) = 0$, for all i, t.

Here, the last assumption crucially connects the distribution of W_{it} conditional on X_{it} with the chosen objective function $m_{it}(z)$. For the LS case we have $\partial_z m_{it}(z_{it}^0) = E_{it}$, and Assumption 2 then becomes the familiar mean independence condition $\mathbb{E}(E_{it}|X) = 0$. This condition excludes a predetermined regressor. Some further examples for data generating processes and corresponding objective functions are

(a) Maximum likelihood: Let Y_{it} conditional on X have probability mass or density function $p(y|z_{it}^0)$, set $W_{it} = Y_{it}$ and $m_{it}(z) = -\log p(Y_{it}|z)$, and assume that $m_{it}(z)$ is strictly convex in z and three times continuously differentiable. A concrete example is a binary choice probit model, where $p(y|z) = \mathbb{1}(y = 1)\Phi(z) + \mathbb{1}(y = 0)[1 - \Phi(z)]$, and $\Phi(.)$ is the cdf of $\mathcal{N}(0, 1)$.

(b) Weighted Least Squares: Let outcomes Y_{it} be generated from the linear model (2) with $\mathbb{E}(E_{it}|X_{it}, S_{it}) = 0$, and let $m_{it}(z) = \frac{1}{2} S_{it}(Y_{it} - z)^2$, and $W_{it} = (Y_{it}, S_{it})$. Here, the $S_{it} \geq 0$ are observed weights for each observation. A special case is $S_{it} \in \{0, 1\}$, where S_{it} is an indicator of a missing outcome Y_{it}.

(c) Quantile Regression: Let outcomes Y_{it} be generated from the linear model (2), but instead of the mean restriction for E_{it} we impose the quantile restriction $\mathbb{E}[\mathbb{1}(E_{it} \leq 0)|X_{it}] = \tau$, and we let $m_{it}(z) = \rho_\tau(Y_{it} - z)$, and $W_{it} = Y_{it}$, where $\rho_\tau(u) = u \cdot [\tau - \mathbb{1}(u < 0)]$ is the quantile regression objective function, and $\tau \in (0, 1)$ is a chosen quantile of interest.

Some additional regularity conditions are needed to guarantee that those examples satisfy Assumption 2. For many models (e.g. quantile regressions and binary choice likelihood) we have $\lim_{z \to \pm\infty} \partial_{z^2} \overline{m}_{it}(z) = 0$. Then, the lower bound on $\partial_{z^2} \overline{m}_{it}(z)$ in Assumption 2(iii) will require us to impose that $\mathcal{Z}$ is a bounded set, which can be guaranteed by assuming

that X_{it} and $\Gamma_{0,it}$ are uniformly bounded. Apart from that it is straightforward to verify Assumption 2 under standard regularity conditions for the respective model. Notice also that Assumption 2(ii) is formulated with the quantile regression case in mind, where $\partial_z m_{it}(z_{it}) = \tau - \mathbb{1}(Y_{it} - z < 0)$ is not well-defined at $z_{it} = Y_{it}$, but that is a probability zero event for continuously distributed Y_{it}.

In the following theorem we show that $\widehat{\beta} - \beta_0 = O_P(\psi^{1/2})$ for $\psi \to 0$ with $\psi\sqrt{NT} \to \infty$ and the regressors have a generalized factor structure. We present in a special case where there exists a single regressor (i.e., $K = 1$) and the regressor is strictly exogenous for technical simplicity.

Theorem 5. *Let Assumption 2 be satisfied. Let $N, T \to \infty$, $\psi \to 0$, and $\sqrt{NT}\psi \to \infty$. Let $K = 1$. Assume that*

(i) $\|\Gamma_0\|_1 = O(\sqrt{NT})$.

(ii) The regressor can be decomposed as $X = X^{(1)} + X^{(2)}$ such that $\|X^{(1)}\|_1 = o_P(\sqrt{NT}\,\psi^{-1/2})$, and $\|X^{(2)}\|_\infty = o_P(\sqrt{NT}\,\psi^{1/2})$.

(iii) $W := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}^{(2)})^2$ *satisfies* $W \to_P W_\infty > 0$.

Then we have $\widehat{\beta}_\psi - \beta_0 = O_P(\psi^{1/2})$.

Condition (i) of the theorem is a restriction on the growth rate of the nuclear norm of Γ_0, which was not required for the results in Section 3, where we assumed only that $R_0 = \text{rank}(\Gamma_0)$ is fixed. However, this condition (i) imposes only an upper bound on the growth of Γ_0, it allows that Γ_0 contains both strong factors and weak factors.[21]

Condition (ii) is satisfied if the regressor has a generalized factor structure,

$$X = \underbrace{\lambda_x f_x'}_{X^{(1)}} + \underbrace{E_x}_{X^{(2)}},$$

where $\|\lambda_x f_x'\|_1 = O_P(\sqrt{NT})$ and $\|E_x\|_\infty = O_P(\sqrt{\max(N,T)}$, and we have $\psi \to 0$ with $\min(N,T)\psi \to \infty$.

The proof of Theorem 5 is presented in the appendix, where we also discuss how the result could in principle be extended to $K > 1$ regressors, which requires some additional technical restrictions. Notice also that the convergence rate of $\psi^{1/2}$ in Theorem 5 is different from the convergence rate ψ obtained in Section 3, but this likely an artifact of our proof strategy for Theorem 5. Finally, the analog of the nuclear-norm minimizing estimator $\widehat{\beta}_*$

[21] For a discussion of weak factors we refer to Onatski (2012).

to non-linear models is given by $\lim_{\psi \to 0} \widehat{\beta}_\psi$ (limit for fixed N, T), but we do not provide results for that limiting estimator here. The goal of this section was not to fully discuss the non-linear case, but to highlight the potential of the nuclear norm penalization approach beyond the linear model that is main focus of this paper.

7 Conclusions

In this paper we analyze two new estimation methods for interactive fixed effect panel regressions that are based on convex objective functions: (i) nuclear norm penalized estimation, and (ii) nuclear norm minimizing estimation. The resulting estimators can also be applied in situations where the LS estimator may not be consistent, in particular when low-rank regressors are present and the true number of factors is unknown. We provide consistency and convergence rate results for the new estimators of the regression coefficients, and we show how to use them as a preliminary estimator to achieve asymptotic equivalence to the local version of the LS estimator. We have focused on the linear model with homogenous coefficients, which is a natural starting point to understand the usefulness of nuclear norm penalization approach for panel regression models, but there are several ongoing extensions, including developing a unified method to deal with non-linear models, heterogeneous coefficients, treatment effect estimation, nonparametric sieve estimation, and high-dimensional regressors, see Section 6 above, and also Athey, Bayati, Doudchenko, Imbens, and Khosravi (2017) and Chernozhukov, Hansen, Liao, and Zhu (2018).

References

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 105(490), 493–505.

——— (2015): "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59(2), 495–510.

Abadie, A., and J. Gardeazabal (2003): "The economic costs of conflict: A case study of the Basque Country," *American economic review*, 93(1), 113–132.

Adams, C. (2017): "Identification of Treatment Effects from Synthetic Controls," .

Ahn, S. C., and A. R. Horenstein (2013): "Eigenvalue ratio test for the number of factors," *Econometrica*, 81(3), 1203–1227.

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

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013): "Panel data models with multiple time-varying individual effects," *Journal of Econometrics*, 174(1), 1–14.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017): "Matrix Completion Methods for Causal Panel Data Models," *Working Paper, Stanford University*.

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

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

——— (2017): "Principal Components and Regularized Estimation of Factor Models," *arXiv preprint arXiv:1708.08137*.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, pp. 1705–1732.

Candes, E., and T. Tao (2007): "The Dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, pp. 2313–2351.

Chamberlain, G., and M. J. Moreira (2009): "Decision theory applied to a linear panel data model," *Econometrica*, 77(1), 107–133.

Chan, M. K., and S. Kwok (2016): "Policy Evaluation with Interactive Fixed Effects," Discussion paper.

Chen, M. (2014): "Estimation of nonlinear panel models with multiple unobserved effects," *Warwick Economics Research Paper Series No. 1120*.

Chen, M., I. Fernandez-Val, and M. Weidner (2014): "Nonlinear panel models with interactive effects," *arXiv preprint arXiv:1412.5647*.

Chen, X., and D. Pouzo (2012): "Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals," *Econometrica*, 80(1), 277–321.

Cheng, X., Z. Liao, and F. Schorfheide (2016): "Shrinkage estimation of high-dimensional factor models with structural instabilities," *The Review of Economic Studies*, 83(4), 1511–1543.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2018): "Inference For Heterogeneous Effects Using Low-Rank Estimations," *arXiv preprint arXiv:1812.08089*.

Gobillon, L., and T. Magnac (2016): "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 98(3), 535–551.

Gobillon, L., and F.-C. Wolff (2017): "The local effects of an innovation: Evidence from the French fish market," .

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56(6), 1371–95.

Hsiao, C., H. S. Ching, and S. K. Wan (2012): "A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong kong with mainland China," *Journal of Applied Econometrics*, 27(5), 705–740.

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

Li, D., J. Qian, and L. Su (2016): "Panel data models with interactive fixed effects and multiple structural breaks," *Journal of the American Statistical Association*, 111(516), 1804–1819.

Li, K. (2018): "Inference for Factor Model Based Average Treatment Effects," .

Lu, X., and L. Su (2016): "Shrinkage estimation of dynamic panel data models with interactive fixed effects," *Journal of Econometrics*, 190(1), 148–175.

Moon, H. R., and M. Weidner (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83(4), 1543–1579.

——— (2017): "Dynamic linear panel regression models with interactive fixed effects," *Econometric Theory*, 33(1), 158–195.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012): "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statisical Science*, 27(4), 538–557.

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

——— (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, 168(2), 244–258.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

Piracha, M., M. Tani, and G. Tchuente (2017): "Immigration Policy and Remittance Behaviour," .

Powell, D. (2017): "Synthetic control estimation beyond case studies: Does the minimum wage reduce employment?," .

Recht, B., M. Fazel, and P. A. Parrilo (2010): "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, 52(3), 471–501.

Rohde, A., and A. B. Tsybakov (2011): "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, 39(2), 887–930.

Su, L., Z. Shi, and P. C. Phillips (2016): "Identifying latent structures in panel data," *Econometrica*, 84(6), 2215–2264.

A Appendix

A.1 An Example of a Non-convex LS Profile Objective Function

As an example for a non-convex LS profile objective function we consider the following linear model with one regressor and two factors:

$$Y_{it} = \beta_0 X_{it} + \sum_{r=1}^{2} \lambda_{0,ir} f_{0,tr} + E_{it},$$
$$X_{it} = 0.04 E_{x,it} + \lambda_{0,i1} f_{0,t2} + \lambda_{x,i} f_{x,t},$$

where

$$\lambda_{0,i} = \binom{\lambda_{0,i1}}{\lambda_{0,i2}} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right),\ f_{0,t} = \binom{f_{0,t1}}{f_{0,t2}} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right),$$

and $\lambda_{x,i} \sim iid\ 2\chi^2(1)$, $f_{x,t} \sim iid\ 2\chi^2(1)$, $E_{x,it}, E_{it} \sim i.i.d.\mathcal{N}(0,1)$, and $\{\lambda_{0,i}\}$, $\{f_{0,t}\}$, $\{\lambda_{x,i}\}$, $\{f_{x,t}\}$, $\{E_{x,it}\}, \{E_{it}\}$ are all independent of each other. For $(N,T) = (200, 200)$, we generate the panel data for (Y_{it}, X_{it}), and plot the LS objective function (3) in Figure 1, which is discussed in the main text.

A.2 Alternative Bias Correction

In this section, we discuss an alternative bias reduction method used in the Monte Carlo simulations in Section 5. The alternative method reduces the bias of the score function of the regularized least squares objective function $Q_\psi(\beta)$. We introduce the procedure in a heuristic way *without* presenting a rigorous proof. We have implemented this alternative method in our Monte Carlo simulations, and while it indeed improves the of the nuclear-norm penalized estimates (see Table 1), it does not perform better than the iteration method described in Section 4.

Recall that $L_R(\beta, \Gamma) = \frac{1}{2NT} \|Y - \beta \cdot X - \Gamma\|_2^2$, where $\Gamma = \lambda f'$. Define

$$\widehat{\Gamma}_R(\beta) := \underset{\Gamma:\text{rank}(\Gamma)\leq R}{\text{argmin}}\ L_R(\beta, \Gamma).$$

We can write

$$L_R(\beta) = L_R(\beta, \widehat{\Gamma}_R(\beta)) = \frac{1}{2} \sum_{r=R+1}^{\min(N,T)} s_r \left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right)^2.$$

Let $\widehat{\Gamma}_\psi(\beta) = \text{argmin}_\Gamma \, Q_\psi(\beta, \Gamma)$, and

$$\widehat{R}(\beta, \psi) := \sum_{r=1}^{\min(N,T)} \mathbb{I}\{s_r(Y - \beta \cdot X) \geq \sqrt{NT}\psi\} = \text{rank}\left(\widehat{\Gamma}_\psi(\beta)\right).$$

Suppose that we choose ψ such that

$$\widehat{R}(\beta, \psi) = R_0 \tag{A.1}$$

Then, in view of (8) and (10), we write the profile objective function of the regularized least squares as

$$\begin{aligned} Q_\psi(\beta) &:= q_\psi\left(\frac{Y - \beta X}{\sqrt{NT}}\right) \\ &= \frac{1}{2}\sum_{r=R_0+1}^{\min(N,T)} s_r\left(\frac{Y - \beta X}{\sqrt{NT}}\right)^2 + \psi \sum_{r=1}^{R(\beta,\psi)} s_r\left(\frac{Y - \beta X}{\sqrt{NT}}\right) - \frac{1}{2}\psi^2 R(\beta, \psi) \\ &= L_{R_0}(\beta) + \psi \sum_{r=1}^{R_0} s_r\left(\frac{Y - \beta X}{\sqrt{NT}}\right) - \frac{1}{2}\psi^2 R_0 \\ &= L_{R_0}(\beta) + \psi \left\|\widehat{\Gamma}_{R_0}(\beta)\right\|_1 - \frac{1}{2}\psi^2 R_0. \end{aligned} \tag{A.2}$$

This shows that the term $\left\|\widehat{\Gamma}_{R_0}(\beta)\right\|_1$ is the main source of the regularization bias. We suggest to approximate $\left\|\widehat{\Gamma}_{R_0}(\beta)\right\|_1$ as follows,

$$\|\Gamma_0 - (\beta - \beta_0) \cdot X + E\|_1 \approx \|\Gamma_0\|_1 - (\beta - \beta_0)' B_{NT}, \tag{A.3}$$

where $B_{NT} = (B_{NT,1}, ..., B_{NT,K})'$, with

$$B_{NT,k} := \frac{1}{\sqrt{NT}} \text{Tr}\left[(\lambda_0'\lambda_0)^{-1/2}\lambda_0' X_k f_0 (f_0' f_0)^{-1/2}\right].$$

From (A.2) and (A.3) we expect that $Q_\psi(\beta) + \psi(\beta - \beta_0)B_{NT}$ should be a good approximation to $L_{R_0}(\beta)$. This heuristics suggests that we may reduce the bias of the nuclear norm regularized estimation by modifying the objective function.

For this, suppose that $\widehat{\psi}$ is a data dependent choice of ψ that satisfies the condition (A.1). Let $\widehat{R}$ be a consistent estimator of R_0. Let $\widehat{\beta}^{(0)}_{\text{alt}}$ be an preliminary estimator. For example, $\widehat{\beta}^{(0)} = \widehat{\beta}_{\widehat{\psi}}$ or $\widehat{\beta}^{(0)} = \widehat{\beta}_*$.

