WORKING PAPERS

N° TSE –1008

April 2019

"Square-root nuclear norm penalized estimator for panel data models with approximately low-rank unobserved Heterogeneity"

Jad Beyhum and Eric Gautier

Toulouse School of Economics

SQUARE-ROOT NUCLEAR NORM PENALIZED ESTIMATOR FOR PANEL DATA MODELS WITH APPROXIMATELY LOW-RANK UNOBSERVED HETEROGENEITY

JAD BEYHUM AND ERIC GAUTIER

Abstract. This paper considers a nuclear norm penalized estimator for panel data models with interactive effects. The low-rank interactive effects can be an approximate model and the rank of the best approximation unknown and grow with sample size. The estimator is solution of a well-structured convex optimization problem and can be solved in polynomial-time. We derive rates of convergence, study the low-rank properties of the estimator, estimation of the rank and of annihilator matrices when the number of time periods grows with the sample size. Two-stage estimators can be asymptotically normal. None of the procedures require knowledge of the variance of the errors.

1. Introduction

Panel data allow to estimate models with flexible unobserved heterogeneity using the fact that each individual is observed repeatedly. The high-dimensional statistics literature enables estimation in the presence of a high-dimensional parameter, provided that it has a low-dimensional structure. This paper studies a model that borrows from the two aforementioned strands of literature. We consider a linear panel data model with interactive effects of the form: for $i = 1, ..., N$ and $t = 1, ..., T$,

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{kit} + \lambda_i^\top f_t + \Gamma_{it}^d + E_{it}, \quad \mathbb{E}[E_{it}] = 0, \tag{1}$$

where Y_{it} is the outcome, X_{kit} is the k^{th} regressor, $\beta \in \mathbb{R}^K$ is a vector of parameters, λ_i and f_t are vectors in $\mathbb{R}^r$ of factor loadings and factors, Γ_{it}^d is a remainder which can account for many weak factors, and E_{it} is an error. Only β is considered nonrandom. Precise assumptions on the joint distribution of the vector of right-hand side variables is given later. Importantly, only the regressors and outcomes are available to the researcher. The regressors correspond to observed heterogeneity and the remaining right-hand side elements are called unobserved heterogeneity. The interactive effects or factor structure generalizes the usual individual plus time effects in where $\lambda_i^\top f_t = c_i + d_t$. It allows for example for group time effects of the form d_{gt} for individuals in group g. One can think that $\lambda_i^\top f_t + \Gamma_{it}^d + E_{it}$ accounts for the contribution

The authors acknowledge financial support from the grant ERC POEMH 337665.
Keywords: panel data, interactive fixed effects, factor models, flexible unobserved heterogeneity, nuclear norm penalization, unknown variance.
MSC 2010 subject classification: 62J99, 62H12, 62H25.

of regressors which are not available to the researcher but have an effect on the outcome if we believe these have an approximate factor structure plus remainder plus error term. In such a case, the error E_{it} is a composite error which accounts for a linear combination of those coming from the approximate factor structure of the missing regressors and the usual error from the model which includes both observed and unobserved regressors. When the regressors and $\lambda_i^\top f_t + \Gamma_{it}^d$ are correlated, the least-squares estimator is inconsistent. This is a situation where we say that the regressors are endogenous or that there is an omitted variable bias. The specification that we analyze is very flexible to model unobserved heterogeneity and can be used in the context of many applications (see, *e.g.*, [13] in the context of public policy evaluation). It is also a challenging one which has mainly been analyzed when the number of factors r is fixed, especially when r is known, and $\Gamma_{it}^d = 0$. In matrix form, (1) becomes

$$Y = \sum_{k=1}^{K} \beta_k X_k + \Gamma^l + \Gamma^d + E, \tag{2}$$

where $Y, X_1, ..., X_K, \Gamma^l, \Gamma^d$ and E are random $N \times T$ matrices. Γ^l is such that $\Gamma_{it}^l = \lambda_i^\top f_t$ and $\text{rank}\left(\Gamma^l\right) = r$ and Γ^d has small nuclear norm. The nuclear norm is the ℓ_1-norm of the vector of singular values. We denote by $\Gamma = \Gamma^l + \Gamma^d$. In this paper, β is most of the time the parameter of interest and Γ^l a nuisance. Many variations on model (1) have been considered and we name only a few. In [9, 24] the regressors have a factor structure and β can vary across individuals. In [14, 18] the number of regressors grows with the sample size. [9, 20] allow for lags of the outcome in (1). [3] proposes a least-squares estimator for the model which equation is (1) when $\Gamma = \Gamma^l$ and r is fixed and known. The least squares criterion involves the product of λ_i and f_t or a rank restriction and is not convex. It is shown to be $\sqrt{NT}$-consistent and asymptotically normal when, among other things, the factors are strong. [19] shows that using the same estimator with an upper bound on the true number of factors leads to the same asymptotic properties.

The tools in this paper are related to those used in matrix completion. There, the problem consists in estimating the unobserved entries of a low-rank matrix from an observed subset of its entries, sometimes with additive noise (see, *e.g.*, [7, 8, 15, 16, 17, 26, 27, 28]). The usual ℓ_0 and ℓ_1-norms are replaced by the rank and nuclear norm, soft and hard thresholding are carried on the singular values. These methods have recently been used in econometrics (see in particular [2, 4, 10]). The problem in this paper differs in that we observe all the entries of the matrices Y and $X_1, \ldots, X_K$ but none of $\Gamma + E$ and both Γ and E are random.

The iterative procedure in [3] could yield a local minimum while the theoretical properties are for the global minimum. In contrast, the estimators in [21] and in this paper involve convex programs for which converge to a global minimum is achieved in polynomial time. The additional novelties of this paper are as follows. This paper considers a square-root nuclear norm penalized estimator (see [5] for the Lasso), where the sum of squared residuals is replaced by its square-root. It can be viewed as the estimator in [21] using a data-driven penalty level so it is directly implementable by the researcher and does not require an additional diverging multiplicative factor which can result in over-penalization and is useful in finite samples. We provide a straightforward iterative algorithm to compute the estimator. Our results do not rely on conditioning on realizations of Γ and we state the conditions on the joint distribution of Γ and the regressors. Moreover, we allow the interactive effect to be an approximate model and hence many non-strong-factors (see [25]) via the additional term Γ^d. The rank of Γ^l is treated as random and can grow with the sample size and be unknown. We obtain low-rank oracle type inequalities for various loss functions and results on the rank of our estimator of Γ, introduce a thresholded estimator which can be used to estimate the rank of Γ^l as well as projectors on the vector spaces spanned by the factors and factor loadings which we analyze theoretically. We also obtain rates of convergence for the estimation of β. These results do not rely on a strong-factor assumption which amounts to assuming that the ratio of any singular value of Γ^l and $\sqrt{NT}$ has a deterministic limit as N goes to infinity and T increases with N. Finally, we propose a two-stage estimator and show its asymptotic normality. Based on our result on the estimation of the rank of Γ^l by the procedures of this paper, we can proceed as analyzed in [21] and use the estimator in [3] as a second stage.

2. Preliminaries

$\mathbb{N}$ denotes the positive integers, $\mathbb{N}_0$ denotes $\mathbb{N}\cup\{0\}$. For $a\in\mathbb{R}$, we set $a_+=\max(a,0)$ and, for $a\neq 0$, $a/0=\infty$. $\{\mu_N\}$ denotes a numerical sequence of generic term μ_N. $\mathcal{M}_{NT}$ is the set of matrices with real coefficients of size $N\times T$. The transpose of a matrix $A\in\mathcal{M}_{NT}$ is written $A^\top$, its trace is $\mathrm{tr}(A)$, and its rank is $\mathrm{rank}(A)$. For $A\in\mathcal{M}_{NT}$, $\mathrm{vec}(A)$ is the operator that vectorizes the columns of A and, for a vector $v\in\mathbb{R}^{NT}$, $\mathrm{mat}(v)$ is the unique matrix in $\mathcal{M}_{NT}$ such that $v=\mathrm{vec}\left(\mathrm{mat}(v)\right)$. When matrices are defined involving capital letters, their vectorization is denoted using lowercase letters. The k^{th} singular value of $A\in\mathcal{M}_{NT}$ (arranged in decreasing order and repeated according to multiplicty) is $\sigma_k(A)$ and $\mathrm{rank}(A)$ is its rank. $A=\sum_{k=1}^{\mathrm{rank}(A)}\sigma_k(A)u_k(A)v_k(A)^\top$ is a singular value decomposition of A, where $\{u_k(A)\}_{k=1}^{\mathrm{rank}(A)}$ is a family of orthonormal vectors of $\mathbb{R}^N$ and $\{v_k(A)\}_{k=1}^{\mathrm{rank}(A)}$ is a family of orthonormal vectors of $\mathbb{R}^T$. The scalar product in $\mathcal{M}_{NT}$ is $\langle A,B\rangle=\mathrm{tr}\left(A^\top B\right)$. The ℓ_2-norm (or Frobenius norm) is $|A|_2^2=\langle A,A\rangle=\sum_{k=1}^{\mathrm{rank}(A)}\sigma_k(A)^2$, the nuclear norm is $|A|_*=\sum_{k=1}^{\mathrm{rank}(A)}\sigma_k(A)$, and the operator norm is $|A|_{\mathrm{op}}=\sigma_1(A)$. $P_{u(A)}$ and $P_{v(A)}$ are the orthogonal projectors onto $\mathrm{span}\{u_1(A),\dots,u_{\mathrm{rank}(A)}(A)\}$ and $\mathrm{span}\{v_1(A),\dots,v_{\mathrm{rank}(A)}(A)\}$ and $M_{u(A)}$ and $M_{v(A)}$ onto the orthogonal complements. For $\Delta\in\mathcal{M}_{NT}$, $\mathcal{P}_A$ is defined as $\mathcal{P}_A(\Delta)=\Delta-M_{u(A)}\Delta M_{v(A)}$ and $\mathcal{P}_A^\perp$ as $\mathcal{P}_A^\perp(\Delta)=M_{u(A)}\Delta M_{v(A)}$. We rely, for $A\in\mathcal{M}_{NT}$ and $c>0$, on the cone $C_{A,c}=\left\{\Delta\in\mathcal{M}_{NT}:\ \left|\mathcal{P}_A^\perp(\Delta)\right|_*\le c\left|\mathcal{P}_A(\Delta)\right|_*\right\}$.

We denote by P_X (resp. M_X) the orthogonal projector on the vector space spanned by $\{X_k\}_{k=1}^K$ (resp. on its orthogonal) and $X=(x_1,\dots,x_K)$. We consider an asymptotic where N goes to infinity and T is a function of N that goes to infinity when N goes to infinity. The probabilistic framework consists of a sequence of data generating processes (henceforth DGPs) that depend on N. We write that an event occurs w.p.a. 1 ("with probaility approaching 1") when its probability converges to 1 as N goes to infinity. All limits are when N goes to infinity. We denote convergence in probability and in distribution by respectively $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{d}$. We allow the researcher to apply annihilator matrices M_u (to the left) and M_v (to the right) on both sides of (2) and still denote by $Y, X_1, \dots, X_K, \Gamma^l, \Gamma^d, E$ the transformed matrices. She can apply a within-group (or first difference or Helmert) transform on the left to annihilate individual effects and a similar on the right to annihilate time effects, two matrices are required to annihilate group specific time effects. This is important if the researcher thinks there are individual and time effects and there could be additional interactive effects and wants to avoid relying on penalisation to figure out that there are classical individual and time effects. The regressors can be transformations of the baseline regressors as developed in Section 4.6 to ensure their operator norm is not too large, a feature sometimes useful in the analysis. We do not write these transformations explicitly to simplify the exposition.