For $s = 0, 1, 2, ...,$ define

$$(\widehat{\lambda}^{(s)}, \widehat{f}^{(s)}) \in \underset{\lambda \in \mathbb{R}^{N \times \widehat{R}}, f \in \mathbb{R}^{T \times \widehat{R}}}{\operatorname{argmin}} \left\| Y - \widehat{\beta}_{\text{alt}}^{(s)} \cdot X - \lambda f' \right\|_2^2,$$

and

$$\widehat{B}_{NT,k}^{(s)} := \frac{1}{\sqrt{NT}} \operatorname{Tr} \left[(\widehat{\lambda}^{(s)\prime} \widehat{\lambda}^{(s)})^{-1/2} \widehat{\lambda}^{(s)\prime} X_k \widehat{f}^{(s)} (\widehat{f}^{(s)\prime} \widehat{f}^{(s)})^{-1/2} \right].$$

We modify the nuclear norm regularized objective function as

$$Q_{\widehat{\psi}}^{\text{bc,s+1}}(\beta) := Q_{\widehat{\psi}}(\beta) + \widehat{\psi}(\beta - \widehat{\beta}_{\text{alt}}^{(s)}) \widehat{B}_{NT}^{(s)}$$

and update the estimator as

$$\widehat{\beta}_{\text{alt}}^{(s+1)} := \underset{\beta}{\operatorname{argmin}}\, Q_{\widehat{\psi}}^{\text{bc,s+1}}(\beta).$$

B Supplementary Appendix

B.1 Proofs for Section 2.1

For matrix A, let the singular value decomposition of A be given by $A = U_A S_A V_A'$, where $S_A = \mathrm{diag}(s_1, \ldots, s_q)$, with $q = \mathrm{rank}(A)$.

Lemma S.1. *For any $\psi > 0$ we have*

$$\min_{\Gamma} \left(\frac{1}{2} \|A - \Gamma\|_2^2 + \psi \|\Gamma\|_1 \right) = q_\psi(A),$$
$$\operatorname*{argmin}_{\Gamma} \left(\frac{1}{2} \|A - \Gamma\|_2^2 + \psi \|\Gamma\|_1 \right) = U_A \mathrm{diag}((s_1 - \psi)_+, \ldots, (s_q - \psi)_+) V_A',$$

where the minimization is over all matrices Γ of the same size as A and $(s)_+ = \max(0, s)$.

Proof of Lemma S.1. The dependence of the various quantities on ψ is not made explicit in this proof. Let $\mathcal{Q}(A) = \min_\Gamma \left(\frac{1}{2} \|A - \Gamma\|_2^2 + \psi \|\Gamma\|_1 \right)$. A possible value for Γ is $\Gamma^* = U_A S^* V_A'$, where $S^* = \mathrm{diag}(s_1^*, \ldots, s_q^*)$ and $s_r^* = \max(0, s_r - \psi)$, and therefore we have

$$\begin{aligned} \mathcal{Q}(A) &\leq \frac{1}{2} \|A - \Gamma^*\|_2^2 + \psi \|\Gamma^*\|_1 = \frac{1}{2} \|S_A - S^*\|_2^2 + \psi \|S_\psi^*\|_1 \\ &= \sum_{r=1}^{q} \left[\frac{1}{2} (s_r - s_r^*)^2 + \psi s_r^* \right] = \sum_{r=1}^{q} q_\psi(s_r) = q_\psi(A). \end{aligned}$$

The nuclear norm satisfies $\|\Gamma\|_1 = \max_{\|B\|_\infty \leq 1} \mathrm{Tr}(\Gamma' B)$. A possible value for B is $B^* = U_A D^* V_A'$, where $D^* = \mathrm{diag}(d_1^*, \ldots, d_q^*)$ and $d_r^* = \min(1, \psi^{-1} s_r)$, which indeed satisfies $\|B^*\|_\infty = \|D^*\|_\infty = \max_r |d_r^*| \leq 1$, and therefore we have

$$\begin{aligned} \mathcal{Q}(A) &\geq \min_{\Gamma} \left[\frac{1}{2} \|A - \Gamma\|_2^2 + \psi \,\mathrm{Tr}(\Gamma' B^*) \right] = \frac{1}{2} \|A - (A - \psi B^*)\|_2^2 + \psi \,\mathrm{Tr}[(A - \psi B^*)' B^*] \\ &= \psi \,\mathrm{Tr}(A' B^*) - \frac{\psi^2}{2} \|B^*\|_2^2 = \psi \,\mathrm{Tr}(S_A' D^*) - \frac{\psi^2}{2} \|D^*\|_2^2 \\ &= \sum_{r=1}^{q} \left[\psi \, s_r d_r^* - \frac{\psi^2}{2} (d_r^*)^2 \right] = \sum_{r=1}^{q} q_\psi(s_r) = q_\psi(A), \end{aligned}$$

where in the second step we found and plugged in the minimizing $\Gamma = A - \psi B^*$. By combing the above upper and lower bound on $\mathcal{Q}(A)$ we obtain $\mathcal{Q}(A) = q_\psi(A)$, which is the first statement of the lemma. Since $\operatorname{argmin}_\Gamma \left(\frac{1}{2} \|A - \Gamma\|_2^2 + \psi \|\Gamma\|_1 \right)$ is unique, we deduce that $\Gamma^* = U_A S^* V_A'$ is the minimizing value, which is the second statement in the lemma. □

Proof of Lemma 1. The lemma follows from the first statement of Lemma S.1 by replacing

A and Γ in Lemma S.1 with $\frac{Y-\beta\cdot X}{\sqrt{NT}}$ and $\frac{1}{\sqrt{NT}}\Gamma$, respectively. □

B.2 Proofs for Section 2.2

The function $q_\psi(s)$ that appears in Lemma S.1 was defined in (8). We now define a similar function $g_\psi : [0,\infty) \to [0,\infty)$ by $g_\psi(s) = \psi^{-1} q_\psi(s)$ for $\psi > 0$, and $g_\psi(s) = s$ for $\psi = 0$, that is, we have

$$g_\psi(s) := \begin{cases} \frac{1}{2\psi}\, s^2, & \text{for } s < \psi, \\ s - \frac{\psi}{2}, & \text{for } s \geq \psi, \end{cases} \tag{S.1}$$

and for matrices A we define $g_\psi(A) := \sum_{r=1}^{\text{rank}(A)} g_\psi(s_r(A))$. Using Lemma S.1 and the definition of the nuclear norm we can write

$$g_\psi(A) = \begin{cases} \min_\Gamma \left(\frac{1}{2\,\psi} \|A-\Gamma\|_2^2 + \|\Gamma\|_1 \right), & \text{for } \psi > 0, \\ \|A\|_1, & \text{for } \psi = 0. \end{cases} \tag{S.2}$$

As already discussed in the main text, it is natural to rescale the profiled nuclear norm penalized objective function by ψ^{-1}, because it then has a non-trivial limit as $\psi \to 0$. Using g_ψ instead q_ψ therefore helps to clarify the scaling with ψ in various expressions. The following lemma summarizes some properties of the function $g_\psi(A)$, which are useful for the subsequent proofs.

Lemma S.2. *Let A and B be $N \times T$ matrices, λ be an $N \times R_1$ matrix, and f be a $T \times R_2$ matrix. We then have*

(i) $g_\psi(A) \geq \|A\|_1 - \frac{\psi}{2}\text{rank}(A)$.

(ii) $g_\psi(A+B) \leq g_\psi(A) + \|B\|_1$, *and* $g_\psi(A+B) \geq g_\psi(A) - \|B\|_1$.

(iii) $g_\psi(A) \geq g_\psi(\mathbf{M}_\lambda\, A\, \mathbf{M}_f) + g_\psi(\mathbf{P}_\lambda\, A\, \mathbf{P}_f)$.

Proof of Lemma S.2. # Part (i): From the definition of $g_\psi(s)$ in (S.1) one finds $g_\psi(s) \geq s - \frac{\psi}{2}$ for all $s \geq 0$. We thus obtain

$$g_\psi(A) = \sum_{r=1}^{\text{rank}(A)} g_\psi(s_r(A)) \geq \sum_{r=1}^{\text{rank}(A)} \left[s_r(A) - \frac{\psi}{2} \right] = \|A\|_1 - \frac{\psi}{2}\text{rank}(A).$$

Part (ii): For $\psi = 0$ this is just the triangle inequality for the nuclear norm. For $\psi > 0$

we use (S.2) to write

$$g_\psi(A+B) = \min_\Gamma \left(\frac{1}{2\,\psi} \|A+B-\Gamma\|_2^2 + \|\Gamma\|_1 \right) = \min_\Gamma \left(\frac{1}{2\,\psi} \|A-\Gamma\|_2^2 + \|\Gamma+B\|_1 \right)$$
$$\leq \min_\Gamma \left(\frac{1}{2\,\psi} \|A-\Gamma\|_2^2 + \|\Gamma\|_1 \right) + \|B\|_1 = g_\psi(A) + \|B\|_1.$$

where in the second step we reparameterized $\Gamma \mapsto \Gamma + B$ in the minimization problem, in the third step we used the triangle inequality for the nuclear norm, and in the final step we employed again (S.2). We have thus shown the first statement of this part. The second statement is obtained from the first statement by replacing $B \mapsto -B$ and $A \mapsto A + B$.

Part (iii): We first show the result for $\psi = 0$. Let $\mathbf{M}_\lambda \, A \, \mathbf{M}_f = U_1 S_1 V_1'$ and $\mathbf{P}_\lambda \, A \, \mathbf{P}_f = U_2 S_2 V_2'$ be the singular value decompositions of those $N \times T$ matrices. We then have $\|\mathbf{M}_\lambda \, A \, \mathbf{M}_f\|_1 = \mathrm{Tr}[V_1(\mathbf{M}_\lambda \, A \, \mathbf{M}_f)U_1']$ and $\|\mathbf{P}_\lambda \, A \, \mathbf{P}_f\|_1 = \mathrm{Tr}[V_2(\mathbf{P}_\lambda \, A \, \mathbf{P}_f)U_2']$. Furthermore, we have $g_0(A) = \|A\|_1 = \max_{\|C\|\leq 1} \mathrm{Tr}(C'A)$. By choosing $C^* = U_1 V_1' + U_2 V_2'$ we obtain

$$\|A\|_1 \geq \mathrm{Tr}(C^{*\prime} A) = \mathrm{Tr}[V_1(\mathbf{M}_\lambda \, A \, \mathbf{M}_f)U_1'] + \mathrm{Tr}[V_2(\mathbf{P}_\lambda \, A \, \mathbf{P}_f)U_2'] = \|\mathbf{M}_\lambda \, A \, \mathbf{M}_f\|_1 + \|\mathbf{P}_\lambda \, A \, \mathbf{P}_f\|_1 , \tag{S.3}$$

which is the statement of part (iii) of the lemma for $\psi = 0$. For $\psi > 0$ we find

$$\begin{aligned}
g_\psi(A) &= \min_\Gamma \left(\frac{1}{2\,\psi} \|A-\Gamma\|_2^2 + \|\Gamma\|_1 \right) \geq \min_\Gamma \left(\frac{1}{2\,\psi} \|A-\Gamma\|_2^2 + \|\mathbf{M}_\lambda \Gamma \mathbf{M}_f\|_1 + \|\mathbf{P}_\lambda \Gamma \mathbf{P}_f\|_1 \right) \\
&= \min_\Gamma \Bigg[\frac{1}{2\,\psi} \left(\|\mathbf{M}_\lambda (A-\Gamma)\mathbf{M}_f\|_2^2 + \|\mathbf{P}_\lambda (A-\Gamma)\mathbf{P}_f\|_2^2 + \|\mathbf{P}_\lambda (A-\Gamma)\mathbf{M}_f\|_2^2 + \|\mathbf{M}_\lambda (A-\Gamma)\mathbf{P}_f\|_2^2 \right) \\
&\qquad\qquad + \|\mathbf{M}_\lambda \Gamma \mathbf{M}_f\|_1 + \|\mathbf{P}_\lambda \Gamma \mathbf{P}_f\|_1 \Bigg] \\
&= \min_\Gamma \left[\frac{1}{2\,\psi} \left(\|\mathbf{M}_\lambda (A-\Gamma)\mathbf{M}_f\|_2^2 + \|\mathbf{P}_\lambda (A-\Gamma)\mathbf{P}_f\|_2^2 \right) + \|\mathbf{M}_\lambda \Gamma \mathbf{M}_f\|_1 + \|\mathbf{P}_\lambda \Gamma \mathbf{P}_f\|_1 \right] \\
&\geq \min_\Gamma \left(\frac{1}{2\,\psi} \|\mathbf{M}_\lambda (A-\Gamma)\mathbf{M}_f\|_2^2 + \|\mathbf{M}_\lambda \Gamma \mathbf{M}_f\|_1 \right) + \min_\Gamma \left(\frac{1}{2\psi} \|\mathbf{P}_\lambda (A-\Gamma)\mathbf{P}_f\|_2^2 + \|\mathbf{P}_\lambda \Gamma \mathbf{P}_f\|_1 \right) \\
&\geq \min_\Gamma \left(\frac{1}{2\,\psi} \|\mathbf{M}_\lambda A \mathbf{M}_f - \Gamma\|_2^2 + \|\Gamma\|_1 \right) + \min_\Gamma \left(\frac{1}{2\,\psi} \|\mathbf{P}_\lambda A \mathbf{P}_f - \Gamma\|_2^2 + \|\Gamma\|_1 \right) \\
&= g_\psi(\mathbf{M}_\lambda A \mathbf{M}_f) + g_\psi(\mathbf{P}_\lambda A \mathbf{P}_f),
\end{aligned}$$

where in the first step we used (S.2); in the second step we used (S.3) with A replaced by Γ; in the third step we decomposed $\|A-\Gamma\|_2^2$ into four parts; in the fourth step we used that the minimization over Γ implies that $\|\mathbf{P}_\lambda (A-\Gamma)\mathbf{M}_f\|_2^2 = 0$ and $\|\mathbf{M}_\lambda (A-\Gamma)\mathbf{P}_f\|_2^2 = 0$ at the optimum, because the components $\mathbf{P}_\lambda \Gamma \mathbf{M}_f$ and $\mathbf{M}_\lambda \Gamma \mathbf{P}_f$ of Γ appear nowhere else in the

objective function, so that choosing $\mathbf{P}_\lambda \Gamma \mathbf{M}_f = \mathbf{P}_\lambda A \mathbf{M}_f$ and $\mathbf{M}_\lambda \Gamma \mathbf{P}_f = \mathbf{M}_\lambda A \mathbf{P}_f$ is optimal; the fifth step is obvious (it is actually an equality, which is less obvious, but not required for our argument); in the sixth step we replaced $\mathbf{M}_\lambda \Gamma \mathbf{M}_f$ and $\mathbf{P}_\lambda \Gamma \mathbf{P}_f$ by an unrestricted Γ in the minimization problems, which can only make the minimizing values smaller (again, this is actually an equality, but $\leq$ is sufficient to show here); and the final step again employs (S.2). We have thus shown the desired result. □

Before presenting the next lemma it is useful to introduce some further notation. For $\beta \in \mathbb{R}^K$ let $\Delta\beta := \beta - \beta_0$. Let λ_X be an $N \times R_{\mathrm{c}}$ matrix such that the column span of λ_X equals the columns span of the $N \times TK$ matrix $[X_1, \ldots, X_K]$. Analogously, let f_X be an $T \times R_{\mathrm{r}}$ matrix such that the column span of f_X equals the columns span of the $T \times NK$ matrix $[X_1', \ldots, X_K']$.

Lemma S.3. *Let model* (1) *hold. Then, the penalized profiled objective function* $Q_\psi(\beta)$ *defined in* (5) *satisfied, for all* $\beta \in \mathbb{R}^K$*, and all* $\psi > 0$*,*

$$\begin{aligned}\frac{Q_\psi(\beta) - Q_\psi(\beta_0)}{\psi} \geq g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) - \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\beta \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 - \frac{\psi}{2}\operatorname{rank}(\Gamma_0) \\ - \left\|\frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1 - \left\|\frac{E - \mathbf{M}_{[\lambda_0,\lambda_X]} E \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1.\end{aligned}$$

For $\psi = 0$ *the same bound holds if one replaces* $\psi^{-1}\left[Q_\psi(\beta) - Q_\psi(\beta_0)\right]$ *by its* $\psi \to 0$ *limit* $\left\|(Y - \beta \cdot X)/\sqrt{NT}\right\|_1 - \left\|(Y - \beta_0 \cdot X)/\sqrt{NT}\right\|_1$.