3. First-stage estimator

The estimator is defined, for $\lambda > 0$, as

$$\left(\widehat{\beta}, \widehat{\Gamma}\right) \in \underset{\beta \in \mathbb{R}^K,\ \Gamma \in \mathcal{M}_{NT}}{\operatorname{argmin}} \frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^K \beta_k X_k - \Gamma \right|_2 + \frac{\lambda}{NT} |\Gamma|_*. \tag{3}$$

The nuclear norm is the ℓ_1-norm of the vector of singular values. Similarly to the ℓ_1-norm in the Lasso estimator, it yields low-rank solutions, that is a sparse vectors of singular value of $\widehat{\Gamma}$ (see Proposition 7 for a formal result). This estimator can be viewed as a type of square-root Lasso estimator of [5] for parameters which are matrices. As for the square-root Lasso, the ℓ_2-norm is not squared in (3) which implies that we do not need to know the variance of E_{it} to choose the parameter λ. Under the assumptions of Proposition 4, the choice of λ amounts to the choice of $\{\phi_N\}$ but this can be made without knowledge of parameters of the class of DGP considered in the two cases analysed in the proposition.

Proposition 1. *A solution* $\left(\widehat{\beta}, \widehat{\Gamma}\right)$ *of* (3) *is such that*

$$\widehat{\Gamma} \in \arg \min_{\Gamma \in \mathcal{M}_{NT}} \frac{1}{\sqrt{NT}} \left| M_X \left(Y - \Gamma \right) \right|_2 + \frac{\lambda}{NT} |\Gamma|_*.$$

For $u \geq 0$, $u = \min_{\sigma > 0} \left\{ \frac{\sigma}{2} + \frac{1}{2\sigma} u^2 \right\}$ and the minimum is attained at $\sigma = u$ if $u > 0$ or using minimizing sequences going to 0 if $u = 0$. Thus any solution $\left(\widehat{\beta}, \widehat{\Gamma}\right)$ of (3) is solution of

$$\left(\widehat{\beta}, \widehat{\Gamma}, \widehat{\sigma}\right) \in \underset{\beta \in \mathbb{R}^k, \Gamma \in \mathcal{M}_{NT}, \sigma > 0}{\operatorname{argmin}} \sigma + \frac{1}{\sigma NT} \left| Y - \sum_{k=1}^K \beta_k X_k - \Gamma \right|_2^2 + \frac{2\lambda}{NT} |\Gamma|_* \tag{4}$$

and

$$\widehat{\sigma} = \frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k - \widehat{\Gamma} \right|_2 . \tag{5}$$

This objective function in (4) has the advantage that the new objective function only has one nonsmooth convex function in (β, Γ): the nuclear norm. Because $f(x, y) = x^2/y$ is convex on the domain $\{(x, y) \in \mathbb{R}^2 | y > 0\}$, the objective function is convex in (β, Γ, σ). This formulation is analogous to the concomitant Lasso or scaled-Lasso for linear regression (see [23, 29]).

3.1. **First-order conditions and consequences.** The formulation is used in Section 3.2 for implementation of our estimator. It is also useful to obtain by subdifferential calculus the first order-conditions of program (3). Indeed, the differential with respect to β_k at (β, Γ, σ) on the domain (hence $\sigma > 0$) is, for $k = 1, \dots, K$,

$$- \frac{2}{\sigma NT} \left\langle X_k, Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma \right\rangle \tag{6}$$

and the subdifferential with respect to Γ at (β, Γ, σ) (see (2.1) in [17]) is

(7)
$$\left\{ -\frac{2}{\sigma NT} \left(Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma \right) + \frac{2\lambda}{NT} Z,\ Z = \sum_{k=1}^{\mathrm{rank}(\Gamma)} u_k(\Gamma) v_k(\Gamma)^\top + M_{u(\Gamma)} W M_{v(\Gamma)},\ |W|_{\mathrm{op}} \le 1 \right\},$$

in particular $|Z|_{\mathrm{op}} \le 1$ and $\langle \Gamma, Z \rangle = |\Gamma|_*$. Due to (5), if $\widehat{\sigma} = 0$ then clearly $\widehat{\beta}$ is the least-squares estimator which minimizes $\left| Y - \sum_{k=1}^{K} \beta_k X_k - \widehat{\Gamma} \right|_2^2$. Else, setting (6) to 0 at $\left(\widehat{\beta}, \widehat{\Gamma}, \widehat{\sigma}\right)$ yields the same conclusion. Hence, if $X^\top X$ is positive definite, we have

$$\widehat{\beta} = \left(X^\top X \right)^{-1} X^\top (y - \widehat{\gamma}). \tag{8}$$

Because, if $\widehat{\sigma} > 0$, 0 belongs to the set defined in (7) at $\left(\widehat{\beta}, \widehat{\Gamma}, \widehat{\sigma}\right)$, there exists $\widehat{W} \in \mathcal{M}_{NT}$ and $\widehat{Z} = \sum_{k=1}^{\mathrm{rank}(\widehat{\Gamma})} u_k\left(\widehat{\Gamma}\right) v_k\left(\widehat{\Gamma}\right)^\top + M_{u\left(\widehat{\Gamma}\right)} \widehat{W} M_{v\left(\widehat{\Gamma}\right)}$ such that $\left|\widehat{W}\right|_{\mathrm{op}} \le 1$ and $Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k - \widehat{\Gamma} = \lambda \widehat{\sigma} \widehat{Z}$, hence, for all $k = 1, \dots, K$, $\left\langle X_k, \widehat{Z} \right\rangle = 0$, thus $M_X\left(\widehat{Z}\right) = \widehat{Z}$ and

$$Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k - \widehat{\Gamma} = M_X\left(Y - \widehat{\Gamma}\right) = \lambda \widehat{\sigma} \widehat{Z}. \tag{9}$$

Again, due to (5), if $\widehat{\sigma} = 0$ then (9) also holds. As a consequence, we have

$$\widehat{\sigma} = \frac{1}{\sqrt{NT}} \left| M_X\left(Y - \widehat{\Gamma}\right) \right|_2$$

and any solution $\left(\widehat{\beta}, \widehat{\Gamma}\right)$ of (3) is also solution of

$$\left(\widehat{\beta}, \widehat{\Gamma}\right) \in \underset{\beta \in \mathbb{R}^K,\ \Gamma \in \mathcal{M}_{NT}}{\operatorname{argmin}} \frac{1}{NT} \left| Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma \right|_2^2 + \frac{2\lambda\widehat{\sigma}}{NT} |\Gamma|_*. \tag{10}$$

So $\left(\widehat{\beta}, \widehat{\Gamma}\right)$ given by (3) is a solution to a type of matrix Lasso estimator with data-driven penalty $\lambda\widehat{\sigma}|\Gamma|_*/NT$. The estimator in [21] corresponds to (10) without the data-driven $\widehat{\sigma}$.

Remark 1. *Due to the nuclear norm,* (9) *and the expression of* $\widehat{Z}$ *yield*

$$\left(Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k\right) M_{v(\widehat{\Gamma})} = \lambda\widehat{\sigma} M_{u(\widehat{\Gamma})} \widehat{W} M_{v(\widehat{\Gamma})}$$

which, unlike [3]*, is not zero. Applying the annihilator* $M_{u(\widehat{\Gamma})}$ *does not change this.*

3.2. **Computational aspect.** Based on (4), where the objective function is convex, we can iteratively minimize over β, Γ, and σ: start from $\left(\beta^{(0)}, \Gamma^{(0)}, \sigma^{(0)}\right)$ and repeat, for $t \in \mathbb{N}_0$ until convergence,

(1) $\beta^{(t+1)}$ is obtained by least-squares minimizing $\left|Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma^{(t)}\right|_2^2$,

(2) Setting $Z^{(t+1)} = Y - \sum_{k=1}^{K} \beta_k^{(t+1)} X_k$, $\Gamma^{(t+1)}$ is obtained by solving the matrix Lasso

$$\min_{\Gamma} \left|Z^{(t+1)} - \Gamma\right|_2^2 + 2\lambda\sigma^{(t)} |\Gamma|_* ,$$

i.e. applying soft-thresholding to the singular value decomposition (henceforth SVD)

$$\Gamma^{(t+1)} = \sum_{k=1}^{\min(N,T)} \left(\sigma_k\left(Z^{(t+1)}\right) - \lambda\sigma^{(t)}\right)_+ u_k\left(Z^{(t+1)}\right) v_k\left(Z^{(t+1)}\right)^\top ,$$

(3) $\sigma^{(t+1)} = \left|Z^{(t+1)} - \Gamma^{(t+1)}\right|_2 / \sqrt{NT}$.

Remark 2. *The estimator in* [21] *can be obtained by repeating (1) and (2) for a fixed value of* $\sigma^{(t)}$*.* $\lambda_N\sigma^{(t)}$ *corresponds to* $\sqrt{NT}\Psi_{NT}$ *in their notations and they assume* $1/\left(\Psi_{NT}\sqrt{\min(N,T)}\right) + \Psi_{NT} \to 0$ *to circumvent the unavailability of an upper bound on the variance of the errors. The method in* [3] *considers the number* r *of factors fixed and iterates step (1) and a modified step (2) where* $\lambda = 0$ *and under the restriction that* $\mathrm{rank}(\Gamma) = r$*, from which we extract the factor and factor loadings. The second step corresponds to hard-thresholding the SVD of* $Z^{(t+1)}$ *to keep only the part corresponding to the* r *largest singular values. This can be written*

$$\left(\widetilde{\beta}, \widetilde{\Gamma}\right) \in \underset{\substack{b\in\mathbb{R}^K \\ \Gamma\in\mathcal{M}_{NT}:\ \mathrm{rank}(\Gamma)=r}}{argmin} \left|Y - \sum_{k=1}^{K} b_k X_k - \Gamma\right|_2^2 .$$

It is argued that iterating (a) least-squares given factors and (b) PCA to obtain the r *common factors is less numerically robust. By partialling out, (a) corresponds to minimizing*

$$\left|\left(Y - \sum_{k=1}^{K} \beta_k X_k\right) M_{v\left(\Gamma^{(t)}\right)}\right|_2^2 .$$

4. Results

4.1. **Error bound on the estimation of Γ.** A key quantity is the compatibility constant (see [6]) defined, for each realization of X and all $A \in \mathcal{M}_{NT}$, by

$$\kappa_{A,c} = \inf_{\Delta \in C_{A,c}:\ \Delta \neq 0} \frac{\sqrt{2\text{rank}\,(A)}|M_X(\Delta)|_2}{|\mathcal{P}_A\,(\Delta)|_*}.$$

A few remarks are in order. First, if $X = 0$, we have $M_X(\Delta) = \Delta$. Second, the denominator of the ratio cannot be 0 because, for $\Delta \in C_{A,c}$, $|\Delta|_* \le (1+c)\,|\mathcal{P}_A\,(\Delta)|_*$, hence the function of Δ in the infimum is continuous. Third, because the ratio involves two linear operators, the infimum is the same if we restrict Δ to have norm 1 and the intersection with the cone is compact. Hence, the infimum is a minimum. Fourth, for all $A \in \mathcal{M}_{NT}$ and $c > 0$, the minimum is the limit of minima over finite sets so it is a measurable function of X. Fifth, we work with $\kappa_{\widetilde{\Gamma},c}$ for a random $\widetilde{\Gamma}$ which depends on the random Γ and X via $\kappa_{\widetilde{\Gamma},c}$ itself and we allow Γ and X to be dependent. We make a slight abuse of notations and consider that $\kappa_{\widetilde{\Gamma},c}$ is a measurable function of both inputs $\widetilde{\Gamma}$ and X. In practice, it is a measurable lower bound on it for every fixed $\widetilde{\Gamma} \in \mathcal{M}_{NT}$ and X in the support of the corresponding random matrix.

Remark 3. *When $X = 0$ one has, for all $A \in \mathcal{M}_{NT}$ and $c > 0$, $\kappa_{A,c} \ge 1$.*

Proposition 2. *The following lower bounds hold*

$$\kappa_{A,c} \ge \min_{\Delta \in C_{A,c}:\ \Delta \neq 0} \frac{|M_X(\Delta)|_2}{|\mathcal{P}_A\,(\Delta)|_2} \ge \min_{\Delta \in C_{A,c}:\ \Delta \neq 0} \frac{|M_X(\Delta)|_2}{|\Delta|_2}. \tag{11}$$

The quantity in the middle is the restricted eigenvalue (see [17]). The smaller one is used in [21]. Throughout the rest of the paper, $\rho \in (0,1)$ and define

$$c\,(\rho,\widetilde{\rho}) = \frac{1+\rho+\widetilde{\rho}}{1-\rho},\ d\,(\rho,\widetilde{\rho}) = \max\left(1+\widetilde{\rho}, \rho\,(1+c\,(\rho,\widetilde{\rho}))\right),\ e\,(\rho,\widetilde{\rho}) = d\,(\rho,\widetilde{\rho}) + \rho\,(1+c\,(\rho,\widetilde{\rho}))\,,$$

$$\theta_\infty\left(\widetilde{\Gamma},\rho,\widetilde{\rho}\right) = 2\left(1-\left(\frac{d\,(\rho,\widetilde{\rho})\sqrt{2\text{rank}\left(\widetilde{\Gamma}\right)}\lambda}{\sqrt{NT}\kappa_{\widetilde{\Gamma},c(\rho,\widetilde{\rho})}}\right)^2\right)_+^{-1} e\,(\rho,\widetilde{\rho})\,,$$

$$\theta\,(\rho,\widetilde{\rho}) = \inf_{\widetilde{\Gamma}\in\mathcal{M}_{NT}} \max\left(\theta_\infty\left(\widetilde{\Gamma},\rho,\widetilde{\rho}\right)\frac{\lambda\text{rank}\left(\widetilde{\Gamma}\right)|M_X(E)|_2}{\sqrt{NT}\kappa^2_{\widetilde{\Gamma},c(\rho,\widetilde{\rho})}}, \frac{1}{\widetilde{\rho}}\left|\Gamma-\widetilde{\Gamma}\right|_*\right),$$

$$\theta_*(\rho) = \inf_{\widetilde{\rho}>0}\,(1+c\,(\rho,\widetilde{\rho}))\,\theta\,(\rho,\widetilde{\rho})\,,\ \theta_\sigma(\rho) = \inf_{\widetilde{\rho}>0} d\,(\rho,\widetilde{\rho})\,\theta\,(\rho,\widetilde{\rho})\,.$$

Theorem 1. *If $\rho\lambda|M_X(E)|_2/\sqrt{NT} \ge |M_X(E)|_{\text{op}}$, we have*

$$\left|\widehat{\Gamma}-\Gamma\right|_* \le 2\theta_*(\rho), \tag{12}$$

$$\left|\widehat{\sigma} - \frac{1}{\sqrt{NT}}\,|M_X\,(E)|_2\right| \le \frac{2\lambda}{NT}\theta_\sigma(\rho). \tag{13}$$

Note that $\theta_*(\rho) \leq \theta_\sigma(\rho)/\rho$. For example, we can take $\widetilde{\rho} = 1$ and $\rho = 2/5$, in which case $c(\rho, \widetilde{\rho}) = 4$, $d(\rho, \widetilde{\rho}) = 2$, $e(\rho, \widetilde{\rho}) = 4$, $\theta_*(\rho) = 5\theta(\rho, \widetilde{\rho})$, and $\theta_\sigma(\rho) = 2\theta(\rho, \widetilde{\rho})$. We state a more general result to allow the theory to handle the case where ρ is close to 1 which allows a smaller λ (what matters is the product $\rho\lambda$) and we find works well in small samples. The result of Theorem 1 is in the spirit of a low-rank oracle inequality. If we use the decomposition $\Gamma = \Gamma^l + \Gamma^d$ in (2), where Γ^l has low-rank and Γ^d could have high-rank but has small nuclear norm, and take $\widetilde{\Gamma} = \Gamma^l$ in the maximum in the expression of θ_* we obtain

$$\max\left(\theta_\infty\left(\Gamma^l, \rho, \widetilde{\rho}\right) \frac{\lambda \mathrm{rank}\left(\Gamma^l\right) |M_X(E)|_2}{\sqrt{NT}\kappa^2_{\Gamma^l, c\left(\rho, \widetilde{\rho}\right)}}, \frac{1}{\widetilde{\rho}}\left|\Gamma^d\right|_*\right)$$

and the upper bounds in Theorem 1 depend on both nuisance parameters. In the usual setup where $\Gamma = \Gamma^l$, we can drop $\left|\Gamma^d\right|_*$ from the maximum and obtain a more classical bound which depends on rank(Γ). The term involving $(\cdot)_+^{-1}$ in the definition of $\theta_\infty\left(\widetilde{\Gamma}, \rho, \widetilde{\rho}\right)$ could be ∞ if $\kappa_{\widetilde{\Gamma}, c\left(\rho, \widetilde{\rho}\right)}$ is too small. This term appears because we do not assume a priori knowledge on the variance of the errors or use a sequence of penalties that diverge too fast. A small constant $c(\rho, \widetilde{\rho})$ implies a small cone and a large value of $\kappa_{\widetilde{\Gamma}, c\left(\rho, \widetilde{\rho}\right)}$. The difference between the upper bound in Theorem 1 and a genuine oracle inequality is that the right-hand side is random due to the randomness of Γ and X.

4.2. **Restriction on the joint distribution of X and E.** We maintain the following baseline assumption on the DGP.

Assumption 1. *The following hold:*

(i) *There exists $\sigma > 0$ such that $|E|_2^2/(NT) \xrightarrow{\mathbb{P}} \sigma^2$,*
(ii) *There exists $\Sigma \in \mathcal{M}_{KK}$ positive definite such that $X^\top X/(NT) \xrightarrow{\mathbb{P}} \Sigma$,*
(iii) *$X^\top e = O_P\left(\sqrt{NT}\right)$,*
(iv) *There exists $\{\mu_N\}$ such that $\mu_N = O(\sqrt{NT})$ and $\sum_{k=1}^K |X_k|_{\mathrm{op}}^2 = O_P\left(\mu_N^2\right)$.*

Condition (iv) is not restrictive if $\mu_N = \sqrt{NT}$ due to (ii). The role of (iv) is to introduce the notation $\{\mu_N\}$.

Proposition 3. *Under Assumption 1 with $\mu_N = \sqrt{NT}$, we have*

$$\left|\frac{|M_X(E)|_2}{\sqrt{NT}} - \sigma\right| = O_P\left(\frac{1}{\sqrt{NT}}\right) \tag{14}$$

$$\left||M_X(E)|_{\mathrm{op}} - |E|_{\mathrm{op}}\right| = O_P\left(\frac{\mu_N}{\sqrt{NT}}\right). \tag{15}$$

Based on Theorem 1 and Proposition 3 the researcher should choose $\{\lambda_N\}$ as follows.

Assumption 2. *Maintain Assumption 1 and, given an upper bound μ_N for Assumption 1* (iv), *take $\{\lambda_N\}$ of the form*

$$\lambda_N = \left(1 - \frac{\phi_{1N}}{\sqrt{NT}}\right)^{-1} \left(\psi_N + \phi_{2N}\frac{\mu_N}{\sqrt{NT}}\right), \tag{16}$$

where $\{\phi_{1N}\}$ and $\{\phi_{2N}\}$ are arbitrary sequence going to infinity, as slowly as we want but no faster than $\sqrt{NT}$ for $\{\phi_{1N}\}$, and

(i) $\psi_N = O\left(\sqrt{NT}\right)$,
(ii) $\lim_{N\to\infty} \mathbb{P}\left(\rho\psi_N\sigma \geq |E|_{\rm op}\right) = 1$.

We can take $\phi_1 = \phi_2$ in which case we write $\phi = \phi_1 = \phi_2$. (16) holds whether $\mu_N = \sqrt{NT}$ or we have a sharper bound on it. Under the premises of Section 4.6, we can take $\mu_N = \lambda_N$ and

$$\lambda_N = \left(1 - \frac{\phi_N}{\sqrt{NT}}\right)^{-1} \psi_N. \tag{17}$$

The event $\mathcal{E} = \left\{\rho\lambda_N \left|M_X(E)\right|_2 / \sqrt{NT} \geq \left|M_X(E)\right|_{\rm op}\right\}$ can be written

$$\mathcal{E} = \left\{\rho\psi_N\sigma + \rho\frac{\phi_{2N}\mu_N}{\sqrt{NT}}\sigma + \rho\frac{\phi_{1N}\lambda_N}{\sqrt{NT}}\sigma \geq |E|_{\rm op} + \left(|M_X(E)|_{\rm op} - |E|_{\rm op}\right) + \rho\lambda_N\left(\sigma - \frac{|M_X(E)|_2}{\sqrt{NT}}\right)\right\},$$

hence

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}\left(\left\{\rho\psi_N\sigma \geq |E|_{\rm op}\right\} \bigcap \left\{\rho\frac{\phi_{2N}\mu_N}{\sqrt{NT}}\sigma \geq |M_X(E)|_{\rm op} - |E|_{\rm op}\right\} \bigcap \left\{\frac{\phi_{1N}}{\sqrt{NT}}\sigma \geq \sigma - \frac{|M_X(E)|_2}{\sqrt{NT}}\right\}\right)$$

and the 3 events have probability going to 1 by (ii) and Proposition 3 so $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}) = 1$.

We can handle large classes of joint distributions of X and E, including ones where the errors have heavy tails. Else, important cases are such that $|E|_{\rm op} = O_P\left(\sqrt{\max(N,T)}\right)$ (see [22, 30] and Appendix A.1 in [19]). For such distributions, it is enough to take $\psi_N = C\sqrt{\max(N,T)}$ for large enough C for Assumption 2 to hold. An easy way to circumvent thr problem that C is unknown is to take $\psi_N = \phi_{2N}\sqrt{\max(N,T)}$ but this results in over penalization. At the cost of additional assumptions on the distribution, one can obtain the following more precise proposal based on Corollary 5.35 and Theorem 5.31 in [30].

Proposition 4. *If $E = M_u\eta M_v$, where M_u and M_v are, possibly random, matrices such that $|M_u|_{\rm op} \leq 1$ and $|M_v|_{\rm op} \leq 1$ and either of the following holds*

(i) *$\{\eta_{it}\}_{i,t}$ are i.i.d. centered Gaussian random variables,*
(ii) *$\{\eta_{it}\}_{i,t}$ are i.i.d. centered random variables with finite fourth moments and T/N converges to a constant in $[0,1]$,*

then the sequence defined by $\psi_N = \left(\sqrt{N} + \sqrt{T}\right)/\rho + \varphi_N$, where $\varphi_N \to \infty$ arbitrarily slowly in case (i) *and $\left\{\varphi_N/\sqrt{T}\right\}$ is bounded away from 0 in case* (ii), *satisfies Assumption 2* (ii).