Proof of Lemma S.3. We have

$$\begin{aligned}g_\psi\left(\frac{Y - \beta \cdot X}{\sqrt{NT}}\right) &= g_\psi\left(\frac{\Gamma_0 - \Delta\beta \cdot X + E}{\sqrt{NT}}\right) \\ &\geq g_\psi\left(\frac{\mathbf{P}_{[\lambda_0,\lambda_X]}(\Gamma_0 - \Delta\beta \cdot X + E)\mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}}\right) + g_\psi\left(\frac{\mathbf{M}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right) \\ &= g_\psi\left(\frac{\Gamma_0 - \Delta\beta \cdot X}{\sqrt{NT}} + \frac{\mathbf{P}_{[\lambda_0,\lambda_X]} E \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}}\right) + g_\psi\left(\frac{\mathbf{M}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right) \\ &\geq g_\psi\left(\frac{\Gamma_0 - \Delta\beta \cdot X}{\sqrt{NT}}\right) - \left\|\frac{\mathbf{P}_{[\lambda_0,\lambda_X]} E \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1 + g_\psi\left(\frac{\mathbf{M}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right).\end{aligned}$$

Here, we first plugged in the model for Y, then used part (iii) of Lemma S.2 with $\lambda = [\lambda_0, \lambda_X]$ and $f = [f_0, f_X]$, and in the final step used part (ii) of Lemma S.2. In the same way we

obtain

$$
\begin{aligned}
g_\psi\left(\frac{\Gamma_0 - \Delta\beta\cdot X}{\sqrt{NT}}\right) &\geq g_\psi\left(\frac{\mathbf{P}_{\lambda_0}(\Gamma_0 - \Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right) + g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) \\
&= g_\psi\left(\frac{\Gamma_0}{\sqrt{NT}} - \frac{\mathbf{P}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right) + g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) \\
&\geq g_\psi\left(\frac{\Gamma_0}{\sqrt{NT}}\right) - \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 + g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) \\
&\geq \left\|\frac{\Gamma_0}{\sqrt{NT}}\right\|_1 - \frac{\psi}{2}\operatorname{rank}(\Gamma_0) - \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 + g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right),
\end{aligned}
$$

where in the last step we also used part (i) of Lemma S.2. Furthermore, we find

$$
\begin{aligned}
g_\psi\left(\frac{Y - \beta_0\cdot X}{\sqrt{NT}}\right) = g_\psi\left(\frac{E + \Gamma_0}{\sqrt{NT}}\right) &= g_\psi\left(\frac{\mathbf{M}_{[\lambda_0,\lambda_X]}E\mathbf{M}_{[f_0,f_X]} + \left(E - \mathbf{M}_{[\lambda_0,\lambda_X]}E\mathbf{M}_{[f_0,f_X]}\right) + \Gamma_0}{\sqrt{NT}}\right) \\
&\leq g_\psi\left(\frac{\mathbf{M}_{[\lambda_0,\lambda_X]}E\mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right) + \left\|\frac{E - \mathbf{M}_{[\lambda_0,\lambda_X]}E\mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1 + \left\|\frac{\Gamma_0}{\sqrt{NT}}\right\|_1,
\end{aligned}
$$

where we used part (ii) of Lemma S.2 and the triangle inequality for the nuclear norm. Combining the inequalities in the last three displays gives

$$
\begin{aligned}
g_\psi\left(\frac{Y - \beta\cdot X}{\sqrt{NT}}\right) - g_\psi\left(\frac{Y - \beta_0\cdot X}{\sqrt{NT}}\right) \geq g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) &- \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 - \frac{\psi}{2}\operatorname{rank}(\Gamma_0) \\
&- \left\|\frac{\mathbf{P}_{[\lambda_0,\lambda_X]}E\mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1 - \left\|\frac{E - \mathbf{M}_{[\lambda_0,\lambda_X]}E\mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}}\right\|_1.
\end{aligned}
$$

The derivation so far was valid for all $\psi \geq 0$. For $\psi = 0$ the left hand side of the last display simply is $\left\|(Y - \beta\cdot X)/\sqrt{NT}\right\|_1 - \left\|(Y - \beta_0\cdot X)/\sqrt{NT}\right\|_1$. For $\psi > 0$ we have, by (S.2),

$$
\frac{Q_\psi(\beta) - Q_\psi(\beta_0)}{\psi} = g_\psi\left(\frac{Y - \beta\cdot X}{\sqrt{NT}}\right) - g_\psi\left(\frac{Y - \beta_0\cdot X}{\sqrt{NT}}\right),
$$

so that we have shown the statement of the lemma. □

Lemma S.4. *Let model (2) hold, and let* $\mathbb{E}(E_{it} \mid X) = 0$, *and* $\mathbb{E}\left(E_{it}^2 \mid X\right) < \infty$, *for all* i, t. *Then we have, for all* $\psi > 0$,

$$
g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) - \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\bar{\beta}_\psi\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 \leq \frac{\psi}{2}\operatorname{rank}(\Gamma_0).
$$

Proof of Lemma S.4. Using the model and the assumptions on E_{it} in the proposition we

find

$$\begin{aligned}\mathbb{E}\left[\|Y-\beta\cdot X-\Gamma\|_2^2\,\middle|\,X\right] &= \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\left(\Gamma_{0,it}-\Gamma_{it}-X_{it}'\,\Delta\beta+E_{it}\right)^2\middle|\,X\right]\\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\Gamma_{0,it}-\Gamma_{it}-X_{it}'\,\Delta\beta\right)^2+\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left(E_{it}^2\middle|\,X\right)\\ &= \|\Gamma_0-\Gamma-\Delta\beta\cdot X\|_2^2+\mathbb{E}\left(\|E\|_2^2\middle|\,X\right),\end{aligned}$$

where the expectation is also implicitly conditional on Γ_0, because Γ_0 is treated as non-random throughout the whole paper. Because $\mathbb{E}\left(\|E\|_2^2\middle|\,X\right)$ is just a constant that does not depend on the parameters β and Γ, we can thus rewrite the definition of $\bar{\beta}_\psi$ in (12) as

$$\bar{\beta}_\psi=\operatorname*{argmin}_\beta \overline{Q}_\psi(\beta),\qquad \overline{Q}_\psi(\beta):=\min_\Gamma\left\{\frac{1}{2NT}\,\|\Gamma_0-\Gamma-\Delta\beta\cdot X\|_2^2+\frac{\psi}{\sqrt{NT}}\,\|\Gamma\|_1\right\}.$$

We can obtain $\overline{Q}_\psi(\beta)$ from the profiled objective function $Q_\psi(\beta)$ that was defined in (5) by simply setting $E=0$ in the model (2). The bound on $\psi^{-1}\left[Q_\psi(\beta)-Q_\psi(\beta_0)\right]$ in Lemma S.3 is therefore applicable to $\overline{Q}_\psi(\beta)$ if we just set $E=0$ in that lemma. We thus have, for all $\beta\in\mathbb{R}^K$,

$$\frac{\overline{Q}_\psi(\beta)-\overline{Q}_\psi(\beta_0)}{\psi}\geq g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right)-\left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\beta\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1-\frac{\psi}{2}\operatorname{rank}(\Gamma_0).$$

We have $Q_\psi(\bar{\beta}_\psi)-Q_\psi(\beta_0)\leq 0$, because $\bar{\beta}_\psi$ minimizes $Q_\psi(\beta)$, and combining this with the result in the last display gives the statement of the lemma. □

Proof of Proposition 1. Let

$$c=\min_{\{\alpha\in\mathbb{R}^K:\,\|\alpha\|=1\}}C(\alpha),\qquad C(\alpha)=\frac{\|\mathbf{M}_{\lambda_0}(\alpha\cdot X)\mathbf{M}_{f_0}\|_1-\|\mathbf{P}_{\lambda_0}(\alpha\cdot X)\mathbf{P}_{f_0}\|_1}{\sqrt{NT}}.$$

Using the absolute homogeneity of the nuclear norm this definition implies that for any $\alpha\in\mathbb{R}^K$ we have

$$c\,\|\alpha\|\leq\left\|\frac{\mathbf{M}_{\lambda_0}(\alpha\cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right\|_1-\left\|\frac{\mathbf{P}_{\lambda_0}(\alpha\cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1. \tag{S.4}$$

Since the ball $\left\{\alpha\in\mathbb{R}^K\,:\,\|\alpha\|=1\right\}$ is a compact set, and $C(\alpha)$ is a continuous function there exists a value $\alpha^*\in\left\{\alpha\in\mathbb{R}^K\,:\,\|\alpha\|=1\right\}$ where the minimum is attained, that is, $c=C(\alpha^*)$. By the assumption on the regressors in Proposition 1 we thus have $c=C(\alpha^*)>0$.

Next, applying part (i) of Lemma S.2 we obtain

$$g_\psi\left(\frac{\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right) \geq \left\|\frac{\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right\|_1 - \frac{\psi}{2}\text{rank}\left[\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}\right], \quad \text{(S.5)}$$

and also using Lemma S.4 we thus find that

$$\begin{aligned}\left\|\frac{\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right\|_1 - \left\|\frac{\mathbf{P}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1 &\leq \frac{\psi}{2}\left\{\text{rank}(\Gamma_0) + \text{rank}\left[\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}\right]\right\} \\ &\leq \frac{\psi}{2}\left\{\text{rank}(\Gamma_0) + \max_{\alpha\in\mathbb{R}^K}\text{rank}\left[\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\right]\right\}.\end{aligned}$$

From this and (S.4) with $\alpha = \Delta\bar{\beta}_\psi$ we obtain for any $\psi > 0$ that[22]

$$\left\|\bar{\beta}_\psi - \beta_0\right\| \leq \frac{\psi}{2c}\left\{\text{rank}(\Gamma_0) + \max_{\alpha\in\mathbb{R}^K}\text{rank}\left[\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\right]\right\}, \quad \text{(S.6)}$$

and therefore $\left\|\bar{\beta}_\psi - \beta_0\right\| = O(\psi)$, as $\psi \to 0$. □

B.3 Proofs for Section 3.1

Lemma S.5. *Let* $R_{\text{c}} := \text{rank}([X_1, \ldots, X_K])$ *and* $R_{\text{r}} := \text{rank}([X_1', \ldots, X_K'])$. *Assume that*

$$C := \min_{\{\alpha\in\mathbb{R}^K\,:\,\|\alpha\|=1\}}\left\|\frac{\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}}\right\|_1 - \left\|\frac{\mathbf{P}_{\lambda_0}(\alpha \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}}\right\|_1$$

satisfies $C > 0$. *Then we have, for all* $\psi > 0$,

$$\left\|\widehat{\beta}_\psi - \beta_0\right\| \leq \frac{1}{C}\left[\left(\frac{\psi}{2} + \frac{\|E\|_\infty}{\sqrt{NT}}\right)[R_0 + \min(R_{\text{c}}, R_{\text{r}})] + \frac{\|E\|_\infty}{\sqrt{NT}}\left(2\,R_0 + R_{\text{c}} + R_{\text{r}}\right)\right],$$

and

$$\left\|\widehat{\beta}_* - \beta_0\right\| \leq \frac{1}{C}\frac{\|E\|_\infty}{\sqrt{NT}}\left[3\,R_0 + R_{\text{c}} + R_{\text{r}} + \min(R_{\text{c}}, R_{\text{r}})\right].$$

Proof of Lemma S.5. By definition we have $Q_\psi(\widehat{\beta}_\psi) - Q_\psi(\beta_0) \leq 0$. Combining this with

[22] The bound (S.6) is sufficient for our purposes since we ultimately consider the limit $\psi \to 0$ here, but for a fixed value of ψ (and N, T) this bound is potentially very crude if high-rank regressors X_k are present. From Lemma S.4 one could then obtain a sharper bound on $\bar{\beta}_\psi - \beta_0$ by not using part (i) of Lemma S.2 to simplify $g_\psi\left[\left(\mathbf{M}_{\lambda_0}(\Delta\bar{\beta}_\psi \cdot X)\mathbf{M}_{f_0}\right)/\sqrt{NT}\right]$.

Lemma S.3 and equation (S.5), and writing $\operatorname{rank}(\Gamma_0) = R_0$, we obtain

$$0 \geq \left\| \frac{\mathbf{M}_{\lambda_0}(\Delta\widehat{\beta}_\psi \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}} \right\|_1 - \left\| \frac{\mathbf{P}_{\lambda_0}(\Delta\widehat{\beta}_\psi \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}} \right\|_1 - \frac{\psi}{2}\left\{ R_0 + \max_{\alpha\in\mathbb{R}^K} \operatorname{rank}\left[\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\right] \right\} - \left\| \frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 - \left\| \frac{E - \mathbf{M}_{[\lambda_0,\lambda_X]} E \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 .$$

The definition of c in the theorem together with the absolute homogeneity of the nuclear norm implies

$$c \left\| \Delta\widehat{\beta}_\psi \right\| \leq \left\| \frac{\mathbf{M}_{\lambda_0}(\Delta\widehat{\beta}_\psi \cdot X)\mathbf{M}_{f_0}}{\sqrt{NT}} \right\|_1 - \left\| \frac{\mathbf{P}_{\lambda_0}(\Delta\widehat{\beta}_\psi \cdot X)\mathbf{P}_{f_0}}{\sqrt{NT}} \right\|_1 .$$

We have

$$\max_{\alpha\in\mathbb{R}^K} \operatorname{rank}\left[\mathbf{M}_{\lambda_0}(\alpha \cdot X)\mathbf{M}_{f_0}\right] \leq \max_{\alpha\in\mathbb{R}^K} \operatorname{rank}(\alpha \cdot X) \leq \min(R_{\mathrm{c}}, R_{\mathrm{r}}),$$

because we have $\alpha \cdot X = [X_1, \ldots, X_K](\alpha \otimes \mathbb{I}_T)$, and therefore $\operatorname{rank}(\alpha \cdot X) \leq R_{\mathrm{c}}$, and also $(\alpha \cdot X)' = [X_1', \ldots, X_K'](\alpha \otimes \mathbb{I}_N)$, and therefore $\operatorname{rank}(\alpha \cdot X) \leq R_{\mathrm{r}}$.

We also have

$$\begin{aligned}
\left\| \frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 &\leq \left\| \frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_\infty \operatorname{rank}\left(\mathbf{P}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}\right) \\
&\leq \frac{\|E\|_\infty}{\sqrt{NT}} \min\left\{ \operatorname{rank}\left(\mathbf{P}_{[\lambda_0,\lambda_X]}\right), \operatorname{rank}\left(\mathbf{P}_{[f_0,f_X]}\right) \right\} \\
&= \frac{\|E\|_\infty}{\sqrt{NT}} \min\left\{ R_0 + R_{\mathrm{c}}, R_0 + R_{\mathrm{r}} \right\} = \frac{\|E\|_\infty}{\sqrt{NT}} \left[R_0 + \min(R_{\mathrm{c}}, R_{\mathrm{r}}) \right],
\end{aligned}$$

and similarly

$$\begin{aligned}
\left\| \frac{E - \mathbf{M}_{[\lambda_0,\lambda_X]} E \mathbf{M}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 &= \left\| \frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E}{\sqrt{NT}} + \frac{\mathbf{M}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 \\
&\leq \left\| \frac{\mathbf{P}_{[\lambda_0,\lambda_X]}\, E}{\sqrt{NT}} \right\|_1 + \left\| \frac{\mathbf{M}_{[\lambda_0,\lambda_X]}\, E\, \mathbf{P}_{[f_0,f_X]}}{\sqrt{NT}} \right\|_1 \\
&\leq \frac{\|E\|_\infty}{\sqrt{NT}} \operatorname{rank}\left(\mathbf{P}_{[\lambda_0,\lambda_X]}\right) + \frac{\|E\|_\infty}{\sqrt{NT}} \operatorname{rank}\left(\mathbf{P}_{[f_0,f_X]}\right) \\
&= \frac{\|E\|_\infty}{\sqrt{NT}} \left(2\, R_0 + R_{\mathrm{c}} + R_{\mathrm{r}}\right).
\end{aligned}$$

Combining the above inequalities gives the finite sample bound in the theorem,

$$c\left\|\widehat{\beta}_{\psi}-\beta_0\right\| \leq \left(\frac{\psi}{2}+\frac{\|E\|_{\infty}}{\sqrt{NT}}\right)\left[R_0+\min(R_{\mathrm{c}},R_{\mathrm{r}})\right]+\frac{\|E\|_{\infty}}{\sqrt{NT}}\left(2\,R_0+R_{\mathrm{c}}+R_{\mathrm{r}}\right),$$

and the same bound holds for $\widehat{\beta}_{*}$ if we set $\psi = 0$, because all bounds above, including Lemma S.3 are applicable for $\psi = 0$ as well. Finally, the asymptotic statements in the theorem are immediate corollaries of the finite sample bounds. □

Proof of Theorem 1. The theorem follows immediately from Lemma S.5, because our assumptions guarantee that $C \geq c > 0$ (and therefore $1/C = O(1)$), $R_0 = O_P(1)$, $R_{\mathrm{c}} = O_P(1)$, $R_{\mathrm{r}} = O_P(1)$, and

$$\frac{\|E\|_{\infty}}{\sqrt{NT}} = O_P\left(\frac{1}{\sqrt{\min(N,T)}}\right).$$

□

B.4 Proofs for Section 3.2.1

Lemma S.6. *Suppose that A and B are two matrices with ranks of A and B are* rank(A) *and* rank(B)*, respectively.*

(i) $\|A\|_{\infty} \leq \|A\|_2 \leq \|A\|_1 \leq \sqrt{\mathrm{rank(A)}}\|A\|_2 \leq \mathrm{rank(A)}\|A\|_{\infty}$.