The matrices M_u and M_v can be known or estimated (see, *e.g.*, Section 4.6) and have been applied to the data. Applying such matrices cannot increase rank $\left(M_u\Gamma^l M_v\right)$, $\left|M_u\Gamma^d M_v\right|_{\text{op}}$, or $|M_u\eta M_v|_{\text{op}}$. These matrices can be unknown and the baseline error E can have temporal and cross-sectional dependence. Because the operator norm of a matrix is equal to the operator norm of its transpose, the role of N and T can be exchanged in (ii). The proposed choice of the penalty level is almost completely explicit and does not depend on the variance of the errors. The remaining sequences are arbitrary. In contrast to (16) where $\left(1-\phi_{1N}/\sqrt{NT}\right)^{-1}$ converges to 1, [21] employs a factor converging to infinity. Hence, the data-driven method of this paper provides less shrinkage, less bias, and a better bias/variance tradeoff.

4.3. **Restriction on the joint distribution of X and Γ.** We now discuss restrictions so that the bounds in Theorem 1 are small.

Assumption 3. *The random matrix Γ can be decomposed as $\Gamma=\Gamma^l+\Gamma^d$, where, for $\{r_N\}$,*

(*i*) rank $\left(\Gamma^l\right)=O_P(r_N)$,
(*ii*) $\left|\Gamma^d\right|_*=O_P\left(\lambda_N r_N\right)$,
(*iii*) *There exists $\kappa>0$ independent of N such that $\kappa_{\Gamma^l}\ge\kappa$ w.p.a. 1.*

Based on the expression of $\theta_*(\rho)$ and $\theta_\sigma(\rho)$, Theorem 1, and Proposition 3, a tight decomposition of the form $\Gamma=\Gamma^l+\Gamma^d$ implies that Γ^l and Γ^d are functions of X and Γ.

Proposition 5. *Assumption 3* (iii) *for a cone with constant c holds with the lower bound κ if, w.p.a. 1, $\kappa^2+2\text{rank}\left(\Gamma^l\right)Q(b,b_\perp)\le 1$, where $b,b_\perp\in\mathbb{R}^K$ are defined, for $k=1,\dots,K$, as $b_k=a\min\left(\left|\mathcal{P}_{\Gamma^l}\left(X_k\right)\right|_{\text{op}},|X_k|_{\text{op}}\right)$, $b_{\perp k}=a\left|\mathcal{P}^\perp_{\Gamma^l}\left(X_k\right)\right|_{\text{op}}$, $a=\left|X^\top X/(NT)\right|_{\text{op}}^{-1}|X|_2/(NT)$,*

$$\begin{aligned}Q(b,b_\perp)=&|b|_2^2\mathbb{1}\left\{p_N\left|b_\perp\right|_2^2\ge 1\right\}+\left(|b+b_\perp c|_2^2-\frac{c^2}{p_N}\right)\mathbb{1}\left\{1-\frac{p_N\langle b_\perp,b\rangle}{c}\le p_N\left|b_\perp\right|_2^2<1\right\}\\&+\left(\left|b+b_\perp\frac{p_N\langle b_\perp,b\rangle}{1-p_N\left|b_\perp\right|_2^2}\right|_2^2-\frac{p_N\langle b_\perp,b\rangle^2}{\left(1-p_N\left|b_\perp\right|_2^2\right)^2}\right)\mathbb{1}\left\{p_N\left|b_\perp\right|_2^2<1-\frac{p_N\langle b_\perp,b\rangle}{c}\right\},\end{aligned}$$

and $p_N=\min\left(N-\text{rank}\left(\Gamma^l\right),T-\text{rank}\left(\Gamma^l\right)\right)$.

Note that $Q(b,b_\perp)<|b+b_\perp c|_2^2$ and, if $K=1$, $a=1/\left|X_1\right|_2$ and

$$|b+b_\perp c|_2^2=\frac{1}{|X_1|_2^2}\left(\min\left(\left|\mathcal{P}_{\Gamma^l}\left(X_1\right)\right|_{\text{op}},|X_1|_{\text{op}}\right)+\left|\mathcal{P}^\perp_{\Gamma^l}\left(X_1\right)\right|_{\text{op}}c\right)^2.$$

The quantity $\left|\mathcal{P}^\perp_{\Gamma^l}\left(X_k\right)\right|_{\text{op}}=\left|M_{u(\Gamma^l)}X_kM_{v(\Gamma^l)}\right|_{\text{op}}$ in the definition of $b_{\perp k}$ can be not too large because the projectors can reduce the operator norm if X_k has a component with a factor structure and shares some factors in common with Γ^l which are annihilated by $M_{v(\Gamma^l)}$ (see Remark 5 for further discussion of this aspect). Due to Assumption 1 (ii), $a=O_P\left(1/\sqrt{NT}\right)$.

In the worst case, by the crude bound $|X_k|_{\mathrm{op}} \le |X_k|_2$, b and $b_\perp$, hence $Q(b, b_\perp)$ are bounded. If $\mu_N = o\left(\sqrt{NT}\right)$, the condition in Proposition holds for arbitrary constants $\kappa < 1$ for N large enough, but this is not necessary. Section 4.6 presents solutions to work with regressors with smaller operator norm. Lemma A.7 in [21] provides an alternative sufficient condition for Assumption 3 (ii). Lemma A.8 is another sufficient condition when $K = 1$. In our framework r_{1N} can grow, c can be different from 3, and we do not work contionnal on Γ^l, condition (iii) has to be modified with a denominator of $\sqrt{NTr_N}$ and the probabilities are with respect to the distribution of (Γ, X_1). It is claimed in Remark (a) in [21] that the condition in Lemma A.8 holds when $X_1 = \Pi_1^l + U_1$, Π_1^l has a fixed rank, and U_1 has iid mean zero normal entries.

4.4. **Rates of convergence.** Theorem 1 and the assumptions on the DGP yield the following.

Theorem 2. *Under assumptions 2 and 3,*

$$\left|\widehat{\Gamma} - \Gamma\right|_* = O_P\left(\lambda_N r_N\right), \tag{18}$$

$$\widehat{\sigma} - \sigma = O_P\left(\frac{\lambda_N^2 r_N}{NT}\right), \tag{19}$$

$$\widehat{\beta} - \beta = O_P\left(\frac{\lambda_N r_N \mu_N}{NT}\right). \tag{20}$$

In (20), we have implicitly assumed that $\sqrt{NT} = O\left(\lambda_N r_N \mu_N\right)$ but this always occurs when $X \neq 0$ and the problem is to have $\lambda_N r_N \mu_N$ as close as possible in rate to $\sqrt{NT}$. Under usual assumptions where we can take λ_N proportional to $\sqrt{\max(N,T)}$, r_N fixed, and make no restriction on $\{\mu_N\}$ so that $\mu_N = O(\sqrt{NT})$, we obtain the rate convergence of $1/\sqrt{\min(N,T)}$ which is the one in [21]. Theorem 2 shows that $\widehat{\beta}$ remains consistent if $r_N = o\left(\sqrt{\min(N,T)}\right)$. Obviously r_N can be larger if μ_N is smaller. The most favorable situation, when $\mu_N = O\left(\sqrt{\max\left(N,T\right)}\right)$ and λ_N is proportional to $\sqrt{\max(N,T)}$, yields

$$\widehat{\beta} - \beta = O_P\left(\frac{\max(N,T) r_N}{NT}\right),$$

hence, when N/T converges to a constant, this becomes $O_P\left(r_N/\sqrt{NT}\right)$. Achieving $\mu_N = o\left(\sqrt{NT}\right)$ and in some cases $\mu_N = O\left(\sqrt{\max\left(N,T\right)}\right)$ using transformed regressors is sometimes possible under the premises of Section 4.6 and this paper allows to obtain such an estimator and transformed regressors in a data-driven way. Section 4.7 proposes an alternative approach where we can obtain the $1/\sqrt{NT}$ rate and to have asymptotic normality.

4.5. **Additional results using the relation to the matrix Lasso.** Recall that any solution $\left(\widehat{\beta}, \widehat{\Gamma}\right)$ of (3) is also solution of (10). Based on this we can prove the following additional results on our estimator which would also apply to (10) even if rather than $\widehat{\sigma}$ we use an upper bound on the standard error of the errors. The results that we state involve $\widehat{\sigma}$ but, under the assumptions of Theorem 2, $\widehat{\sigma}$ is a consistent estimator of σ. In order to guarantee $\mathbb{P}\left(\rho\lambda_N \min\left(\widehat{\sigma}, \sigma\right) \ge |M_X(E)|_{\mathrm{op}}\right) \to 1$ we need the following assumption.

Assumption 4. *Assumption 2 holds and* $\{\phi_{1N}\}$ *satisfies the additional restriction that, for* N *large enough,*

$$\left(1-\frac{\phi_{1N}}{\sqrt{NT}}\right)^2\phi_{1N}\geq\phi_{2N}\frac{r_N}{\sqrt{NT}}\left(\psi_N+\phi_{2N}\frac{\mu_N}{\sqrt{NT}}\right)^2.$$

Indeed, we can replace $\left\{\phi_{1N}\sigma/\sqrt{NT}\geq\sigma-|M_X(E)|_2/\sqrt{NT}\right\}$ by $\{\phi_{2N}\sigma\lambda_N^2 r_N/NT\geq\sigma-\widehat{\sigma}\}$ in the previous analysis which converges to 1 due to (19) because, due to Assumption 4, $\phi_{1N}\geq\phi_{2N}\lambda_N^2 r_N/\sqrt{NT}$, hence

$$\left(1-\phi_{2N}\frac{\lambda_N^2 r_N}{NT}\right)\lambda_N\geq\left(1-\frac{\phi_{1N}}{\sqrt{NT}}\right)\lambda_N=\psi_N+\phi_{2N}\frac{\mu_N}{\sqrt{NT}}.$$

A conservative choice is $\phi_{1N}=c_1\sqrt{NT}$ for a small $c_1\in(0,1)$. Now on, we use cones with constant $c=c(\rho)=(1+\rho)/(1-\rho)$. First, with a proof similar to the computations in [17], we obtain a result which is an oracle inequality with constant 1 if X and Γ are not random.

Proposition 6. *If* $\rho\lambda\min(\widehat{\sigma},\sigma)\geq|M_X(E)|_{\mathrm{op}}$, *we have*

$$\frac{1}{NT}\left|M_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2\leq\inf_{\widetilde{\Gamma}}\left\{\frac{1}{NT}\left|M_X\left(\Gamma-\widetilde{\Gamma}\right)\right|_2^2+\frac{2(\lambda(1+\rho)\min(\widehat{\sigma},\sigma))^2}{NT}\frac{\mathrm{rank}\left(\widetilde{\Gamma}\right)}{\kappa^2_{\widetilde{\Gamma},c(\rho)}}\right\}.$$

This inequality yields a slightly different notion of approximately sparse solution because the first term in the maximum involves $\left|M_X\left(\Gamma-\widetilde{\Gamma}\right)\right|_2^2/(NT)$ rather than $\left|\Gamma-\widetilde{\Gamma}\right|_*$. The next result provides a bound on $\mathrm{rank}\left(\widehat{\Gamma}\right)$ as a function of the previous bound.

Proposition 7. *If* $\rho\lambda\widehat{\sigma}\geq|M_X(E)|_{\mathrm{op}}$ *then we have*

$$\left(\lambda(1-\rho)\widehat{\sigma}-\left|\Gamma^d\right|_{\mathrm{op}}\right)_+^2\mathrm{rank}\left(\widehat{\Gamma}\right)\leq\left|P_{u(\widehat{\Gamma})}M_X\left(\Gamma^l-\widehat{\Gamma}\right)P_{v(\widehat{\Gamma})}\right|_2^2\leq\left|M_X\left(\Gamma^l-\widehat{\Gamma}\right)\right|_2^2.$$

As a result, under the above conditions and Assumtion 3 (ii),

$$\mathrm{rank}\left(\widehat{\Gamma}\right)\leq 2\left((1+\rho)/((1-\rho)\kappa_{\Gamma^l,c(\rho)})\right)^2\mathrm{rank}\left(\Gamma^l\right).$$

We can combine propositions 6 and 7 with Proposition 11 in the appendix to obtain results for other loss functions, in particular the Frobenius norm.

Our estimator has desirable low-rank properties but it can fail to obtain rank(Γ), $\mathrm{rank}\left(\Gamma^l\right)$, or annihilator matrices. Thus, we introduce the hard-thresholded estimator

$$\widehat{\Gamma}^t=\sum_{k=1}^{\mathrm{rank}(\widehat{\Gamma})}\sigma_k\left(\widehat{\Gamma}\right)\mathbb{1}\left\{\sigma_k\left(\widehat{\Gamma}\right)\geq t\right\}u_k\left(\widehat{\Gamma}\right)v_k\left(\widehat{\Gamma}\right)^\top.$$

Proposition 8. *Under the assumptions of Theorem 2 and Assumption 4, if* $\left|\Gamma^d\right|_{\mathrm{op}} = o_P\left(\lambda_N \sigma\right)$, *we have*

$$\max\left(\left|\Gamma - \widehat{\Gamma}\right|_{\mathrm{op}}, \left|\Gamma^l - \widehat{\Gamma}\right|_{\mathrm{op}}\right) \le (\rho+1)\lambda_N\left(\sigma + O_P\left(\frac{r_N \mu_N^2}{NT}\right)\right). \tag{21}$$

Assumption 5. *Let $h > 1$. The following conditions hold*

(*i*) $r_N \mu_N^2 = o(NT)$,
(*ii*) $\mathbb{P}\left(\sigma_{\mathrm{rank}(\Gamma^l)}\left(\Gamma^l\right) \ge (\rho+1)\lambda_N h^2(h+1)\sigma\right) \to 1$.

Condition (i) guarantees the O_P in (21) is $o_P(1)$. It allows the pivotal thresholding methods below but imposes a slight restriction on the operator norms of the regressors. Section 4.6 allows to come back to a case where (i) holds for a large class of regressors. Without (i)

$$\max\left(\left|\Gamma - \widehat{\Gamma}\right|_{\mathrm{op}}, \left|\Gamma^l - \widehat{\Gamma}\right|_{\mathrm{op}}\right) = O_P\left(\lambda_N\right)$$

and can adapt the results which follow at the expense of a theoretical but unfeasible thresholding level or using conservative levels $\lambda_N/t = o(1)$. Condition (ii) is weaker than a strong-factor assumption on Γ^l. We now show that we can recover $\mathrm{rank}(\Gamma)$ with a data-driven threshold.

Proposition 9. *Under the assumptions of Proposition 8 and Assumption 5, then setting $t = (\rho+1)\lambda_N h^2 \widehat{\sigma}$ yields*

$$\mathbb{P}\left(\mathrm{rank}\left(\widehat{\Gamma}^t\right) = \mathrm{rank}\left(\Gamma^l\right)\right) \to 1.$$

Moreover, if we remove (ii)*, then we have*

$$\mathbb{P}\left(\mathrm{rank}\left(\widehat{\Gamma}^t\right) \le \mathrm{rank}\left(\Gamma^l\right)\right) \to 1,$$

if we replace (ii) *by the weaker assumption* $\mathbb{P}\left(\sigma_{\mathrm{rank}(\Gamma^l)}\left(\Gamma^l\right) \ge (\rho+1)\lambda_N h^3\sigma\right) \to 1$, *we have*

$$\mathbb{P}\left(\mathrm{rank}\left(\widehat{\Gamma}^t\right) \ge \mathrm{rank}\left(\Gamma^l\right)\right) \to 1,$$

and

$$\max\left(\left|\Gamma - \widehat{\Gamma}^t\right|_{\mathrm{op}}, \left|\Gamma^l - \widehat{\Gamma}^t\right|_{\mathrm{op}}\right) \le (\rho+1)\lambda_N(h^2+1)\left(\sigma + o_P(1)\right). \tag{22}$$

We strengthen Assumption 5 (ii) as follows. When v_N increases like $\sqrt{NT}$, it is a strong-factor assumption.

Assumption 6. *Let $\{v_N\}$ be such that $v_N \ge (\rho+1)\lambda_N h^2(h+1)\sigma$. Assume that*

$$\mathbb{P}\left(\sigma_{\mathrm{rank}(\Gamma^l)}\left(\Gamma^l\right) \ge v_N\right) \to 1.$$

Proposition 10. *Under the assumptions of Proposition 9 and Assumption 6, we have*

$$\left|P_{v(\widehat{\Gamma}^t)} - P_{v(\Gamma^l)}\right|_2 = \left|M_{v(\widehat{\Gamma}^t_v)} - M_{v(\Gamma^l)}\right|_2 \le (\rho+1)\frac{\sqrt{2r_N}\lambda_N}{v_N}\left((h^2+1)\sigma + o_P(1)\right)$$
$$\left|P_{u(\widehat{\Gamma}^t)} - P_{u(\Gamma^l)}\right|_2 = \left|M_{u(\widehat{\Gamma}^t)} - M_{u(\Gamma^l)}\right|_2 \le (\rho+1)\frac{\sqrt{2r_N}\lambda_N}{v_N}\left((h^2+1)\sigma + o_P(1)\right).$$

Under a strong-factor assumption, when λ_N is proportional to $\sqrt{\max(N,T)}$ and r_N is fixed, we obtain the same rate of convergence as using PCA and as in Lemma A.7 in [3]. Here we obtain an upper bound with known constant. The rates that we obtain are also more general because we do not need to maintain the strong-factor assumption or that r_N is fixed, $\{\lambda_N\}$ could also allow for errors with larger tails of the operator norm.

4.6. **Working with transformed regressors.** In the previous sections, $\{\mu_N\}$ sometimes plays an important role and we might want it to be not too large. However, this can be as large as $O(\sqrt{NT})$ if the next assumption holds. So we devote this section to the analysis of this difficult situation.

Assumption 7. *For all* $k \in \{1, \dots, K\}$,

$$X_k = \Pi_k^l + \Pi_k^d + U_k, \tag{23}$$

and Π_k^d, U_k, σ_k, r_{kN}, λ_{kN}, *and* v_{kN} *play the role of* Γ^d, E, σ, r_N, λ_N *and* v_N *and satisfy the assumptions of Proposition 4, Assumption 3* (i) *and* (ii)*, and Assumption 5* (ii)*. We assume that for at least one* $k \in \{1, \dots, K\}$, $\Pi_k^l \neq 0$ *and* $|\Pi_k|_{\mathrm{op}} + |\Pi_k|_{\mathrm{op}}^{-1} = O_P\left(\sqrt{NT}\right)$.

The problem is difficult due to $\Pi_k^l \neq 0$ and $|\Pi_k|_{\mathrm{op}}^{-1} = O_P\left(\sqrt{NT}\right)$. No transformation is required if $\Pi_k^l = 0$ or if $\left|\Pi_k^l\right|_{\mathrm{op}} = o_P\left(\sqrt{NT}\right)$. The problem would be even harder if Π_k^l does not have a small rank (*i.e.*, with "many" strong factors) and there is obviously a problem related to identification when $X_k = \Pi_k^l$ and Π_k^l has small rank. Under the aforementioned assumptions, we can take $\lambda_{kN} = \lambda_N$. The matrix Π_k^l, σ_k, and the annihilators $M_{u(\Pi_k^l)}$ and $M_{v(\Pi_k^l)}$ can be estimated like in the previous sections and one can replace X_k by $\widetilde{X}_k$, where $X_k - \widetilde{X}_k$ has low rank, and Γ^l by $\widetilde{\Gamma}^l = \Gamma^l + \sum_{k=1}^K \beta_k \left(X_k - \widetilde{X}_k\right)$. For simplicity of exposition, we apply a transformation to all regressors. When $X = 0$, (3) can be computed as an iterated soft-thresholding estimator.

One can work with an estimator $\widetilde{\Pi}_k$ of Π_k of the form $\widetilde{\Pi}_k = \widehat{\Pi}_k$ or $\widetilde{\Pi}_k = \widehat{\Pi}_k^t$ obtained as described in the previous sections, with (1) $\widetilde{X}_k = X_k - \widetilde{\Pi}_k$, (2) $\widetilde{X}_k = M_{u\left(\widetilde{\Pi}_k\right)} X_k$, (3) $\widetilde{X}_k = X_k M_{v\left(\widetilde{\Pi}_k\right)}$, (4) $\widetilde{X}_k = \mathcal{P}^{\perp}_{\widetilde{\Pi}_k}(X_k)$, (5) $\widetilde{X}_k = X_k - X_k^{(l_k)}$ where $X_k^{(l_k)}$ is obtained from X_k by keeping the low rank component from a SVD corresponding to the $l_k = \mathrm{rank}\left(\widetilde{\Pi}_k\right)$ largest singular values. An alternative is to rely on Principal Component Analysis (henceforth PCA) using the eigenvalue-ratio (see [1]) to select the number of factors. By the previous results, using such transformed regressors gives rise to additional terms in $\widetilde{\Gamma}$ of rank each at most $18 r_{kN} + o_P(1)$ if $\Pi_k = \Pi_k^l$ or of same rank as $\widetilde{X}_k^l$ w.p.a. 1 if we use hard-thresholding as well. Assuming we transform all regressors, the rank of $\widetilde{\Gamma}$ is at most $\widetilde{r}_N + o_P(1)$, where $\widetilde{r}_N = r_N + 2((1+\rho)/(1-\rho))^2 \sum_{k=1}^K r_{kN}$ if $\widetilde{\Pi}_k = \widehat{\Pi}_k$ and $l_k = \mathrm{rank}\left(\widehat{\Pi}_k\right)$ and else $\widetilde{r}_N = r_N + \sum_{k=1}^K r_{kN}$. Using $\widetilde{\Pi}_k = \widehat{\Pi}_k^t$ has the advantage that if $\Pi_k^d \neq 0$ we have guarantees on the low rank of $\widetilde{\Gamma}$.

Remark 4. *In Assumption 7 we have assumed that we maintain the assumption of Proposition 4 and Assumption 5* (ii) *for simplicity of exposition. But we can also handle heavy tailed errors U_k by choosing a penalty level λ_{kN} large enough as disscussed before Proposition 4 . We maintain Assumption 5* (ii) *to allow for a simple thresholding rule but it is enough to use a thresholding at any level of smaller order than $\sqrt{NT}$ to obtain $\mu_N = o\left(\sqrt{N}\right)$.*

4.7. **Second-stage estimator of β.** As seen at the end of Section 4.4, the estimator $\widehat{\beta}$ could sometimes achieve the $1/\sqrt{NT}$ rate. But under weaker conditions we obtain a slower rate of convergence. This section presents three different two-stage approaches which deliver an asymptotically normal estimator of β.