(ii) $\|AB\|_{\infty} \leq \|A\|_{\infty}\|B\|_{\infty}$.

(iii) $\|AB\|_2 \leq \|A\|_{\infty}\|B\|_2 \leq \|A\|_2\|B\|_2$.

(iv) If $AB' = 0$ *and* $A'B = 0$*, then* $\|A+B\|_{\infty} = \max(\|A\|_{\infty}, \|B\|_{\infty})$.

(v) If $A'B = 0$ *(or equivalently* $B'A = 0$*), then* $\|A+B\|_{\infty}^2 \leq \|A\|_{\infty}^2 + \|B\|_{\infty}^2$.

Recall that the rank of $\Gamma_0 = \lambda_0 f_0'$ is R_0, which is fixed. Throughout the rest of the appendix, we use the following singular value decomposition of Γ_0,

$$\Gamma_0 = USV', \tag{S.7}$$

where $U \in \mathbb{R}^{N\times R_0}$ with $U'U = \mathbf{I}_{R_0}$, $V \in \mathbb{R}^{T\times R_0}$ with $V'V = \mathbf{I}_{R_0}$, S is the $R_0 \times R_0$ diagonal matrix of singular values of Γ_0.

Suppose that f_0 is normalized as $\frac{1}{T}f_0'f_0 = \mathbf{I}_{R_0}$. Then, we have

$$f_0 = \sqrt{T}V \quad \lambda_0 = \frac{US}{\sqrt{T}}.$$

Some further notation:

$$L(\beta,\Gamma)=\frac{1}{2NT}\|Y-\beta\cdot X-\Gamma\|_2^2,\quad Q_\psi(\beta,\Gamma)=\frac{1}{2NT}\|Y-\beta\cdot X-\Gamma\|_2^2+\frac{\psi}{\sqrt{NT}}\|\Gamma\|_1.$$

Let

$$Q_\psi(\Gamma):=\inf_\beta Q_\psi(\beta,\Gamma),\qquad\qquad L(\Gamma):=\inf_\beta L(\beta,\Gamma).$$

These are the profile objective functions of $Q_\psi(\beta,\Gamma)$ and $L(\beta,\Gamma)$, respectively, which concentrate out parameter the β. We also use the notation $\Theta:=\Gamma-\Gamma_0$ and $\theta:=\mathrm{vec}(\Theta)$.

Proof of Lemma 2.

Step 1: Use (15) to show $\widehat{\Theta}_\psi\in\mathbb{C}$

By definition, we have

$$\begin{aligned}0&\geq Q_\psi(\Gamma_0+\widehat{\Theta}_\psi)-Q_\psi(\Gamma_0)\\&=L(\Gamma_0+\widehat{\Theta}_\psi)-L(\Gamma_0)+\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_\psi\|_1-\|\Gamma_0\|_1\right),\end{aligned}$$

where $\widehat{\Theta}_\psi:=\widehat{\Gamma}_\psi-\Gamma_0$. Let $\widehat{\theta}_\psi:=\mathrm{vec}(\widehat{\Theta}_\psi)$, $\widehat{\Theta}_{\psi,1}:=\mathbf{M}_{U_0}\widehat{\Theta}_\psi\mathbf{M}_{V_0}$ and $\widehat{\Theta}_{\psi,2}:=\widehat{\Theta}_\psi-\mathbf{M}_{U_0}\widehat{\Theta}_\psi\mathbf{M}_{V_0}$. Then

$$\begin{aligned}L(\Gamma_0+\widehat{\Theta}_\psi)-L(\Gamma_0)&=\frac{1}{2NT}\widehat{\theta}_\psi'\mathbf{M}_x\widehat{\theta}_\psi-\frac{1}{NT}e'\mathbf{M}_x\widehat{\theta}_\psi\\&\geq-\frac{1}{NT}e'\mathbf{M}_x\widehat{\theta}_\psi\\&=-\frac{1}{NT}\,\mathrm{Tr}(\widehat{\Theta}_\psi'\,\mathrm{mat}(\mathbf{M}_xe))\\&\geq-\frac{\|\widehat{\Theta}_\psi\|_1}{\sqrt{NT}}\,\frac{\|\mathrm{mat}(\mathbf{M}_xe)\|_\infty}{\sqrt{NT}}\\&\geq-\frac{\psi}{2}\frac{\|\widehat{\Theta}_\psi\|_1}{\sqrt{NT}}\\&\geq-\frac{\psi}{2}\frac{\|\widehat{\Theta}_{\psi,1}\|_1}{\sqrt{NT}}-\frac{\psi}{2}\frac{\|\widehat{\Theta}_{\psi,2}\|_1}{\sqrt{NT}}.\end{aligned}$$

Here the first inequality holds since $\widehat{\theta}_\psi'\mathbf{M}_x\widehat{\theta}_\psi\geq0$, the second inequality holds by the Hölder inequality, the third inequality holds by (15), and the last inequality holds by the triangle

inequality. We furthermore have

$$\begin{aligned}
&\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_\psi\|_1-\|\Gamma_0\|_1\right)\\
&=\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_{\psi,1}+\widehat{\Theta}_{\psi,2}\|_1-\|\Gamma_0\|_1\right)\\
&\geq\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_{\psi,1}\|_1-\|\Gamma_0\|_1\right)-\frac{\psi}{\sqrt{NT}}\|\widehat{\Theta}_{\psi,2}\|_1\\
&=\frac{\psi}{\sqrt{NT}}\|\widehat{\Theta}_{\psi,1}\|_1-\frac{\psi}{\sqrt{NT}}\|\widehat{\Theta}_{\psi,2}\|_1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
0&\geq L(\Gamma_0+\widehat{\Theta}_\psi)-L(\Gamma_0)+\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_\psi\|_1-\|\Gamma_0\|_1\right)\\
&\geq-\frac{\psi}{2}\frac{\|\widehat{\Theta}_{\psi,1}\|_1}{\sqrt{NT}}-\frac{\psi}{2}\frac{\|\widehat{\Theta}_{\psi,2}\|_1}{\sqrt{NT}}+\psi\frac{\|\widehat{\Theta}_{\psi,1}\|_1}{\sqrt{NT}}-\psi\frac{\|\widehat{\Theta}_{\psi,2}\|_1}{\sqrt{NT}}\\
&=\frac{\psi}{2}\frac{1}{\sqrt{NT}}\left(\|\widehat{\Theta}_{\psi,1}\|_1-3\|\widehat{\Theta}_{\psi,2}\|_1\right).
\end{aligned}$$

Thus, we have

$$\widehat{\Theta}_\psi\in\mathbb{C}:=\left\{B\in\mathbb{R}^{N\times T}\mid\|\mathbf{M}_UB\mathbf{M}_V\|_1\leq3\|B-\mathbf{M}_UB\mathbf{M}_V\|_1\right\}.$$

Step 2: Also use Assumption 1 to show the final result: Using Assumption 1 and the same derivation as above, we find

$$\begin{aligned}
Q_\psi(\Gamma_0+\widehat{\Theta}_\psi)-Q_\psi(\Gamma_0)&=\frac{1}{2NT}\widehat{\theta}_\psi'\mathbf{M}_x\widehat{\theta}_\psi-\frac{1}{NT}e'\mathbf{M}_x\widehat{\theta}_\psi+\frac{\psi}{\sqrt{NT}}\left(\|\Gamma_0+\widehat{\Theta}_\psi\|_1-\|\Gamma_0\|_1\right)\\
&\geq\frac{\mu}{2NT}\|\widehat{\Theta}_\psi\|_2^2+\frac{\psi}{2}\frac{1}{\sqrt{NT}}\left(\|\widehat{\Theta}_{\psi,1}\|_1-3\|\widehat{\Theta}_{\psi,2}\|_1\right)\\
&\geq\frac{\mu}{2NT}\|\widehat{\Theta}_\psi\|_2^2-\frac{3\,\psi}{2}\frac{1}{\sqrt{NT}}\|\widehat{\Theta}_{\psi,2}\|_1.
\end{aligned}$$

Because $0\geq Q_\psi(\Gamma_0+\widehat{\Theta}_\psi)-Q_\psi(\Gamma_0)$ we thus have

$$\frac{\mu}{2NT}\|\widehat{\Theta}_\psi\|_2^2-\frac{3\,\psi}{2}\frac{1}{\sqrt{NT}}\|\widehat{\Theta}_{\psi,2}\|_1\leq0.$$

Since the rank of $\widehat{\Theta}_{\psi,2}$ is at most $2R_0$ (e.g., see Recht, Fazel, and Parrilo (2010)), we have

$$\|\widehat{\Theta}_{\psi,2}\|_1\leq\sqrt{2R_0}\|\widehat{\Theta}_{\psi,2}\|_2$$

and we also have

$$\|\widehat{\Theta}_{\psi,2}\|_2 \leq \|\widehat{\Theta}_{\psi}\|_2.$$

Therefore,

$$\frac{1}{NT}\|\widehat{\Theta}_{\psi}\|_2^2 - \frac{3\,\psi\,\sqrt{2R_0}}{\mu}\frac{1}{\sqrt{NT}}\|\widehat{\Theta}_{\psi}\|_2 \leq 0,$$

and

$$\frac{\|\widehat{\Theta}_{\psi}\|_2}{\sqrt{NT}} \leq \frac{3\,\sqrt{2R_0}\,\psi}{\mu}.$$

□

Proof of Theorem 2.

Part (i). Part (i) follows by Lemma 2 and the condition on ψ in Theorem 2.

Part (ii). Let $\widehat{\beta}(\Gamma) = (x'x)^{-1}x'(y-\gamma)$. Then, by definition we have

$$\widehat{\beta}_{\psi} - \beta_0 \;:=\; \widehat{\beta}(\widehat{\Gamma}_{\psi}) - \beta_0 = \left(\frac{1}{NT}x'x\right)^{-1}\left(\frac{1}{NT}x'e - \frac{1}{NT}x'(\widehat{\gamma}_{\psi} - \gamma_0)\right).$$

Under the assumption of the theorem we have $\left(\frac{1}{NT}x'x\right)^{-1} = O_P(1)$ and $\frac{1}{NT}e'x = O_P(\frac{1}{\sqrt{NT}})$. Also, by Part (a) we have

$$\begin{aligned}\left\|\frac{1}{NT}x'(\widehat{\gamma}_{\psi} - \gamma_0)\right\|_2 &\leq \frac{1}{\sqrt{NT}}\|X\|_2\frac{1}{\sqrt{NT}}\|\widehat{\Gamma}_{\psi} - \Gamma_0\|_2 \\ &= O_P(1)\psi.\end{aligned}$$

Combining these, we can deduce the required result for Part (b). □

Proof of (16).

Since $\mathbf{M}_x$ is positive semi-definite, $|e'M_x\widehat{\gamma}_{\psi}| \leq \|\widehat{\Gamma}_{\psi}\|_1\|\mathrm{mat}(\mathbf{M}_x e)\|_{\infty}$ by Hölder inequality,

and $\Gamma_0 = 0$, we have

$$\begin{aligned}
0 &\geq Q(\widehat{\Gamma}_\psi) - Q(\Gamma_0) \\
&= \frac{1}{2NT}(\widehat{\gamma}_\psi - \gamma_0)'\mathbf{M}_x(\widehat{\gamma}_\psi - \gamma_0) - \frac{1}{NT}e'\mathbf{M}_x(\widehat{\gamma}_\psi - \gamma_0) + \frac{\psi}{\sqrt{NT}}\|\widehat{\Gamma}_\psi - \Gamma_0\|_1 \\
&\geq -\frac{1}{NT}e'\mathbf{M}_x(\widehat{\gamma}_\psi - \gamma_0) + \frac{\psi}{\sqrt{NT}}\|\widehat{\Gamma}_\psi - \Gamma_0\|_1 \\
&\geq -\frac{1}{\sqrt{NT}}\|\widehat{\Gamma}_\psi - \Gamma_0\|_1 \frac{1}{\sqrt{NT}}\|\mathrm{mat}(\mathbf{M}_x e)\|_\infty + \frac{\psi}{\sqrt{NT}}\|\widehat{\Gamma}_\psi - \Gamma_0\|_1 \\
&= \left(\psi - \frac{\|\mathrm{mat}(\mathbf{M}_x e)\|_\infty}{\sqrt{NT}}\right)\frac{\|\widehat{\Gamma}_\psi - \Gamma_0\|_1}{\sqrt{NT}}.
\end{aligned}$$

The required result follows since $\psi - \|\mathrm{mat}(\mathbf{M}_x e)\|_\infty > 0$. □

B.5 Sufficient Conditions for Restricted Strong Convexity

In this section we discuss Assumption 1 in more detail. Define the distance $\mathcal{H}(A, \mathbb{C})$ between a matrix $A \in \mathbb{R}^{N\times T}$ and the cone $\mathbb{C}$ by

$$\mathcal{H}(A, \mathbb{C}) := \left[\min_{B\in\mathbb{C}} \mathrm{Tr}(A - B)'(A - B)\right]^{1/2}.$$

The following lemma provides an alternative formulation for our restricted strong convexity assumption.

Lemma S.7. *Let there exists a positive constant $\mu > 0$ such that for any $\alpha \in \mathbb{R}^K$ with $\alpha'\left(\frac{x'x}{NT}\right)\alpha = 1$, the regressors $X_1, ..., X_K$ satisfy*

$$\mathcal{H}\left(\alpha \cdot \frac{X}{\sqrt{NT}}, \mathbb{C}\right)^2 \geq \mu > 0, \quad \textit{wpa1.}$$

Then Assumption 1 holds.

Proof of Lemma S.7. Recall the definition $x = [x_1, ..., x_K], (NT \times K)$, where $x_k = \mathrm{vec}(X_k)$. Firstly, if $\theta = 0$, then the required result holds for any constant $\mu > 0$. Secondly, if $\theta' x = 0$, then the required result holds for $\mu = 1$ because $(\theta'\theta - \theta' x(x'x)^{-1}x'\theta) = \theta'\theta$. Thus, in the following we only need to consider the case $\theta \neq 0$ and $\theta' x \neq 0$. Also let $x \neq 0$.

Define $\tilde{x}_\theta = \frac{\mathbf{P}_x\theta}{\|\mathbf{P}_x\theta\|}$, and $\widetilde{X}_\theta := \text{mat}(\tilde{x}_\theta)$. Then, for any $\Theta \in \mathbb{C}$ and $\Theta \neq 0$, we have

$$
\begin{aligned}
&\frac{1}{2NT}\left(\theta'\theta - \theta' x (x'x)^{-1} x'\theta\right) \\
&= \frac{1}{2NT}\left(\theta'\theta - \theta'\widetilde{x}_\theta\widetilde{x}_\theta'\theta\right) \quad \text{(by the definition of } \tilde{x}_\theta) \\
&= \frac{1}{2NT}\|\Theta\|_2^2\left(1 - \frac{\theta'\widetilde{x}_\theta\widetilde{x}_\theta'\theta}{\theta'\theta}\right) \quad \text{(since } \theta \neq 0) \\
&= \frac{1}{2NT}\|\Theta\|_2^2\left(1 - \widetilde{x}_\theta'\frac{\theta\theta'}{\theta'\theta}\widetilde{x}_\theta\right) = \frac{1}{2NT}\|\Theta\|_2^2\left(\widetilde{x}_\theta'\widetilde{x}_\theta - \widetilde{x}_\theta'\frac{\theta\theta'}{\theta'\theta}\widetilde{x}_\theta\right) \\
&= \frac{1}{2NT}\|\Theta\|_2^2\left(\|\widetilde{x}_\theta - \mathbf{P}_\theta\widetilde{x}_\theta\|_2^2\right) \\
&\geq \frac{1}{2NT}\|\Theta\|_2^2\left(\min_{A\in\mathbb{C}}\|\widetilde{x}_\theta - \text{vec}(A)\|^2\right) \\
&= \frac{1}{2NT}\|\Theta\|_2^2\left(\mathcal{H}(\widetilde{X}_\theta, \mathbb{C})^2\right), \qquad \text{(S.8)}
\end{aligned}
$$

where the inequality holds because $\text{mat}(\mathbf{P}_\theta\widetilde{x}_\theta) \in \mathbb{C}$ since $\Theta \in \mathbb{C}$ and C is a cone. Notice that

$$\tilde{x}_\theta = \frac{\mathbf{P}_x\theta}{\|\mathbf{P}_x\theta\|_2} = \frac{x}{\sqrt{NT}}\alpha_*,$$

where $\alpha_* = \frac{\left(\frac{x'x}{NT}\right)^{-1}\frac{x'}{\sqrt{NT}}\theta}{\left(\theta'\frac{x}{\sqrt{NT}}\left(\frac{x'x}{NT}\right)^{-1}\frac{x'}{\sqrt{NT}}\theta\right)^{1/2}}$ and $\alpha_*'\left(\frac{x'x}{NT}\right)\alpha_* = 1$. This implies

$$\widetilde{X}_\theta = \alpha_* \cdot \frac{X}{\sqrt{NT}}$$

with $\alpha_*'\alpha_* = 1$. Therefore, we have

$$(S.8) \geq \frac{1}{2NT}\|\Theta\|_2^2\left(\min_{\alpha'\left(\frac{x'x}{NT}\right)\alpha=1}\mathcal{H}\left(\alpha\cdot\frac{X}{\sqrt{NT}}, \mathbb{C}\right)^2\right).$$

Then, the required result of the lemma follows by the assumptions in the lemma. □

Lemma S.8. *Consider $K = 1$. Let $s_1 \geq s_2 \geq s_3 \geq \ldots \geq s_{\min(N,T)} \geq 0$ be the singular values of the $N \times T$ matrix $\mathbf{M}_{\lambda_0}X_1\mathbf{M}_{f_0}$. Assume that there exists a sequence $q_{NT} \geq 2$ such that*

(i) $\frac{1}{\sqrt{NT}}\|X_1\|_2 = O_P(1)$.

(ii) $\frac{1}{NT}\sum_{r=q_{NT}}^{\min(N,T)} s_r^2 \geq c > 0$ *wpa1.*

(iii) $\frac{1}{\sqrt{NT}}\sum_{r=1}^{q_{NT}-2}(s_r - s_{q_{NT}}) \to_P \infty$.

Then Assumption 1 is satisfied with $\mu = c$.

This lemma could be generalized to $K > 1$. We would then need to impose the conditions for X_1 in the lemma for all linear combination $\alpha \cdot X$, in an appropriate uniform sense over all α with $\|\alpha\| = 1$.

Proof of Lemma S.8. For given $N \times T$ matrix X, and $N \times R_0$ matrix λ_0, and $T \times R_0$ matrix f_0, we want to find a lower bound on

$$\begin{aligned}\nu_{NT} &:= NT \ \mathcal{H}\left(\frac{X_1}{\sqrt{NT}}, \mathbb{C}\right)^2 = NT \min_{\Theta \in \mathbb{C}} \left\| X_1/\sqrt{NT} - \Theta \right\|_2^2 \\ &= \min_{\Theta \in \mathbb{R}^{N \times T}} \|X_1 - \Theta\|_2^2 \qquad \text{s.t.} \qquad \|\mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \leq 3 \|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 .\end{aligned}$$

By definition, we have

$$\|X_1 - \Theta\|_2^2 = \|\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0} - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2^2 + \|(X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}) - (\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0})\|_2^2 .$$

Also, $\text{rank}(\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}) \leq 2R_0$ (e.g., see Lemma 3.4 of Recht, Fazel, and Parrilo (2010)), and therefore $\|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \leq \sqrt{2R_0} \|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2$. Using this we find

$$\begin{aligned}\nu_{NT} \geq \min_{\Theta \in \mathbb{R}^{N \times T}} &\left\{ \|\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0} - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2^2 + \|(X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}) - (\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0})\|_2^2 \right\} \\ &\text{s.t.} \ \|\mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \leq 3 \sqrt{2R_0} \ \|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2 .\end{aligned}$$

Here, we have weakened the constraint (allowing more values for Θ), and the minimizing value therefore weakly decreases. It is easy to see that for $\omega \geq 0$ we have

$$\begin{aligned}\left(\|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_2 - \omega \right)^2 = \min_{\Theta \in \mathbb{R}^{N \times T}} &\|(X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}) - (\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0})\|_2^2 \\ &\text{s.t.} \qquad \|\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2 = \omega,\end{aligned}$$

because the optimal $\Theta - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}$ here equals $X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}$ rescaled by a non-negative number. We therefore have

$$\begin{aligned}\nu_{NT} \geq \min_{\omega \geq 0} \min_{\Theta \in \mathbb{R}^{N \times T}} &\left(\|\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0} - \mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_2 \right)^2 + \left(\|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_2 - \omega \right)^2 \\ &\text{s.t.} \ \|\mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}\|_1 \leq 3 \sqrt{2R_0}\, \omega.\end{aligned}$$

Let

$$\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0} = \sum_{r=1}^{\min(N,T)-R_0} s_r \, v_r w_r',$$

be the singular value decomposition of $\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}$ with singular values $s_r \geq 0$ and normalized singular vectors $v_r \in \mathbb{R}^N$ and $w_r \in \mathbb{R}^T$. The optimal $\mathbf{M}_{\lambda_0} \Theta \mathbf{M}_{f_0}$ in the last optimization problem has the form

$$\sum_{r=1}^{\min(N,T)-R_0} \max(0, s_r - \xi)\, v_r w_r',$$

for some $\xi \geq 0$ (see Lemma S.1). Here, $\xi = 0$ occurs if the constraint is not binding, that is, if $\|\mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_1 \leq 3\sqrt{2R_0}\,\omega$. We therefore have

$$\begin{aligned}\nu_{NT} \geq \min_{\omega \geq 0, \xi \geq 0} & \sum_{r=1}^{\min(N,T)-R_0} (s_r - \max(0, s_r - \xi))^2 + \left(\|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_2 - \omega\right)^2 \\ \text{s.t.} & \sum_{r=1}^{\min(N,T)-R_0} \max(0, s_r - \xi) \leq 3\sqrt{2R_0}\,\omega.\end{aligned}$$

Here, the optimal ω equals $\max\left\{\|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0})\|_2\,,\ \frac{1}{3\sqrt{2R_0}} \sum_{r=1}^{\min(N,T)-R_0} \max(0, s_r - \xi)\right\}$, and we thus have

$$\begin{aligned}\nu_{NT} \geq \min_{\xi \geq 0} & \sum_{r=1}^{\min(N,T)-R_0} \Bigg[\min(s_r^2, \xi^2) \\ & + \left(\max\left\{ 0, \frac{1}{3\sqrt{2R_0}} \left(\sum_{r=1}^{\min(N,T)-R_0} \max(0, s_r - \xi) \right) - \|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_2 \right\} \right)^2 \Bigg].\end{aligned}$$

Let $\infty = s_0 > s_1 \geq \ldots \geq s_{\min(N,T)-R_0} \geq s_{\min(N,T)-R_0+1} = 0$. For any $\xi \geq 0$ there exists q be

such that $\xi \in [s_{q+1}, s_q]$. We can therefore write

$$\begin{aligned}
\nu_{NT} \geq & \min_{q \in \{0,1,2,\ldots,\min(N,T)-R_0\}} \min_{\xi \in [s_{q+1}, s_q]} \left[q\,\xi^2 + \sum_{r=q+1}^{\min(N,T)-R_0} s_r^2 \right. \\
& \left. + \left(\max\left\{ 0, \frac{1}{3\sqrt{2R_0}} \left(\sum_{r=1}^{q} (s_r - \xi) \right) \mathbb{1}\{q \geq 1\} - \|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0})\|_2 \right\} \right)^2 \right] \\
\geq & \min_{q \in \{0,1,2,\ldots,\min(N,T)-R_0\}} \left[\left(\min_{\xi \in [s_{q+1}, s_q]} q\,\xi^2 \right) + \sum_{r=q+1}^{\min(N,T)-R_0} s_r^2 \right. \\
& \left. + \left(\max\left\{ 0, \frac{1}{3\sqrt{2R_0}} \left(\min_{\xi \in [s_{q+1}, s_q]} \sum_{r=1}^{q} (s_r - \xi) \right) \mathbb{1}\{q \geq 1\} - \|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0})\|_2 \right\} \right)^2 \right] \\
= & \min_{q \in \{0,1,2,\ldots,\min(N,T)-R_0\}} \left[q\,s_{q+1}^2 + \sum_{r=q+1}^{\min(N,T)-R_0} s_r^2 \right. \\
& \left. + \left(\max\left\{ 0, \frac{1}{3\sqrt{2R_0}} \left(\sum_{r=1}^{q-1} (s_r - s_q) \right) \mathbb{1}\{q \geq 2\} - \|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0})\|_2 \right\} \right)^2 \right].
\end{aligned}$$

Shifting $q \mapsto q - 1$ we can rewrite this as

$$\frac{\nu_{NT}}{NT} \geq \min_{q \in \{1,2,\ldots,\min(N,T)-R_0\}} \left(a(q) + \left[\max\{0, b(q)\} \right]^2 \right),$$

where

$$\begin{aligned}
a(q) &= \frac{1}{NT} \left[(q-1)\, s_q^2 + \sum_{r=q}^{\min(N,T)} s_r^2 \right], \\
b(q) &= \frac{1}{\sqrt{NT}} \left[\frac{1}{3\sqrt{2R_0}} \left(\sum_{r=1}^{q-2} (s_r - s_q) \right) \mathbb{1}\{q \geq 3\} - \|X_1 - \mathbf{M}_{\lambda_0} X_1 \mathbf{M}_{f_0}\|_2 \right].
\end{aligned}$$

Notice that $a(q)$ is nonnegative and weakly decreasing and $b(q)$ is weakly increasing. Then,

for any integer valued sequence q_{NT} between 1 and $\min(N,T) - R_0$ such that $b(q_{NT}) > 0$,

$$\begin{aligned}
&\min_{q\in\{1,2,...,\min(N,T)-R_0\}} \left(a(q) + [\max\{0, b(q)\}]^2 \right) \\
&= \min\left\{ \min_{q\in\{1,2,...,q_{NT}\}} \left(a(q) + [\max\{0, b(q)\}]^2 \right), \min_{q\in\{q_{NT}+1,...,\min(N,T)-R_0\}} \left(a(q) + [\max\{0, b(q)\}]^2 \right) \right\} \\
&\geq \min\left\{ \min_{q\in\{1,2,...,q_{NT}\}} a(q), \min_{q\in\{q_{NT}+1,...,\min(N,T)-R_0\}} [\max\{0, b(q)\}]^2 \right\} \\
&\geq \min\left\{ a(q_{NT}), b(q_{NT}+1)^2 \right\}.
\end{aligned}$$

The assumptions of the lemma thus guarantee that $\nu_{NT}/(NT) \geq c$. The definition of ν_{NT} together with Lemma S.7 thus guarantees that Assumption 1 is satisfied with $\mu = c$. □

Remarks

(a) When X is a"high-rank" regressor and s_q's are of an order $O_P(\sqrt{\max(N,T)})$, we can choose, for example, $q_{NT} = \lfloor \min(N,T)/2 \rfloor$, for N, T converging to infinity at the same rate, where $\lfloor a \rfloor$ is the integer part of a. Then, it is easy to verify those sufficient condition (i), (ii) and (iii) for e.g. $X_{it} \sim i.i.d.\mathcal{N}(0, \sigma^2)$ from well-known random matrix theory results. More generally, we can explicitly verify (i), (ii) and (iii) if X has an approximate factor structure

$$X = \lambda_x f_x' + E_x,$$

where $\lambda_x f_x'$ is an arbitrary low-rank factor structure, and $E_x \sim i.i.d.\mathcal{N}(0, \sigma^2)$.

(b) For a low-rank regressor with $\text{rank}(X) = 1$, we have singular values $s_1 = \|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_2$ and $s_r = 0$ for all $r \geq 2$. In that case we find that $a(1) = \frac{1}{NT} s_1^2$ and $a(q) = 0$ for $q > 1$, and we have $b(1) = b(2) = 0$ and $b(q) = b(3) = \frac{1}{\sqrt{NT}} \left[\frac{1}{3\sqrt{2R_0}} s_1 - \|X - \mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_2 \right]$ for all $q \geq 3$. Also, $a(1) \geq b(2)$. Therefore

$$\min_{q\in\{1,2,...,\min(N,T)\}} \left[a(q) + (\max\{0, b(q)\})^2 \right] = \min\left\{ a(1), (\max\{0, b(3)\})^2 \right\}$$

Thus, the assumptions of Lemma S.8 are satisfied if wpa1 we have

$$\frac{1}{\sqrt{NT}} \left[\|\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_2 - 3\sqrt{2R_0} \|X - \mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0}\|_2 \right] \geq c_1 > 0$$

for some constant c_1. This last condition simply demands that the part of X that cannot be explained by λ_0 and f_0 needs to be sufficiently larger than the part of X that can be explained by either λ_0 or f_0. This is a sufficient condition for Assumption 1.

An analysis that is specialized towards low-rank regressors will likely give a weaker condition for Assumption 1 in this case.

B.6 Proofs for Section 3.2.2

Proof of Theorem 3 . Remember the following singular value decompositions: $\Gamma_0 = USV'$, $\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0} = \mathbf{M}_U E \mathbf{M}_V = U_E S_E V_E'$, and $\mathbf{M}_{\lambda_0} X \mathbf{M}_{f_0} = \mathbf{M}_U X \mathbf{M}_V = U_x S_x V_x'$. The proof consists of two steps. In the first step, we show that the local minimizer that minimizes the objective function $Q_*(\beta)$ in a convex neighborhood of β_0 defined by

$$\mathbf{B} := \left\{ \beta : \frac{c_x \, c_{\text{low}}}{c_{\text{up}}} \, |\Delta\beta| \leq 1 \right\}$$

is $\sqrt{T}$- consistent. In the second step, we show that the local minimizer is the global minimizer, for which we use convexity of the objective function $Q_*(\beta)$.

Step 1. By definition of the nuclear norm, we have

$$Q_*(\beta) = \|\Gamma_0 + E - \Delta\beta \cdot X\|_1 = \sup_{\{A \,:\, \|A\|_\infty \leq 1\}} \text{Tr}\left[(\Gamma_0 + E - \Delta\beta \cdot X)' \, A \right].$$

To obtain a lower bound on $Q_*(\beta)$ we choose the following matrix A in the above minimization,

$$A_\beta = UV' + \sqrt{1 - a_\beta^2} \, U_E V_E' - a_\beta \, (\text{sgn}\, \Delta\beta) \, \mathbf{M}_{U_E} U_x V_x',$$

where $\mathbf{M}_{U_E} = I_N - U_E U_E'$ and $a_\beta \in [0, 1]$ is given by

$$a_\beta = \frac{c_x \, c_{\text{low}}}{c_{\text{up}}} \, |\Delta\beta|.$$

We have $\|A_\beta\|_\infty \leq 1$, because

$$\begin{aligned} \|A_\beta\|_\infty^2 &= \max\left\{ \|UV'\|_\infty^2 \,, \left\| \sqrt{1 - a_\beta^2} \, U_E V_E' - a_\beta \, (\text{sgn}\, \Delta\beta) \, \mathbf{M}_{U_E} U_x V_x' \right\|_\infty^2 \right\} \\ &\leq \max\left\{ \|UV'\|_\infty^2 \,, (1 - a_\beta^2) \, \|U_E V_E'\|_\infty^2 + a_\beta^2 \, \|\mathbf{M}_{U_E} U_x V_x'\|_\infty^2 \right\} \\ &= 1. \end{aligned}$$

Here, for the first line, we used that UV' is orthogonal to $\sqrt{1 - a_\beta^2} \, U_E V_E' - a_\beta \, (\text{sgn}\, \Delta\beta) \, \mathbf{M}_{U_E} U_x V_x'$ in both matrix dimensions (that is, $U'U_E, U'U_k, V'V_E, V'V_x = 0$) and applied Lemma S.6(iv). For the second line, we used that the columns of $U_E V_E'$ are orthogonal to the columns of

$\mathbf{M}_{U_E} U_x V_x'$ since $U_E' \mathbf{M}_{U_E} = 0$, and applied Lemma S.6(v). In the final line we used that $\|UV'\|_\infty = \|U_E V_E'\|_\infty = 1$ and that $\|\mathbf{M}_{U_E} U_x V_x'\|_\infty \leq 1$.