4.7.1. *Approach 1: Annihilation of low-rank components of Γ and the regressors.* This section analyzes another approach under Assumption 7 where, for simplicity of exposition, the last statement holds for all regressors, and we use the transformed regressors with transformation (1) or (2). We obtain estimators of $\Pi_u^l = \left(\Gamma^l, \Pi_1^l, \dots, \Pi_K^l\right)$ and $\Pi_v^l = \left(\left(\Gamma^l\right)^\top, \left(\Pi_1^l\right)^\top, \dots, \left(\Pi_K^l\right)^\top\right)^\top$ by plug-in using $\widetilde{\Pi}_k = \widehat{\Pi}_k$ or $\widetilde{\Pi}_k = \widehat{\Pi}_k^t$ (preferably) for $k = 1, \dots, K$ and

$$\widehat{\Gamma} = \widehat{\widetilde{\Gamma}} - \sum_{k=1}^{K} \widehat{\beta}_k \widetilde{\Pi}_k. \tag{24}$$

We denote by $\widehat{\Pi}_u$ and $\widehat{\Pi}_v$ the estimators, by $\overline{\sigma}^2 = \sigma^2 + \sum_{k=1}^K \sigma_k^2$ and $\widehat{\overline{\sigma}}^2 = \widehat{\sigma}^2 + \sum_{k=1}^K \widehat{\sigma}_k^2$, by $\widetilde{\sigma} = \overline{\sigma}$ and $\widehat{\widetilde{\sigma}} = \widehat{\overline{\sigma}}$ if $\widetilde{\Pi}_k = \widehat{\Pi}_k$, and by $\widetilde{\sigma} = (h^2+1)\overline{\sigma}$ and $\widehat{\widetilde{\sigma}} = (h^2+1)\widehat{\overline{\sigma}}$ if $\widetilde{\Pi}_k = \widehat{\Pi}_k^t$. Because, for $K \in \mathbb{N}$ and $A_1, \dots, A_K$ with same number of rows, $|(A_1, \dots, A_K)|_{\text{op}}^2 \le \sum_{k=1}^K |A_k|_{\text{op}}^2$, and

$$\widehat{\Gamma} - \Gamma^l = \widehat{\widetilde{\Gamma}} - \widetilde{\Gamma}^l + \sum_{k=1}^{K} \left(\beta_k - \widehat{\beta}_k\right)\left(\widetilde{\Pi}_k - \Pi_k\right) + \sum_{k=1}^{K} \left(\beta_k - \widehat{\beta}_k\right)\Pi_k,$$

we obtain the following corollary of Proposition 8 and (22).

Corollary 1. *Under the assumptions 1, 3, where in* (iii) *we have $\widetilde{\Gamma}^l$ instead of Γ^l, 4, 7, $\lambda_N^2 \widetilde{r}_N = o(NT)$, and $\left|\Gamma^d\right|_{\text{op}} = o_P(\lambda_N \sigma)$, we have*

$$\left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}} \le (\rho+1)\lambda_N\left(\sigma + o_P(1)\right)$$

$$\max\left(\left|\Pi_u^l - \widehat{\Pi}_u^l\right|_{\text{op}}, \left|\Pi_v^l - \widehat{\Pi}_v^l\right|_{\text{op}}\right) \le (\rho+1)\lambda_N\left(\widetilde{\sigma} + o_P(1)\right).$$

Based on this corollary, we can rely on hard-thresholding of these estimators that we denote by $\widehat{\Gamma}^t$, $\widehat{\Pi}_u^t$ and $\widehat{\Pi}_v^t$ and estimate the rank of Γ^l and the annihilator matrices $M_{u(\Gamma^l)}$, $M_{v(\Gamma^l)}$, $M_{u(\Pi_u^l)}$, and $M_{v(\Pi_v^l)}$ by $M_{u\left(\widehat{\Gamma}^t\right)}$, $M_{v\left(\widehat{\Gamma}^t\right)}$, $M_{u\left(\widehat{\Pi}_u^t\right)}$, and $M_{v\left(\widehat{\Pi}_v^t\right)}$. Again, the first two annihilators are estimated at the same rate as in Lemma A.7 in [3] if Γ^l satisfies a strong factor assumption. Proposition 9 and Proposition 10 hold with the annihilator matrices of this section replacing σ by $\widetilde{\sigma}$ and $\widehat{\sigma}$ by $\widehat{\widetilde{\sigma}}$ and Assumption 5 (ii) by

$$\mathbb{P}\left(\min\left(\sigma_{\mathrm{rank}\left(\Pi_u^l\right)}\left(\Pi_u^l\right),\sigma_{\mathrm{rank}\left(\Pi_v^l\right)}\left(\Pi_v^l\right)\right)\geq(\rho+1)\lambda_N h^2(h+1)\widetilde{\sigma}\right)\to 1$$

and Assumption 6 by $\lambda_N^2\widetilde{r}_N=o(NT)$ maintained in Corollary 1 and the next assumption.

Assumption 8. *Let $\{\overline{v}_N\}$ be such that $\overline{v}_N\geq(\rho+1)\lambda_N h^2(h+1)\widetilde{\sigma}$, we have*

$$\mathbb{P}\left(\min\left(\sigma_{\mathrm{rank}\left(\Pi_u^l\right)}\left(\Pi_u^l\right),\sigma_{\mathrm{rank}\left(\Pi_v^l\right)}\left(\Pi_v^l\right)\right)\geq\overline{v}_N\right)\to 1$$

and, for a sequence $\{\overline{r}_N\}$, $\max\left(\mathrm{rank}\left(\Pi_u^l\right),\mathrm{rank}\left(\Pi_v^l\right)\right)=O_P(\overline{r}_N)$.

We denote by $\mathcal{P}_{\widehat{\Pi}^t}^{\perp}$ (resp. $\mathcal{P}_{\Pi}^{\perp}$) the operator which applied to $A\in\mathcal{M}_{NT}$ is $\mathcal{P}_{\widehat{\Pi}}^{\perp}(A)=M_{u\left(\widehat{\Pi}_u^t\right)}AM_{v\left(\widehat{\Pi}_v^t\right)}$ (resp. $\mathcal{P}_{\Pi}^{\perp}(A)=M_{u(\Pi_u)}AM_{v(\Pi_v)}$) and define the estimator

$$\widetilde{\beta}^{(1)}\in\underset{\beta\in\mathbb{R}^K}{\mathrm{argmin}}\left|\mathcal{P}_{\widehat{\Pi}^t}^{\perp}\left(Y-\sum_{k=1}^K\beta_kX_k\right)\right|_2^2. \tag{25}$$

Also $\mathcal{P}_{\widehat{\Pi}^t}^{\perp}(X)$ (resp. $\mathcal{P}_{\widehat{\Pi}^t}^{\perp}(U)$, $\mathcal{P}_{\Pi}^{\perp}(X)$, and $\mathcal{P}_{\Pi}^{\perp}(U)$) is the matrix formed like X, replacing the matrices X_k by $\mathcal{P}_{\widehat{\Pi}^t}^{\perp}(X_k)$ (resp. $\mathcal{P}_{\Pi}^{\perp}(X_k)$, $\mathcal{P}_{\Pi}^{\perp}(U_k)$, and $\mathcal{P}_{\Pi}^{\perp}(U_k)$) for $k=1,\dots,K$.

Assumption 9. *Maintain the assumptions of Corollary 1 and Assumption 8 and*

(i) $\overline{r}_N\lambda_N^2\left(\lambda_N+\sqrt{\overline{r}_N}\mu_N^2/\overline{v}_N\right)/\overline{v}_N=o\left(NT\right)$,
(ii) $\overline{r}_N\lambda_N^3/\overline{v}_N=o\left(\sqrt{NT}\right)$,
(iii) $\overline{r}_N^{3/2}\lambda_N^3\left(|\Gamma|_{\mathrm{op}}+\lambda_N\right)/\overline{v}_N^2=o_P\left(\sqrt{NT}\right)$,
(iv) $\left|\mathcal{P}_{\Pi^l}^{\perp}(\Pi^d)\right|_2^2=o_P(NT)$,
(v) *There exists $\Sigma_{\perp}\in\mathcal{M}_{KK}$ positive definite such that $\mathcal{P}_{\Pi^l}^{\perp}(U)^{\top}\mathcal{P}_{\Pi^l}^{\perp}(U)/(NT)\xrightarrow{\mathbb{P}}\Sigma_{\perp}$,*
(vi) $\mathcal{P}_{\Pi^l}^{\perp}(U)^{\top}e/\sqrt{NT}\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\Sigma_{\perp}\right)$.

Regarding Assumption 9 (iii), $|\Gamma|_{\mathrm{op}}$ is usually $O_P\left(\sqrt{NT}\right)$ if it has a nontrivial low-rank component. (i)-(iii) can be satisfied under weaker assumptions than a strong factor assumption ($\overline{v}_N$ is of the order of $\sqrt{NT}$) and when $\overline{r}_N$ goes to infinity. (v) is satisfied, for example, if (Π_u^l,Π_v^l) and U are independent and (vi) when (X,Γ^l) and E are independent.

Theorem 3. *Let Assumption 9 holds. We have*

$$\frac{\sqrt{NT}}{\widehat{\sigma}}\left(\widetilde{\beta}^{(1)}-\beta\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_{\perp}^{-1}\right),$$
$$\mathcal{P}_{\widehat{\Pi}^t}^{\perp}(X)^{\top}\mathcal{P}_{\widehat{\Pi}^t}^{\perp}(X)/(NT)\xrightarrow{\mathbb{P}}\Sigma_{\perp}.$$

Also, if $|\mathcal{P}_{\Pi^l}(U)|_2^2=o_P(|U|_2^2)$ then $\Sigma_{\perp}=\mathbb{E}[U^{\top}U]$. This occurs if $\mathbb{E}\left[\max\left(\mathrm{rank}\left(\Pi_u^l\right),\mathrm{rank}\left(\Pi_v^l\right)\right)\right]=o\left(\sqrt{\min(N,T)}\right)$ and U and (Π_u^l,Π_v^l) are independent.