With this choice of $A = A_\beta$ we obtain the following lower bound for the objective function; for all $\beta \in \mathbf{B}$,

$$\begin{aligned}
Q_*(\beta) &\geq \operatorname{Tr}\left[(\Gamma_0 + E - \Delta\beta \cdot X)' A_\beta\right] \\
&= \|\Gamma_0\|_1 + \operatorname{Tr}\left(E'UV'\right) + \operatorname{Tr}\left[(-\Delta\beta \cdot X)' UV'\right] \\
&\quad + \sqrt{1 - a_\beta^2}\, \|\mathbf{M}_U E \mathbf{M}_V\|_1 + \sqrt{1 - a_\beta^2}\, \operatorname{Tr}\left[(-\Delta\beta \cdot X)' U_E V_E'\right] \\
&\quad + a_\beta\, |\Delta\beta|\, \operatorname{Tr}\left[X' \mathbf{M}_{U_E} U_x V_x'\right],
\end{aligned}$$

where we used the following:

$$\begin{aligned}
\operatorname{Tr}\left(\Gamma_0' UV'\right) &= \operatorname{Tr}\left(VSU'UV'\right) = \operatorname{Tr}(S) = \|\Gamma_0\|_1, \\
\operatorname{Tr}\left(E' U_E V_E'\right) &= \operatorname{Tr}\left((E - \mathbf{M}_U E \mathbf{M}_V + \mathbf{M}_U E \mathbf{M}_V)' U_E V_E'\right) \\
&= \operatorname{Tr}((\mathbf{M}_U E \mathbf{M}_V)' U_E V_E')) = \operatorname{Tr}(S_E) = \|\mathbf{M}_U E \mathbf{M}_V\|_1, \\
\operatorname{Tr}\left(\Gamma_0' U_E V_E'\right) &= \operatorname{Tr}\left(VSU' U_E V_E'\right) = 0, \\
\operatorname{Tr}\left[\Gamma_0' \mathbf{M}_{U_E} U_x V_x'\right] &= \operatorname{Tr}\left[VSU \mathbf{M}_{U_E} U_x V_x'\right] = 0, \\
\operatorname{Tr}\left[E' \mathbf{M}_{U_E} U_x V_x'\right] &= \operatorname{Tr}\left[\mathbf{M}_V E' \mathbf{M}_U \mathbf{M}_{U_E} U_x V_x'\right] + \operatorname{Tr}\left[(E' - \mathbf{M}_V E' \mathbf{M}_U) \mathbf{M}_{U_E} U_x V_x'\right] = 0.
\end{aligned}$$

We furthermore have $Q(\beta_0) = \|\Gamma_0 + E\|_1$. Thus, applying the assumptions of the theorem and also using $\sqrt{1 - a_\beta^2} \geq 1 - \frac{1}{2} a_\beta^2 - \frac{1}{2} a_\beta^4$, we obtain for $\beta \in \mathbf{B}$,

$$\begin{aligned}
Q_*(\beta) - Q_*(\beta_0) &\geq \operatorname{Tr}\left[(\Gamma_0 + E - \Delta\beta \cdot X)' A_\beta\right] - \|\Gamma_0 + E\|_1 \\
&\geq a_\beta\, |\Delta\beta|\, \operatorname{Tr}\left[X' \mathbf{M}_{U_E} U_x V_x'\right] \\
&\quad - \frac{1}{2} a_\beta^2\, \|\mathbf{M}_U E \mathbf{M}_V\|_1 - (\|\Gamma_0 + E\|_1 - \|\Gamma_0\|_1 - \|\mathbf{M}_U E \mathbf{M}_V\|_1) \\
&\quad + \operatorname{Tr}\left(E'UV'\right) - \frac{1}{2} a_\beta^4 \|\mathbf{M}_U E \mathbf{M}_V\|_1 \\
&\quad + \sqrt{1 - a_\beta^2}\, \operatorname{Tr}\left[(-\Delta\beta \cdot X)' U_E V_E'\right] + \operatorname{Tr}\left[(-\Delta\beta \cdot X)' UV'\right] \\
&=: B_1 - B_2 - B_3 + B_4 - B_5 + B_6. \qquad \text{(S.9)}
\end{aligned}$$

Here we bound B_1 from below by

$$\begin{aligned}
B_1 &= a_\beta \, |\Delta\beta| \, \mathrm{Tr}\left(X' \mathbf{M}_{U_E} U_x V_x'\right) \\
&= a_\beta \, |\Delta\beta| \left[\mathrm{Tr}\left(\mathbf{M}_V X' \mathbf{M}_U \mathbf{M}_{U_E} U_x V_x'\right) - \mathrm{Tr}\left((X' - \mathbf{M}_V X' \mathbf{M}_U) \mathbf{M}_{U_E} U_x V_x'\right)\right] \\
&= a_\beta \, |\Delta\beta| \, \mathrm{Tr}\left(V_x S_x U_x' \mathbf{M}_{U_E} U_x V_x'\right) \\
&= a_\beta \, |\Delta\beta| \; |\Delta\beta| \left[\mathrm{Tr}\left(S_x\right) - \mathrm{Tr}\left(U_E' U_x S_x U_x' U_E\right)\right] \\
&\geq a_\beta \, c_x \; |\Delta\beta| \; \mathrm{Tr}\left(S_x\right) = a_\beta \, c_x \; |\Delta\beta| \; \|\mathbf{M}_U X \mathbf{M}_V\|_1 \\
&\geq a_\beta \, c_x \, c_{\text{low}} \, T \sqrt{N} \; |\Delta\beta| \, .
\end{aligned}$$

Here the first inequality holds by assumption (vi), and the second inequality holds by assumption (v).

We bound B_2 from above by

$$\begin{aligned}
B_2 &= \frac{1}{2} a_\beta^2 \, \|\mathbf{M}_U E \mathbf{M}_V\|_1 \\
&\leq \frac{1}{2} a_\beta^2 \; (\|E\|_1 + \|\mathbf{P}_U E\|_1 + \|E \mathbf{P}_V\|_1 + \|\mathbf{P}_U E \mathbf{P}_V\|_1) \\
&\leq \frac{1}{2} a_\beta^2 \; (\|E\|_1 + 3R_0 \|E\|_\infty) \\
&\leq \frac{1}{2} a_\beta^2 \, T \sqrt{N} \left(\frac{c_{\text{up}}}{2} + \frac{1}{T} O_P(1)\right) \quad \text{wpa1} \\
&\leq \frac{1}{2} a_\beta^2 \, T \sqrt{N} \, c_{\text{up}} \quad \text{wpa1},
\end{aligned}$$

where the first inequality holds by the triangle inequality, the second inequality holds by Lemma S.6(i) and the third and the fourth inequalities follow by assumption (i) and (ii).

We bound term B_3 from above by

$$\begin{aligned}
B_3 &= \|\Gamma_0 + E\|_1 - \|\Gamma_0\|_1 - \|\mathbf{M}_U E \mathbf{M}_V\|_1 \\
&\leq \|E - \mathbf{M}_U E \mathbf{M}_V\|_1 = \|\mathbf{P}_U E + E \mathbf{P}_V - \mathbf{P}_U E \mathbf{P}_V\|_1 \\
&\leq \|\mathbf{P}_U E\|_1 + \|E \mathbf{P}_V\|_1 + \|\mathbf{P}_U E \mathbf{P}_V\|_1 \\
&\leq 3R_0 \|E\|_\infty \\
&\leq O_P(\sqrt{N})
\end{aligned}$$

where the second inequality holds by the triangle inequality and the third inequality holds by Lemma S.6(i).

For B_4, by Hölder's inequality we have

$$B_4 = \|\mathrm{Tr}\,(E'UV')\,\| \leq \|E\|_\infty \|UV'\|_1 = O_P(\sqrt{N}).$$

For B_5, denoting $O_{P+}(\cdot)$ as a stochastically strictly positive and bounded term and using similar arguments for the bound of term B_2, we obtain

$$B_5 = \frac{1}{2} a_\beta^4 \|\mathbf{M}_U E \mathbf{M}_V\|_1 = O_{P+}(1) a_\beta^4 \, T\sqrt{N} = O_{P+}(1)(\Delta\beta)^4 \, T\sqrt{N}.$$

For B_6, we have

$$B_6 = \sqrt{1-a_\beta^2}\,\mathrm{Tr}\left[(-\Delta\beta \cdot X)' U_E V_E'\right] + \mathrm{Tr}\left[(-\Delta\beta \cdot X)' UV'\right] = O_P\left(\sqrt{NT}|\Delta\beta|\right),$$

where the last equality holds since $\mathrm{Tr}(X_k U_E V_E') = O_P(\sqrt{NT})$ by assumption (vi), and $\mathrm{Tr}(X_k UV') \leq \|X\|_\infty \|UV'\|_1 = O_P(\sqrt{NT})$ under assumption (iii).

Notice that our choice for a_β above is such that $a_\beta \, c_x \, c_{\text{low}}|\Delta\beta| - \frac{1}{2} a_\beta^2 \, c_{\text{up}}$ is maximized, which guarantees that $B_1 - B_2$ is positive, namely

$$\frac{B_1 - B_2}{T\sqrt{N}} \geq \frac{c_x^2 c_{\text{low}}^2}{2\, c_{\text{up}}} |\Delta\beta|^2.$$

Combining the above, for any $\beta \in \mathbf{B}$, we have

$$\frac{1}{T\sqrt{N}} \left\{Q_*(\beta) - Q_*(\beta_0)\right\} \geq \frac{c_x^2 c_{\text{low}}^2}{2\, c_{\text{up}}} |\Delta\beta|^2 + O_P\left(\frac{1}{\sqrt{T}}|\Delta\beta|\right) + O_P(T^{-1}) + O_{P+}(1)|\Delta\beta|^4,$$

which holds uniformly over $\beta \in \mathbf{B}$ (i.e. none of the constants hidden in the $O_P(.)$ notation depends on β).

Let

$$\widetilde{\beta}_* := \underset{\beta \in \mathbf{B}}{\mathrm{argmin}}\, Q_*(\beta)$$

be the local minimizer in a convex neighborhood $\mathbf{B}$ of β_0. Notice that since $\beta_0 \in \mathbf{B}$, $Q_*(\widehat{\beta}_*) \leq Q_*(\beta_0)$ by definition. Therefore, we have

$$\begin{aligned} 0 &\geq \frac{1}{T\sqrt{N}} (Q_*(\widetilde{\beta}_*) - Q_*(\beta_0)) \\ &\geq \frac{c_x^2 c_{\text{low}}^2}{2\, c_{\text{up}}} |\widetilde{\beta}_* - \beta_0|^2 + O_P\left(\frac{1}{\sqrt{T}}|\widetilde{\beta}_* - \beta_0|\right) + O_P\left(\frac{1}{T}\right) + O_{P+}\left(|\widetilde{\beta}_* - \beta_0|^4\right). \end{aligned}$$

This implies

$$O_{P+}\left(\frac{1}{T}\right) \geq \left(\frac{c_x^2 c_{\text{low}}^2}{2\, c_{\text{up}}} + O_{P+}(1)|\widetilde{\beta}_* - \beta_0|^2\right)|\widetilde{\beta}_* - \beta_0|^2 + O_P\left(\frac{1}{\sqrt{T}}\right)|\widetilde{\beta}_* - \beta_0|$$
$$\geq \frac{c_x^2 c_{\text{low}}^2}{2\, c_{\text{up}}}|\widetilde{\beta}_* - \beta_0|^2 + O_P\left(\frac{1}{\sqrt{T}}\right)|\widetilde{\beta}_* - \beta_0|.$$

From this we deduce

$$|\widetilde{\beta}_* - \beta_0| = O_P\left(\frac{1}{\sqrt{T}}\right). \tag{S.10}$$

Step 2. Let $\bar{\beta} \in \partial\mathbf{B}$, that is, $\alpha_{\bar{\beta}} = 1$. Write $\Delta\bar{\beta} := \bar{\beta} - \beta_0$. From (S.9) with $a_{\bar{\beta}} = 1$, we can bound $Q_*(\bar{\beta}) - Q_*(\beta_0)$ from below by

$$\begin{aligned}
&\frac{1}{T\sqrt{N}}(Q_*(\bar{\beta}) - Q_*(\beta_0)) \\
&\geq c_x\, c_{\text{low}}\, |\Delta\bar{\beta}| - \frac{1}{2}\, c_{\text{up}} + O_P\left(\frac{1}{\sqrt{T}}|\Delta\bar{\beta}|\right) + O_P\left(\frac{1}{T}\right) + O_{P+}(1)|\Delta\bar{\beta}|^4 \\
&= \frac{1}{2}\, c_{\text{up}} + O_P\left(\frac{1}{T}\right) + O_P\left(\frac{1}{\sqrt{T}}\right)\frac{c_{\text{up}}}{c_x\, c_{\text{low}}} + O_{P+}(1)\left(\frac{c_{\text{up}}}{c_x\, c_{\text{low}}}\right)^4 \\
&> 0 \quad \text{wpa1},
\end{aligned}$$

where the equality holds since $|\Delta\bar{\beta}| = \frac{c_{\text{up}}}{c_x\, c_{\text{low}}}$.

Since $Q_*(\beta)$ is convex and has unique minimum, the local minimum at $\widetilde{\beta}_*$ is also the global minimum asymptotically. Therefore, asymptotically

$$\widetilde{\beta}_* = \widehat{\beta}_* \quad \text{wpa1}.$$

Combining this with the $\sqrt{T}-$ consistency result of the local minimizer in (S.10) gives the statement of the theorem. □

B.6.1 Extension of Theorem 3

Theorem 3 is the special case of one regressors ($K = 1$). We can extend this to a more general case with K regressors. The proof of the following general theorem is similar to that of Theorem 3, and we skip it.

Theorem S.1 (Generalization of Theorem 3 to multiple regressors). *Let there exist symmetric idempotent $T \times T$ matrices $Q_k = Q_{k,NT}$ such that $Q_k V = 0$, for all $k \in \{1, \ldots, K\}$, and $Q_k Q_\ell = 0$, for all $k, \ell \in \{1, \ldots, K\}$. Suppose that $N > T$. As $N, T \to \infty$, we assume the following conditions hold.*

(i) $\|E\|_\infty = O_P(\sqrt{N})$.

(ii) *There exists a finite positive constant* c_{up} *such that* $\frac{1}{T\sqrt{N}}\|E\|_1 \leq \frac{1}{2}c_{\text{up}}$, *wpa1.*

(iii) $\|X_k\|_\infty = O_P(\sqrt{NT})$, *for* $k \in \{1, \ldots, K\}$.

(vi) *Let* $U_E S_E V_E'$ *be the singular value decomposition of* $\mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0}$. *We assume* $\text{Tr}\left(X_k' U_E V_E'\right) = O_P(\sqrt{NT})$ *for all* $k \in \{1, \ldots, K\}$.

(v) *We assume that there exists a constant* $c_{\text{low}} > 0$ *such that wpa1*

$$T^{-1}N^{-1/2}\|\mathbf{M}_U X_k \mathbf{M}_V Q_k\|_1 \geq c_{\text{low}},$$

for all $k \in \{1, \ldots, K\}$.

(vi) *For* $k = 1, \ldots, K$ *let* $U_k S_k V_k' = \mathbf{M}_U X_k \mathbf{M}_V Q_k (= \mathbf{M}_U X_k Q_k)$ *be the singular value decomposition of the matrix* $\mathbf{M}_U X_k \mathbf{M}_V Q_k$. *We assume that there exists* $c_x \in (0,1)$ *such that wpa1* $\|U_k' U_E\|_\infty^2 \leq (1 - c_x)$ *for all* $k = 1, \ldots, K$.

We then have $\sqrt{T}\left(\widehat{\beta}_* - \beta_0\right) = O_P(1)$.

Remark For $t \in \{1, 2, \ldots, T\}$, let $\mathfrak{e}_t$ be the t'th unit vector of dimension T. For $k \in \{1, \ldots, K\}$, let $A_k = (\mathfrak{e}_{\lfloor (k-1)T/K \rfloor + 1}, \mathfrak{e}_{\lfloor (k-1)T/K \rfloor + 2}, \ldots \mathfrak{e}_{\lfloor kT/K \rfloor})$ be a $T \times \lfloor T/K \rfloor$ matrix, and let $\mathbf{P}_{A_k}$ be the projector onto the column space of A_k. Also define $f_{0,k} = \mathbf{P}_{A_k} f_0$ and $B_k = \mathbf{M}_{f_{0,k}} A_k$. Then, for $K > 1$ one possible choice for Q_k in assumption (vi) of Theorem S.1 is given by

$$Q_k = \mathbf{P}_{B_k} = \mathbf{M}_{f_{0,k}} \mathbf{P}_{A_k}.$$

The discussion of assumption (vi) of Theorem S.1 is then analogous to the $K = 1$ case, except that for the k'th regressor only the time periods $\lfloor (k-1)T/K \rfloor + 1$ to $\lfloor kT/K \rfloor$ are used in the assumption, that is, we need enough variation in the k'th regressor within those time periods. Other choices of Q_k are also conceivable.

B.7 Proofs for Section 4

For $\beta \in \mathbb{R}^K$ we define

$$\left\{\widehat{\lambda}(\beta), \widehat{f}(\beta)\right\} := \underset{\lambda \in \mathbb{R}^{N \times R_0}, f \in \mathbb{R}^{T \times R_0}}{\text{argmin}} \|Y - \beta \cdot X\|_2^2,$$

and the corresponding projection matrices

$$\mathbf{M}_{\widehat{\lambda}}(\beta) := \mathbb{I}_N - \widehat{\lambda}(\beta)\left(\widehat{\lambda}(\beta)'\widehat{\lambda}(\beta)\right)^{-1}\widehat{\lambda}(\beta)', \qquad \mathbf{M}_{\widehat{f}}(\beta) := \mathbb{I}_T - \widehat{f}(\beta)\left(\widehat{f}(\beta)'\widehat{f}(\beta)\right)^{-1}\widehat{f}(\beta)'.$$

Lemma S.9. *Under the assumptions (i) and (ii) of Theorem 4 we have*

$$\mathbf{M}_{\widehat{\lambda}}(\beta) = \mathbf{M}_{\lambda_0} + \mathbf{M}^{(1)}_{\widehat{\lambda},E} + \mathbf{M}^{(2)}_{\widehat{\lambda},E} - \sum_{k=1}^{K} (\beta_k - \beta_{0,k})\, \mathbf{M}^{(1)}_{\widehat{\lambda},k} + \mathbf{M}^{(\mathrm{rem})}_{\widehat{\lambda},E} + \mathbf{M}^{(\mathrm{rem})}_{\widehat{\lambda}}(\beta)\ ,$$
$$\mathbf{M}_{\widehat{f}}(\beta) = \mathbf{M}_{f_0} + \mathbf{M}^{(1)}_{\widehat{f},E} + \mathbf{M}^{(2)}_{\widehat{f},E} - \sum_{k=1}^{K} (\beta_k - \beta_{0,k})\, \mathbf{M}^{(1)}_{\widehat{f},k} + \mathbf{M}^{(\mathrm{rem})}_{\widehat{f},E} + \mathbf{M}^{(\mathrm{rem})}_{\widehat{f}}(\beta)\ ,$$

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

$$\sup_{\beta\in\mathcal{B}(\beta_0,r_{NT})} \frac{\left\|\mathbf{M}^{(\mathrm{rem})}_{\widehat{\lambda}}(\beta)\right\|_\infty}{\|\beta-\beta_0\|^2 + (NT)^{-1/2}\|E\|_\infty\|\beta-\beta_0\|} = O_P(1)\,, \quad \sup_{\beta\in\mathcal{B}(\beta_0,r_{NT})} \frac{\left\|\mathbf{M}^{(\mathrm{rem})}_{\widehat{\lambda},E}\right\|_\infty}{(NT)^{-3/2}\|E\|_\infty^3} = O_P(1)\,,$$
$$\sup_{\beta\in\mathcal{B}(\beta_0,r_{NT})} \frac{\left\|\mathbf{M}^{(\mathrm{rem})}_{\widehat{f}}(\beta)\right\|_\infty}{\|\beta-\beta_0\|^2 + (NT)^{-1/2}\|E\|_\infty\|\beta-\beta_0\|} = O_P(1)\,, \quad \sup_{\beta\in\mathcal{B}(\beta_0,r_{NT})} \frac{\left\|\mathbf{M}^{(\mathrm{rem})}_{\widehat{f},E}\right\|_\infty}{(NT)^{-3/2}\|E\|_\infty^3} = O_P(1)\,,$$

and the expansion coefficients are given by

$$\begin{aligned}
\mathbf{M}^{(1)}_{\widehat{\lambda},E} &= -\,\mathbf{M}_{\lambda_0}\, E\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\lambda_0' \;-\; \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, E'\, \mathbf{M}_{\lambda_0}\ ,\\
\mathbf{M}^{(1)}_{\widehat{\lambda},k} &= -\,\mathbf{M}_{\lambda_0}\, X_k\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\lambda_0' \;-\; \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, X_k'\, \mathbf{M}_{\lambda_0}\ ,\\
\mathbf{M}^{(2)}_{\widehat{\lambda},E} &= \mathbf{M}_{\lambda_0}\, E\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\lambda_0'\, E\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\lambda_0'\\
&\quad + \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, E'\, \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, E'\, \mathbf{M}_{\lambda_0}\\
&\quad - \mathbf{M}_{\lambda_0}\, E\, \mathbf{M}_{f_0}\, E'\, \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\, \lambda_0'\\
&\quad - \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\, \lambda_0'\, E\, \mathbf{M}_{f_0}\, E'\, \mathbf{M}_{\lambda_0}\\
&\quad - \mathbf{M}_{\lambda_0}\, E\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, E'\, \mathbf{M}_{\lambda_0}\\
&\quad + \lambda_0\, (\lambda_0'\lambda_0)^{-1}\, (f_0'f_0)^{-1}\, f_0'\, E'\, \mathbf{M}_{\lambda_0}\, E\, f_0\, (f_0'f_0)^{-1}\, (\lambda_0'\lambda_0)^{-1}\lambda_0'\,,
\end{aligned}$$

analogously

$$
\begin{aligned}
\mathbf{M}^{(1)}_{\widehat{f},E} &= -\mathbf{M}_{f_0} E' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' - f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' E \mathbf{M}_{f_0} , \\
\mathbf{M}^{(1)}_{\widehat{f},k} &= -\mathbf{M}_{f_0} X_k' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' - f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' X_k \mathbf{M}_{f_0} , \\
\mathbf{M}^{(2)}_{\widehat{f},E} &= \mathbf{M}_{f_0} E' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' E' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' \\
&\quad + f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' E f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' E \mathbf{M}_{f_0} \\
&\quad - \mathbf{M}_{f_0} E' \mathbf{M}_{\lambda_0} E f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' \\
&\quad - f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' E' \mathbf{M}_{\lambda_0} E \mathbf{M}_{f_0} \\
&\quad - \mathbf{M}_{f_0} E' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' E \mathbf{M}_{f_0} \\
&\quad + f_0 (f_0'f_0)^{-1} (\lambda_0'\lambda_0)^{-1} \lambda_0' E \mathbf{M}_{f_0} E' \lambda_0 (\lambda_0'\lambda_0)^{-1} (f_0'f_0)^{-1} f_0' .
\end{aligned}
$$

Proof. This lemma is a restatement of Theorem S.9.1 in the supplementary appendix of Moon and Weidner (2017), and the proof is given there. However, in the presentation here we split the remainder terms of the expansions into two components, e.g. $\mathbf{M}^{(\text{rem})}_{\widehat{\lambda},E} + \mathbf{M}^{(\text{rem})}_{\widehat{\lambda}}(\beta)$, where $\mathbf{M}^{(\text{rem})}_{\widehat{\lambda},E}$ summarizes all higher order expansion terms depending on E only, and $\mathbf{M}^{(\text{rem})}_{\widehat{\lambda}}(\beta)$ summarizes all higher order terms also involving $\beta - \beta_0$. The reason for this change in presentation is that we will consider differences of the form $\mathbf{M}_{\widehat{\lambda}}(\beta_1) - \mathbf{M}_{\widehat{\lambda}}(\beta_2)$ below, and the remainder terms $\mathbf{M}^{(\text{rem})}_{\widehat{\lambda},E}$ cancel in those differences. □

Proof of Theorem 4. # The first statement of the theorem is an almost immediate consequence of Theorem 4.1 in Moon and Weidner (2017). That theorem shows that, under the assumptions we impose here, we have the following approximate quadratic expansion of the profile LS objective function,

$$
L_{R_0}(\beta) = L_{R_0}(\beta_0) - \frac{1}{\sqrt{NT}} (\beta - \beta_0)' C_{NT} + \frac{1}{2} (\beta - \beta_0)' W_{NT} (\beta - \beta_0) + \frac{1}{NT} R_{NT}(\beta) ,
$$

where the remainder $R_{NT}(\beta)$ is such that for any sequence $r_{NT} \to 0$ we have

$$
\sup_{\beta \in \mathcal{B}(\beta_0, r_{NT})} \frac{|R_{NT}(\beta)|}{\left(1 + \sqrt{NT}\, \|\beta - \beta_0\|\right)^2} = o_p(1) ,
$$

and $W_{NT} = \frac{1}{NT} x' (\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}) x$, and $C_{NT} = C^{(1)}_{NT} + C^{(2)}_{NT}$, with $C^{(1)}_{NT} = \frac{1}{NT} x' (\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}) x$,

and the K-vector $C_{NT}^{(2)}$ has entries, $k = 1, \ldots, K$,

$$\begin{aligned} C_{NT,k}^{(2)} = & -\frac{1}{\sqrt{NT}} \Big[\operatorname{Tr}\left(E M_{f_0} E' M_{\lambda_0} X_k f_0 \left(f_0' f_0\right)^{-1} \left(\lambda_0' \lambda_0\right)^{-1} \lambda_0' \right) \\ & + \operatorname{Tr}\left(E' M_{\lambda_0} E M_{f_0} X_k' \lambda_0 \left(\lambda_0' \lambda_0\right)^{-1} \left(f_0' f_0\right)^{-1} f_0' \right) \\ & + \operatorname{Tr}\left(E' M_{\lambda_0} X_k M_{f_0} E' \lambda_0 \left(\lambda_0' \lambda_0\right)^{-1} \left(f_0' f_0\right)^{-1} f_0' \right) \Big] . \end{aligned}$$

We have assumed that $\operatorname{plim}_{N,T\to\infty} W_{NT} > 0$ and $C_{NT}^{(1)} = O_P(1)$, and using our assumptions (i) and (ii) we also find that

$$\left| C_{NT,k}^{(2)} \right| \leq \frac{3\, R_0}{\sqrt{NT}} \|E\|_\infty^2 \|X_k\|_\infty \|\lambda_0\|_\infty \|f_0\|_\infty \left\| \left(\lambda_0' \lambda_0\right)^{-1} \right\|_\infty \left\| \left(f_0' f_0\right)^{-1} \right\|_\infty = O_P(1),$$

and therefore $C_{NT} = 0$. From this approximate quadratic expansion we conclude that $L_{R_0}(\beta)$ has indeed at least one local minimizer within $\mathcal{B}(\beta_0, r_{NT})$, and that any such local minimizer within $\mathcal{B}(\beta_0, r_{NT})$ satisfied

$$\sqrt{NT} \left(\widehat{\beta}_{\mathrm{LS},R_0}^{\mathrm{local}} - \beta_0 \right) = W_{NT}^{-1} C_{NT} = O_P(1).$$

Next, we want to show the second statement of the theorem. Let $\widehat{\lambda} := \widehat{\lambda}\left(\widehat{\beta}_{\mathrm{LS},R_0}^{\mathrm{local}}\right)$ and $\widehat{f} := \widehat{f}\left(\widehat{\beta}_{\mathrm{LS},R_0}^{\mathrm{local}}\right)$. By definition we have $\widehat{\lambda}^{(s+1)} = \widehat{\lambda}\left(\widehat{\beta}^{(s)}\right)$ and $\widehat{f}^{(s+1)} = \widehat{f}\left(\widehat{\beta}^{(s)}\right)$, and

$$\begin{aligned} \left(x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) x \right) \widehat{\beta}^{(s+1)} &= x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} \right) y, \\ \left(x' \left(\mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) x \right) \widehat{\beta}_{\mathrm{LS},R_0}^{\mathrm{local}} &= x' \left(\mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) y. \end{aligned}$$

By taking the difference of those last equations we obtain

$$\begin{aligned} \left(x' \left(\mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) x \right) \left(\widehat{\beta}^{(s+1)} - \widehat{\beta}_{\mathrm{LS},R_0}^{\mathrm{local}} \right) &= x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) \left(y - x\, \widehat{\beta}^{(s+1)} \right) \\ &= x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) \left[e - x \left(\widehat{\beta}^{(s+1)} - \beta_0 \right) + (f_0 \otimes \lambda_0) \mathrm{vec}(\mathbf{I}_R) \right], \end{aligned}$$

where in the last step we plugged in the model for y. Applying Lemma S.9, the result from the first part of the theorem, and our assumptions we find that

$$\frac{1}{NT} x' \left(\mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}} \right) x = \frac{1}{NT} x' \left(\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0} \right) x + o_P(1),$$

and since the probability limit of $\frac{1}{NT} x' (\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}) x$ is assumed to be invertible we obtain

$$\widehat{\beta}^{(s+1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0} = \left[\frac{1}{NT} x' (\mathbf{M}_{f_0} \otimes \mathbf{M}_{\lambda_0}) x\right]^{-1} \frac{1}{NT} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}}\right) \times \left[e - x\left(\widehat{\beta}^{(s+1)} - \beta_0\right) + (f_0 \otimes \lambda_0)\text{vec}(\mathbf{I}_R)\right][1 + o_P(1)].$$

Again applying Lemma S.9 and our assumptions one can show that

$$\left\|\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}}\right\|_\infty = O_P\left(\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|\right),$$

and therefore

$$\begin{aligned}\frac{1}{NT} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}}\right) e &= O_P\left(\frac{\|E\|_\infty \max_k \|X_k\|_\infty}{NT}\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|\right) \\ &= O_P\left(\frac{\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|}{\sqrt{\min(N,T)}}\right),\end{aligned}$$

and

$$\begin{aligned}\frac{1}{NT} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}}\right) x \left(\widehat{\beta}^{(s+1)} - \beta_0\right) &= O_P\left(\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| \left\|\widehat{\beta}^{(s+1)} - \beta_0\right\|\right) \\ = O_P\left(\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| \left\|\widehat{\beta}^{(s+1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| + \frac{\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|}{\sqrt{NT}}\right),\end{aligned}$$

where in the last step we used that part of the theorem implies that $\widehat{\beta}^{(s+1)} - \beta_0 = \widehat{\beta}^{(s+1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0} + O_P(1/\sqrt{NT})$. Finally, using one more time Lemma S.9 and our assumptions we can also show that

$$\begin{aligned}&\frac{1}{NT} x' \left(\mathbf{M}_{\widehat{f}^{(s+1)}} \otimes \mathbf{M}_{\widehat{\lambda}^{(s+1)}} - \mathbf{M}_{\widehat{f}} \otimes \mathbf{M}_{\widehat{\lambda}}\right) (f_0 \otimes \lambda_0)\text{vec}(\mathbf{I}_R) \\ &\qquad = O_P\left(\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|^2 + \frac{\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\|}{\sqrt{\min(N,T)}}\right).\end{aligned}$$

Combining the above gives

$$\begin{aligned}&\widehat{\beta}^{(s+1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0} \\ &= O_P\left\{\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| \left[\left\|\widehat{\beta}^{(s+1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| + \left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| + \frac{1}{\sqrt{\min(N,T)}}\right]\right\}[1 + o_P(1)].\end{aligned}$$

Starting from the assumptions $\|\widehat{\beta}^{(0)} - \beta_0\| = O_P(c_{NT})$, for $c_{NT} \to 0$, we thus conclude that

$$\left\|\widehat{\beta}^{(1)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| = O_P\left\{c_{NT}\left(c_{NT} + \frac{1}{\sqrt{\min(N,T)}}\right)\right\},$$

and then also

$$\left\|\widehat{\beta}^{(2)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| = O_P\left\{c_{NT}\left(c_{NT} + \frac{1}{\sqrt{\min(N,T)}}\right)^2\right\},$$

and by induction over s we conclude in this way that

$$\left\|\widehat{\beta}^{(s)} - \widehat{\beta}^{\text{local}}_{\text{LS},R_0}\right\| = O_P\left\{c_{NT}\left(c_{NT} + \frac{1}{\sqrt{\min(N,T)}}\right)^s\right\}.$$

□

B.8 Proof of Section 6

Proof of Theorem 5. Like in the previous section, let $Q_\psi(\beta) := \min_\Gamma Q_\psi(\beta, \Gamma)$. Let $\mathbf{B}_\psi(M) := \{\beta \in \mathbf{B} : \|\beta - \beta_0\| = M\psi^{1/2}\}$ be the restricted parameter set consisting of β's whose distance to β_0 is less than or equal to $M\psi^{1/2}$. In the special case where β is a scalar (i.e., $K = 1$) which is assumed in the theorem, $\mathbf{B}_\psi(M)$ is a finite discrete set consisting of two points,

$$\mathbf{B}_\psi(M) = \{\beta_0 - M\psi^{1/2},\, \beta_0 + M\psi^{1/2}\} \tag{S.11}$$

Since $Q_\psi(\beta)$ is convex, if we show that there exists a finite constant M such that

$$\min_{\beta \in \mathbf{B}_\psi(M)} Q_\psi(\beta) - Q_\psi(\beta_0, \Gamma_0) > 0 \quad \text{wp1}, \tag{S.12}$$

then we can deduce

$$\|\widehat{\beta}_\psi - \beta_0\| \leq M\psi^{1/2} \quad \text{wp1},$$

which is required for the theorem.