4.7.2. *Approach 2: Using* [3]*'s estimator as a second stage.* An alternative approach put forward by [21] is to rely on a preliminary estimator like their matrix Lasso as a first-step to initialize [3]'s non convex estimator. Among other conditions, using such a two-stage approach requires that the rate of convergence of the first-step estimator of β is at least $(NT)^{1/6}$, a consistent estimator of rank(Γ), which is assumed constant, a strong-factor assumption on Γ, and $\Gamma^d = 0$. This methodology can be applied using as a first-stage the thresholded or nonthresholded square-root estimator of this paper. We denote this estimator by $\left(\widetilde{\beta}^{(2)}, \widetilde{\Gamma}^{(2)}\right)$. This paper provides a consistent estimator of rank $\left(\Gamma^l\right)$ via hard-thresholding of (24) or an upper bound on it without thresholding. Lemma 3 in [21] proposes an other consistent estimator but probably has a typo due to contradictory assumptions. The advantage of the estimator of this paper is that the level of thresholding is less conservative and makes use of the consistent estimator of the variance of errors. Recall that if $\Gamma^d = 0$ and $\Pi_1^l = \dots, \Pi_K^l$, from the discussion after Proposition 7 and (24),

$$\operatorname{rank}\left(\widehat{\Gamma}\right) \le 2\left(\frac{1+\rho}{1-\rho}\right)^2\left(\frac{\widetilde{r}_N}{\kappa^2_{\widetilde{\Gamma}^l}} + \sum_{k=1}^K r_{kN}\right) + o_P(1).$$

An estimator of the asymptotic covariance matrix of the second-stage estimator, given a consistent estimator of $\widehat{r} = \operatorname{rank}\left(\Gamma^l\right)$, is given by (see page 1552 of [19]) $\widehat{\sigma}_B\widehat{\Sigma}_B$, where

$$\widehat{\sigma}_B = \frac{1}{\sqrt{(N-\widehat{r})(T-\widehat{r})-K}}\left|Y - \sum_{k=1}^K \widetilde{\beta}_k^{(2)} X_k - \widetilde{\Gamma}^{(2)}\right|_2$$

$$\left(\widehat{\Sigma}_B\right)_{kl} = \frac{1}{NT}\left\langle M_{u\left(\widetilde{\Gamma}^{(2)}\right)} X_k M_{v\left(\widetilde{\Gamma}^{(2)}\right)}, X_l\right\rangle \quad \forall k,l \in \{1,\dots,K\}^2.$$

5. Simulations

We take the same data generating process as in [21] with a single regressor and two factors:

$$Y_{it} = X_{1it} + \sum_{l=1}^2 \left(1 + \lambda_{0,il}\right) f_{0,tl} + E_{it},$$

$$X_{1it} = 1 + \sum_{l=1}^2 (2 + \lambda_{0,il} + \lambda_{1,il})(f_{0,tl} + f_{0,t-1\ r}) + U_{it},$$

where $f_{0,tl}$, $\lambda_{0,il}$, $\lambda_{1,il}$, U_{it}, and E_{it} for all indices are mutually independent and i.i.d. standard normal. The matrix X_1 has an approximate factor structure with a low-rank component of rank 3 due to the constant 1. The least-squares estimator $\widehat{\beta}^{LS}$ which ignores the presence of Γ is inconsistent because X_{it} and Γ_{it} are correlated. By the analysis of the paper, the square-root estimator coincides with the estimator in [21] with a smaller penalization. The results in [21] are obtained with a penalty much smaller than allowed by the theory. We compare the performance of the least-squares estimator $\widehat{\beta}^{LS}$, the square-root estimator $\widehat{\beta}$ obtained with the baseline regressors, the square-root estimator $\widehat{\beta}_{pt}$ obtained with the transformed regressors, where we apply (2) from Section 4.6 with $\widetilde{\Pi}_1 = \widehat{\Pi}_1^t$, and the two-stage estimators from Section

4.7. We use $\widehat{\beta}^{LS}$ to initialize the iterative estimators. The number of iterations is 100 to obtain the estimator of rank (Γ), as explained after Corollary 1, useful to compute $\widetilde{\beta}^{(2)}$. We use the same number of iterations to obtain $\widehat{\beta}_{pt}$. We consider an additional 100 iterations for $\widehat{\beta}$, $\widehat{\beta}_{pt}$, and $\widetilde{\beta}^{(2)}$. As a result, $\widetilde{\beta}^{(1)}$ and $\widetilde{\beta}^{(2)}$ have been computed with the same number of iterations. We consider two sample sizes: (a) $N = T = 50$ and (b) $N = T = 150$. We use 7300 Monte-Carlo replications to allow for an accuracy of ± 0.005 with 95% for the coverage probabilities of 95% confidence intervals. We choose $\lambda_N = 1.01\left(\sqrt{N} + \sqrt{T}\right)$ and the hard-thresholding levels are $2\lambda_N$ times an estimator of the standard error from the first-stages.

A first approach is to not apply any matrix to the data as described after Proposition 4. The results in tables 1 and 2 compare the performance of the estimators in terms of Mean Squared Error (henceforth MSE), bias, and standard error (henceforth std). In case (a), rank $\left(\widehat{\Pi}_1^t\right)$ has been found to be always equal to 2 while rank $\left(\widehat{\Pi}_1\right)$ to 3 (the true rank), rank $\left(\widehat{\Gamma}^t\right)$ has been found to be always equal to 2 (the true rank) in 89% of the cases and else to 1. We used rank $\left(\widehat{\Pi}_1^t\right)$ for $\widehat{\beta}_{pt}$ and subsequently rank $\left(\widehat{\Gamma}^t\right)$, $\widetilde{\beta}^{(1)}$ and $\widetilde{\beta}^{(2)}$, even though it did not perform well for such small sample size. In case (b), rank $\left(\widehat{\Pi}_1^t\right)$ has been found to be always equal to 3 while rank $\left(\widehat{\Pi}_1\right)$ and rank $\left(\widehat{\Gamma}^t\right)$ have been found to be always equal to 2 (the true rank).

TABLE 1. $N = T = 50$

	$\widehat{\beta}^{LS}$	$\widehat{\beta}$	$\widehat{\beta}_{pt}$	$\widetilde{\beta}^{(1)}$	$\widetilde{\beta}^{(2)}$
MSE	0.053	0.020	$5\ 10^{-4}$	0.002	$9\ 10^{-4}$
bias	0.230	0.142	-10^{-4}	0.019	0.009
std	0.017	0.015	0.023	0.035	0.029

TABLE 2. $N = T = 150$

	$\widehat{\beta}^{LS}$	$\widehat{\beta}$	$\widehat{\beta}_{pt}$	$\widetilde{\beta}^{(1)}$	$\widetilde{\beta}^{(2)}$
MSE	0.053	0.011	$4\ 10^{-5}$	$4\ 10^{-5}$	$1\ 10^{-5}$
bias	0.231	0.103	$4\ 10^{-4}$	$2\ 10^{-5}$	$-8\ 10^{-5}$
std	0.009	0.008	0.006	0.006	0.003

A second approach is to apply Within transforms $M_u = I_N - J_N/N$ and $M_v = I_T - J_T/T$ to the left and right of Y and X_1, where $J_N \in \mathcal{M}_{NN}$ (resp. $J_T \in \mathcal{M}_{TT}$) has all entries equal to 1. These allow to get rid of the mean 1 of X_1 but more generally of any individual and time effects in both Π^l and Γ^l. The results are in tables 3 and 4. In case (a), rank $\left(\widehat{\Pi}_1^t\right)$ and rank $\left(\widehat{\Pi}_1\right)$ has been found to be always equal to 2 (the true rank), rank $\left(\widehat{\Gamma}^t\right)$ has been found to be equal to 2 (the true rank) in 81% of the cases and else to 1. In case (b), rank $\left(\widehat{\Pi}_1^t\right)$, rank $\left(\widehat{\Pi}_1\right)$, rank $\left(\widehat{\Gamma}^t\right)$ have been found to be always equal to 2 (the true ranks).

Table 5 assesses the coverage probabilities in the different cases.

REFERENCES

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

[2] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khashayar. Matrix completion methods for causal panel data models. arXiv preprint 1710.10251, 2017.

Table 3. $N = T = 50$, Within

	$\widehat{\beta}^{LS}$	$\widehat{\beta}$	$\widehat{\beta}_{pt}$	$\widetilde{\beta}^{(1)}$	$\widetilde{\beta}^{(2)}$
MSE	0.049	0.016	$5\ 10^{-4}$	0.001	0.002
bias	0.221	0.124	$-4\ 10^{-5}$	0.024	0.020
std	0.025	0.018	0.023	0.025	0.044

Table 4. $N = T = 150$, Within

	$\widehat{\beta}^{LS}$	$\widehat{\beta}$	$\widehat{\beta}_{pt}$	$\widetilde{\beta}^{(1)}$	$\widetilde{\beta}^{(2)}$
MSE	0.049	0.007	$5\ 10^{-5}$	$5\ 10^{-5}$	$2\ 10^{-5}$
bias	0.222	0.081	$-1\ 10^{-4}$	$4\ 10^{-4}$	$7\ 10^{-7}$
std	0.014	0.007	0.007	0.007	0.004

Table 5. Coverage of 95% confidence intervals based on the two-stage approaches.

Within transforms	(N,T)	$\widetilde{\beta}^{(1)}$	$\widetilde{\beta}^{(2)}$
No	(50,50)	0.87	0.84
Yes	(50,50)	0.81	0.76
No	(150,150)	0.95	0.94
Yes	(150,150)	0.95	0.94

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

[4] J. Bai and S. Ng. Principal components and regularized estimation of factor models. arXiv preprint 1708.08137, 2017.

[5] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[6] P. Bühlmann and S. van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

[7] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[8] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[9] A. Chudik and M. H. Pesaran. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, 188(2):393–420, 2015.

[10] A. Dupuy, A. Galichon, and Y. Sun. Estimating matching affinity matrix under low-rank constraints. arXiv preprint 1612.09585, 2017.

[11] E. Gautier, A. Tsybakov, and C. Rose. High-dimensional instrumental variables regression and confidence sets. arXiv preprint 1105.2454v5.

[12] C. Giraud. *Introduction to High-Dimensional Statistics.* Chapman and Hall/CRC, 2014.

[13] L. Gobillon and T. Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98:535–551, 2016.

[14] C. Hansen and Y. Liao. The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, pages 1–45, 2018.

[15] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[16] O. Klopp and S. Gaiffas. High dimensional matrix estimation with unknown variance of the noise. *Statistica Sinica*, 27:115–145, 2017.

[17] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

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

[19] H. R. Moon and M. Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579, 2015.

[20] H. R. Moon and M. Weidner. Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195, 2017.

[21] H. R. Moon and M. Weidner. Nuclear norm regularized estimation of panel regression models. arXiv preprint 1810.10987, 2018.

[22] A. Onatski. Asymptotic analysis of the squared estimation error in misspecified factor models. *Journal of Econometrics*, 186(2):388–406, 2015.

[23] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

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

[25] M. H. Pesaran. *Time Series and Panel Data Econometrics*. Oxford University Press, 2015.

[26] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

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

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

[29] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[30] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*, pages 210–268, 2012.

Appendix

Recall that, for all $A, M, N \in \mathcal{M}_{NT}$ (see lemma 2.3 and 3.4 in [27] for the last two),

$$\mathcal{P}_A(M) = M_{u(A)} M P_{v(A)} + P_{u(A)} M, \tag{26}$$

$$\operatorname{rank}\left(\mathcal{P}_A(M)\right) \leq 2 \min\left(\operatorname{rank}\left(M\right), \operatorname{rank}(A)\right), \tag{27}$$

$$\left\langle \mathcal{P}_A(M), \mathcal{P}_A(N) \right\rangle = \left\langle \mathcal{P}_A(M), N \right\rangle, \tag{28}$$

$$\left\langle \mathcal{P}_A(M), \mathcal{P}_A^{\perp}(M) \right\rangle = 0, \tag{29}$$

$$\left| A + \mathcal{P}_A^{\perp}(M) \right|_* = |A|_* + \left| \mathcal{P}_A^{\perp}(M) \right|_*. \tag{30}$$

Proof of Proposition 1. By definition of $\widehat{\beta}$ and $\widehat{\Gamma}$, we have, for all $\beta \in \mathbb{R}^K$ and $\Gamma \in \mathcal{M}_{NT}$,

$$\frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k - \widehat{\Gamma} \right|_2 + \frac{\lambda}{NT} \left| \widehat{\Gamma} \right|_* \leq \frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma \right|_2 + \frac{\lambda}{NT} |\Gamma|_*.$$

By definition of P_X and of the estimator, for all $\beta \in \mathbb{R}^K$ and $\Gamma \in \mathcal{M}_{NT}$, we have

$$\frac{1}{\sqrt{NT}} \left| M_X \left(Y - \widehat{\Gamma} \right) \right|_2 + \frac{\lambda}{NT} \left| \widehat{\Gamma} \right|_* \leq \frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^{K} \widehat{\beta}_k X_k - \widehat{\Gamma} \right|_2 + \frac{\lambda}{NT} \left| \widehat{\Gamma} \right|_*$$

$$\leq \frac{1}{\sqrt{NT}} \left| Y - \sum_{k=1}^{K} \beta_k X_k - \Gamma \right|_2 + \frac{\lambda}{NT} |\Gamma|_*.$$

By choosing β such that $\sum_{k=1}^{K} \beta_k X_k = P_X (Y - \Gamma)$, we obtain, for all $\Gamma \in \mathcal{M}_{NT}$,

$$\frac{1}{\sqrt{NT}} \left| M_X \left(Y - \widehat{\Gamma} \right) \right|_2 + \frac{\lambda}{NT} \left| \widehat{\Gamma} \right|_* \leq \frac{1}{\sqrt{NT}} \left| M_X (Y - \Gamma) \right|_2 + \frac{\lambda}{NT} |\Gamma|_*,$$

hence the result.