For (S.12), we find a function $Q^*_\psi(\beta, \Gamma)$ such that $Q_\psi(\beta, \Gamma) \geq Q^*_\psi(\beta, \Gamma)$ for all β, Γ. With $Q^*_\psi(\beta) := \min_\Gamma Q^*_\psi(\beta, \Gamma)$, we show that there exists a finite constant M such that

$$\min_{\beta \in \mathbf{B}_\psi(M)} Q^*_\psi(\beta) - Q^*_\psi(\beta_0, \Gamma_0) > 0 \quad \text{wp1}. \tag{S.13}$$

A lower bound objective function, $Q^*_\psi(\beta,\Gamma)$: For every pair i,t we define the function $m^*_{it}:\mathbb{R}\to\mathbb{R}$ as the function that satisfies

$$m^*_{it}(z^0_{it}) = m_{it}(z^0_{it}), \qquad \partial_z m^*_{it}(z^0_{it}) = \partial_z m_{it}(z^0_{it}), \qquad \forall z\in\mathbb{R}:\ \frac{\partial_{z^2} m^*_{it}(z)}{b} = \frac{\partial_{z^2} m_{it}(z)}{\partial_{z^2}\overline{m}_{it}(z)}.$$

Here, the last condition on the second derivative should be interpreted in terms of "generalized functions" in cases where $m_{it}(z)$ is not twice differentiable. For example, in the quantile regression example we have $m_{it}(z)=\rho_\tau(Y_{it}-z)$, and therefore $\partial_{z^2} m_{it}(z)=\delta(Y_{it}-z)$, where $\delta(.)$ denotes the Dirac delta function. In general, solving for $m^*_{it}(z)$ we find that

$$m^*_{it}(z) := m_{it}(z^0_{it}) + \left(z-z^0_{it}\right)\,\partial_z m_{it}(z^0_{it}) + b\int_{z^0_{it}}^{z}\int_{z^0_{it}}^{\zeta}\frac{\partial_{z^2} m_{it}(\xi)}{\partial_{z^2}\overline{m}_{it}(\xi)}d\xi d\zeta,$$

where for $z<z^0_{it}$ the integral should be interpreted as $\int_{z^0_{it}}^{z} q(\zeta)d\zeta = -\int_{z}^{z^0_{it}} q(\zeta)d\zeta$, and analogously for the integral over ξ.[23] Let $\overline{m}^*_{it}(z)=\mathbb{E}(m^*_{it}(z)|X)$. Our definition of $m^*_{it}(z)$ together with $\mathbb{E}[\partial_z m_{it}(z^0_{it})|X]=0$ imply that

$$\overline{m}^*_{it}(z) = \overline{m}_{it}(z^0_{it}) + \frac{b}{2}\left(z-z^0_{it}\right)^2,$$

that is, $\overline{m}^*_{it}(z)$ is a quadratic function with second derivative equal to b. Our assumption $\partial_{z^2}\overline{m}_{it}(z)\geq b$ for all $z\in\mathcal{Z}$ (Assumption 2(iii)) together with convexity of $m_{it}(z)$ (Assumption 2(ii)) imply furthermore that $0\leq\partial_{z^2} m^*_{it}(z)\leq\partial_{z^2} m_{it}(z)$. Therefore, $m^*_{it}(z)$ is a convex function and satisfies

$$m_{it}(z)-m_{it}(z^0_{it}) \geq m^*_{it}(z)-m^*_{it}(z^0_{it}), \tag{S.15}$$

because $m^*_{it}(z^0_{it})=m_{it}(z^0_{it})$ and the convex function $m_{it}(z)$ has a steeper curvature than the convex function $m^*_{it}(z)$ everywhere.

Next, we define

$$Q^*_\psi(\beta,\Gamma) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X'_{it}\,\beta+\Gamma_{it}\right) + \frac{\psi}{\sqrt{NT}}\underbrace{\max_{\left\{A\in\mathbb{R}^{N\times T}\,\middle|\,\|A\|_\infty\leq 1\right\}}\mathrm{Tr}(\Gamma' A)}_{=\|\Gamma\|_1}, \tag{S.16}$$

[23] In the quantile regression case we have

$$\int_{z^0_{it}}^{\zeta}\frac{\partial_{z^2} m_{it}(\xi)}{\partial_{z^2}\overline{m}_{it}(\xi)}d\xi = \frac{\mathbb{1}(z^0_{it}<Y_{it}<\zeta)-\mathbb{1}(\zeta<Y_{it}<z^0_{it})}{\partial_{z^2}\overline{m}_{it}(Y_{it})}. \tag{S.14}$$

and $Q^*_\psi(\beta) := \min_{\Gamma\in\mathbb{R}^{N\times T}} Q^*_\psi(\beta,\Gamma)$. From (S.15) we obtain that

$$Q_\psi(\beta,\Gamma) - Q_\psi(\beta_0,\Gamma_0) \geq Q^*_\psi(\beta,\Gamma) - Q^*_\psi(\beta_0,\Gamma_0).$$

Additional definitions: We already defined the expected objective function $\overline{m}_{it}(z)$ in the main text. We now also define the deviation from the expectation $\widetilde{m}_{it}(z) := m_{it}(z) - \overline{m}_{it}(z)$. We drop the argument z whenever those function and their derivatives are evaluated at the true values z^0_{it}, for example, $m_{it} = m_{it}(z^0_{it})$, $\partial_z \widetilde{m}_{it} = \partial_z \widetilde{m}_{it}(z^0_{it})$, $\partial_{z^2}\overline{m}_{it} = \partial_{z^2}\overline{m}_{it}(z^0_{it})$. We use the same notation for $m^*_{it}(z)$, for example, $\partial_{z^2}\overline{m}^*_{it} = \partial_{z^2}\overline{m}^*_{it}(z^0_{it})$. In addition, we define the $N\times T$ matrix $\Gamma^* := \Gamma_0 + X^{(1)}\cdot\beta_0$, and we let $z_{it}(\beta) := X^{(2)\prime}_{it}\beta + \Gamma^*_{it}$.

Deriving a lower bound on $Q^*_\psi(\beta)$ within the shriking neighborhood: Our goal here is to find a lower bound on $Q^*_\psi(\beta)$ that is valid within the shrinking neighborhood of $\beta_0, \mathbf{B}_\psi(M)$. To obtain such a lower bound we choose the matrix A in equation (S.16) to be the $N\times T$ matrix $A(\beta)$ with elements

$$\begin{aligned}
A_{it}(\beta) &:= -\frac{1}{\sqrt{NT}\psi}\,\partial_z m^*_{it}\left(z_{it}(\beta)\right) \\
&= -\frac{1}{\sqrt{NT}\psi}\left[\partial_z \widetilde{m}^*_{it}\left(z_{it}(\beta)\right) + \partial_z \overline{m}^*_{it}\left(z_{it}(\beta)\right)\right], \\
&= -\frac{1}{\sqrt{NT}\psi}\left[\partial_z \widetilde{m}^*_{it}\left(z_{it}(\beta)\right) + b\, X^{(2)\prime}_{it}(\beta-\beta_0)\right],
\end{aligned}$$

where in the final step we used that $\partial_z \overline{m}^*_{it}(z) = \partial_z \overline{m}^*_{it} + b(z - z^0_{it})$, and $\partial_z \overline{m}^*_{it} = 0$, and $z_{it}(\beta) - z^0_{it} = X^{(2)\prime}_{it}(\beta-\beta_0)$. For the mean zero $N\times T$ matrix $\partial_z \widetilde{m}^*\left(z(\beta)\right) := \left[\partial_z \widetilde{m}^*_{it}\left(z_{it}(\beta)\right)\right]$ we have

$$\begin{aligned}
\sup_{\beta\in\mathbf{B}_\psi(M)} \left\|\partial_z \widetilde{m}^*\left(z(\beta)\right)\right\|_\infty &\leq \left\|\partial_z \widetilde{m}^*\left(z(\beta_0 - M\psi^{1/2})\right)\right\|_\infty + \left\|\partial_z \widetilde{m}^*\left(z(\beta_0 + M\psi^{1/2})\right)\right\|_\infty \\
&= O_P\left(\sqrt{\max(N,T)}\right).
\end{aligned}$$

We thus find that

$$\sup_{\beta\in\mathbf{B}_\psi(M)} \|A(\beta)\|_\infty \leq o_P(1) + \sup_{\beta\in\mathbf{B}_\psi(M)} o_P\left(\frac{\|\beta-\beta_0\|}{\psi^{1/2}}\right) \leq o_P(M).$$

A sufficient condition for $\|A(\beta)\|_\infty \leq 1$ wp1 uniformly in $\beta\in\mathbf{B}_\psi(M)$ is therefore satisfied. From now on, we use $\leq_{\text{u.p.}}$ to denote that the inequality holds wp1 uniformly in $\beta\in\mathbf{B}_\psi(M)$.

Under that condition we thus have

$$
\begin{aligned}
Q^*_\psi(\beta) &\geq \min_{\Gamma\in\mathbb{R}^{N\times T}} \left\{ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X'_{it}\,\beta+\Gamma_{it}\right) + \frac{\psi}{\sqrt{NT}}\mathrm{Tr}[\Gamma' A(\beta)] \right\} \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) + \frac{\psi}{\sqrt{NT}}\mathrm{Tr}[\left(\Gamma_0 - X^{(1)}\cdot(\beta-\beta_0)\right)' A(\beta)] \\
&\geq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) - \frac{\psi}{\sqrt{NT}}\left\|\Gamma_0 - X^{(1)}\cdot(\beta-\beta_0)\right\|_1 \|A(\beta)\|_\infty \\
&\geq_{\text{u.p.}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) - \psi\frac{\|\Gamma_0\|_1}{\sqrt{NT}} - \psi\frac{\left\|X^{(1)}\cdot(\beta-\beta_0)\right\|_1}{\sqrt{NT}} \qquad \text{(S.17)}
\end{aligned}
$$

where the second line (the equality part) holds because we used that our choice of $A(\beta)$ implies that the FOC for the minimization over Γ are satisfied for $\Gamma = \Gamma^* - X^{(1)}\cdot\beta = \Gamma_0 - X^{(1)}\cdot(\beta-\beta_0)$. The third line holds by the Holder inequality $|\,\mathrm{Tr}\,A'B| \leq \|A\|_\infty \|B\|_1$, and the last line holds by the triangle inequality and $\|A(\beta)\|_1 \leq_{\text{u.p.}} 1$.

Next, by expanding $X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}$ around $z^0_{it} = X^{(2)\prime}_{it}\,\beta_0+\Gamma^*_{it}$ and by definition of $\overline{m}^*_{it}$, we obtain

$$
\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{m}^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{m}^*_{it}\left(X^{(2)\prime}_{it}\,\beta+\Gamma^*_{it}\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{m}^*_{it} + b(\beta-\beta_0)'W(\beta-\beta_0) + O_P(1/\sqrt{NT}), \qquad \text{(S.18)}
\end{aligned}
$$

where the $O_P(1/\sqrt{NT})$ holds uniformly over β in $\mathbf{B}_\psi(M)$.

Consistency of $\widehat{\beta}_\psi$:

Using the low bounds of (S.17) and (S.18), and the definition of $Q^*_\psi(\beta_0, \Gamma_0)$, we have

$$
\begin{aligned}
&\min_{\beta \in \mathbf{B}_\psi(M)} Q^*_\psi(\beta) - Q^*_\psi(\beta_0, \Gamma_0) \\
&\geq \min_{\beta \in \mathbf{B}_\psi(M)} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(m^*_{it} \left(X^{(2)\prime}_{it} \beta + \Gamma^*_{it} \right) - m^*_{it} \left(X^{(2)\prime}_{it} \beta_0 + \Gamma^*_{it} \right) \right) \right] \\
&\qquad - \psi \frac{2 \left\| \Gamma_0 \right\|_1}{\sqrt{NT}} - \psi \max_{\beta \in \mathbf{B}_\psi(M)} \frac{\left\| X^{(1)} \cdot (\beta - \beta_0) \right\|_1}{\sqrt{NT}} \\
&= \min_{\beta \in \mathbf{B}_\psi(M)} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(m^*_{it} \left(X^{(2)\prime}_{it} \beta + \Gamma^*_{it} \right) - \overline{m}^*_{it} \right) \right] - O_+(1)\psi - o_{P+}(1) M\psi \\
&\geq b \min_{\beta \in \mathbf{B}_\psi(M)} (\beta - \beta_0)' W (\beta - \beta_0) - \frac{O_{P+}(1)}{\sqrt{NT}} - O_+(1)\psi - o_{P+}(1) M\psi \\
&\geq b \lambda_{min}(W) M^2 \psi - \frac{O_{P+}(1)}{\sqrt{NT}} - O_+(1)\psi - o_{P+}(1) M\psi \\
&\geq M\psi \left(b \lambda_{min}(W) M - \frac{O_{P+}(1)}{M\psi\sqrt{NT}} - \frac{O_+(1)}{M} - o_{P+}(1) \right).
\end{aligned}
$$

Since $\lambda_{min}(W) \to_p \lambda_{min}(W_\infty) > 0$ and $\psi\sqrt{NT} \to 0$, we can choose a large constant M such that

$$
b \lambda_{min}(W) M - \frac{O_{P+}(1)}{M\psi\sqrt{NT}} - \frac{O_+(1)}{M} - o_{P+}(1) > 0 \quad \text{wp1.}
$$

Then, we have the required result for the theorem. □

To establish the consistency result in the theorem in a more general case where $K > 1$, the proof requires some additional technical restrictions. The first technical requirement is the uniform bound, $\sup_{\beta \in \mathbf{B}_\psi(M)} \left\| A(\beta) \right\|_\infty$. For this, we may use a recent random matrix theory result in Moon (2019) which requires further regularity conditions such as the tail condition of the distribution of $A_{it}(\beta)$ and a restriction of the entrophy of the parameter set $\mathbf{B}_\psi(M)$. Secondly, we need additional technical restrictions for a uniform stochastic bound of $\sup_{\beta \in \mathbf{B}_\psi(M)} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{m}^*_{it} \left(X^{(2)\prime}_{it} \beta + \Gamma^*_{it} \right) = O_P(1/\sqrt{NT})$.