Proof of Proposition2. The first inequality is obtained using trace duality and (27). The second is obtained by (29) and the Pythagorean theorem.

Proof of Theorem 1. The techniques are similar to those in [5, 11]. Take $\widetilde{\Gamma} \in \mathcal{M}_{NT}$ and denote by $\Delta = \widehat{\Gamma} - \Gamma$. Remark that

$$\left| \widehat{\Gamma} \right|_* = \left| \Gamma - \widetilde{\Gamma} + \widetilde{\Gamma} + \mathcal{P}_{\widetilde{\Gamma}}(\Delta) + \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_*$$

$$\geq \left| \widetilde{\Gamma} + \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* - \left| \Gamma - \widetilde{\Gamma} \right|_* - \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_* \tag{31}$$

$$\geq \left| \widetilde{\Gamma} \right|_* + \left| \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* - \left| \Gamma - \widetilde{\Gamma} \right|_* - \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_* \quad \text{(by (30))}. \tag{32}$$

Now, by (9) and the definition of $\widehat{\Gamma}$, we have

$$\frac{1}{\sqrt{NT}} \left| M_X \left(Y - \widehat{\Gamma} \right) \right|_2 + \frac{\lambda}{NT} \left| \widehat{\Gamma} \right|_* \leq \frac{1}{\sqrt{NT}} \left| M_X (Y - \Gamma) \right|_2 + \frac{\lambda}{NT} \left| \Gamma \right|_* . \tag{33}$$

By convexity, trace duality, and $\lambda\rho |M_X(E)|_2 / \sqrt{NT} \geq |M_X(E)|_{\mathrm{op}}$, if $M_X(E) \neq 0$, we have

$$\frac{1}{\sqrt{NT}} \left| M_X \left(Y - \widehat{\Gamma} \right) \right|_2 - \frac{1}{\sqrt{NT}} \left| M_X (Y - \Gamma) \right|_2 \geq -\frac{1}{\sqrt{NT} |M_X(E)|_2} \left\langle M_X(E), \widehat{\Gamma} - \Gamma \right\rangle$$

$$\geq -\frac{\lambda\rho}{NT} |\Delta|_*. \tag{34}$$

(34) also holds if $M_X(E) = 0$ because $\left| M_X \left(Y - \widehat{\Gamma} \right) \right|_2 \geq 0$. This and (33) imply

$$\left| \widehat{\Gamma} \right|_* \leq \rho |\Delta|_* + |\Gamma|_*. \tag{35}$$

Using (32), we get

$$\left| \widetilde{\Gamma} \right|_* + \left| \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* - \left| \Gamma - \widetilde{\Gamma} \right|_* - \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_* \leq \rho \left| \Delta \right|_* + \left| \Gamma \right|_*$$

and $|\Gamma|_* \leq \left| \Gamma - \widetilde{\Gamma} \right|_* + \left| \widetilde{\Gamma} \right|_*$ yields

$$\left| \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* - \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_* \leq \rho \left| \Delta \right|_* + 2 \left| \Gamma - \widetilde{\Gamma} \right|_* .$$

Then, because $|\Delta|_* \leq \left| \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* + \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_*$, we have

$$(1 - \rho) \left| \mathcal{P}_{\widetilde{\Gamma}}^{\perp}(\Delta) \right|_* \leq (1 + \rho) \left| \mathcal{P}_{\widetilde{\Gamma}}(\Delta) \right|_* + 2 \left| \Gamma - \widetilde{\Gamma} \right|_* . \tag{36}$$

Also, by (33),

$$\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\leq\frac{\lambda}{NT}\left(|\Gamma|_*-\left|\widehat{\Gamma}\right|_*\right)$$

and

$$\begin{aligned}|\Gamma|_*-\left|\widehat{\Gamma}\right|_*&\leq\left|\widetilde{\Gamma}\right|_*+\left|\Gamma-\widetilde{\Gamma}\right|_*-\left|\widehat{\Gamma}\right|_*\\&=2\left|\Gamma-\widetilde{\Gamma}\right|_*+\left|\widetilde{\Gamma}\right|_*-\left|\Gamma-\widetilde{\Gamma}\right|_*-\left|\widehat{\Gamma}\right|_*\\&\leq2\left|\Gamma-\widetilde{\Gamma}\right|_*+\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*-\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}(\Delta)\right|_*\quad\text{(by (32))},\end{aligned}$$

hence we have

$$\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\leq\frac{\lambda}{NT}\left(2\left|\Gamma-\widetilde{\Gamma}\right|_*+\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*\right).\tag{37}$$

Let $\widetilde{\rho}>0$ and consider two cases.
Case 1. If $\widetilde{\rho}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*\leq2\left|\Gamma-\widetilde{\Gamma}\right|_*$, we have, by (36),

$$\begin{aligned}|\Delta|_*&\leq\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*+\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}(\Delta)\right|_*\\&\leq\frac{2}{1-\rho}\left(\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*+\left|\Gamma-\widetilde{\Gamma}\right|_*\right)\\&\leq\frac{2}{1-\rho}\left(\frac{2}{\widetilde{\rho}}+1\right)\left|\Gamma-\widetilde{\Gamma}\right|_*.\end{aligned}$$

This yields the first part of the first inequality of Theorem 1. The first part of the second inequality is obtained by combining (34) and (37).
Case 2. Otherwise, if $\widetilde{\rho}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*>2\left|\Gamma-\widetilde{\Gamma}\right|_*$, we obtain, by (36), that

$$\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}(\Delta)\right|_*\leq c\left(\rho,\widetilde{\rho}\right)\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*,$$

which implies that $\Delta\in C_{\widetilde{\Gamma}}$ and $|\Delta|_*\leq(1+c(\rho,\widetilde{\rho}))\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*$. We have

$$\frac{1}{NT}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2^2-\frac{1}{NT}|M_X\left(Y-\Gamma\right)|_2^2=\frac{1}{NT}\left|M_X\left(\widehat{\Gamma}-\Gamma\right)\right|_2^2-\frac{2}{NT}\left\langle M_X(E),\widehat{\Gamma}-\Gamma\right\rangle$$

hence, because $\lambda\rho|M_X(E)|_2/\sqrt{NT}\geq|M_X(E)|_{\mathrm{op}}$,

$$\begin{aligned}&\frac{1}{NT}\left|M_X\left(\widehat{\Gamma}-\Gamma\right)\right|_2^2\\&\leq\frac{1}{NT}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2^2-\frac{1}{NT}|M_X\left(Y-\Gamma\right)|_2^2+2\lambda\rho\left(1+c\left(\rho,\widetilde{\rho}\right)\right)\frac{|M_X(E)|_2}{(NT)^{\frac{3}{2}}}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*\end{aligned}\tag{38}$$

and, by (37),

$$\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\leq\frac{(1+\widetilde{\rho})\lambda}{NT}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*$$

which, combined with (34), yields

$$\left|\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\right|\leq d\left(\rho,\widetilde{\rho}\right)\frac{\lambda}{NT}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*. \tag{39}$$

Now, using

$$\begin{aligned}&\frac{1}{NT}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2^2-\frac{1}{NT}\left|M_X\left(Y-\Gamma\right)\right|_2^2\\&=\left(\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\right)\\&\quad\times\left(\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2-\frac{1}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2+\frac{2}{\sqrt{NT}}\left|M_X\left(Y-\Gamma\right)\right|_2\right)\end{aligned}$$

and (39), we obtain

$$\begin{aligned}&\frac{1}{NT}\left|M_X\left(Y-\widehat{\Gamma}\right)\right|_2^2-\frac{1}{NT}\left|M_X\left(Y-\Gamma\right)\right|_2^2\\&\leq d\left(\rho,\widetilde{\rho}\right)\frac{\lambda}{NT}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*\left(d\left(\rho,\widetilde{\rho}\right)\frac{\lambda}{NT}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*+\frac{2\left|M_X\left(E\right)\right|_2}{\sqrt{NT}}\right).\end{aligned} \tag{40}$$

Combining (38) and (40), we get

$$\frac{1}{NT}\left|M_X\left(\widehat{\Gamma}-\Gamma\right)\right|_2^2\leq\left(d\left(\rho,\widetilde{\rho}\right)\frac{\lambda}{NT}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*\right)^2+2e\left(\rho,\widetilde{\rho}\right)\frac{\lambda\left|M_X\left(E\right)\right|_2}{(NT)^{\frac{3}{2}}}\left|\mathcal{P}_{\widetilde{\Gamma}}(\Delta)\right|_*.$$

By definition of $\kappa_{\widetilde{\Gamma},c\left(\rho,\widetilde{\rho}\right)}$, this implies

$$\left|M_X\left(\Delta\right)\right|_2\leq 2\left(1-\left(d\left(\rho,\widetilde{\rho}\right)\frac{\sqrt{2\mathrm{rank}\left(\widetilde{\Gamma}\right)}\lambda}{\sqrt{NT}\kappa_{\widetilde{\Gamma},c\left(\rho,\widetilde{\rho}\right)}}\right)^2\right)_+^{-1}e\left(\rho,\widetilde{\rho}\right)\frac{\lambda\sqrt{2\mathrm{rank}\left(\widetilde{\Gamma}\right)}\left|M_X(E)\right|_2}{\sqrt{NT}\kappa_{\widetilde{\Gamma},c\left(\rho,\widetilde{\rho}\right)}},$$

$$\left|\mathcal{P}_\Gamma(\Delta)\right|_*\leq 4\left(1-\left(d\left(\rho,\widetilde{\rho}\right)\frac{\sqrt{2\mathrm{rank}\left(\widetilde{\Gamma}\right)}\lambda}{\sqrt{NT}\kappa_{\widetilde{\Gamma},c\left(\rho,\widetilde{\rho}\right)}}\right)^2\right)_+^{-1}e\left(\rho,\widetilde{\rho}\right)\frac{\lambda\mathrm{rank}\left(\widetilde{\Gamma}\right)\left|M_X(E)\right|_2}{\sqrt{NT}\kappa^2_{\widetilde{\Gamma},c\left(\rho,\widetilde{\rho}\right)}}, \tag{41}$$

which yields the first result. The second result follows from (39) and (41).

Proof of Proposition 3.

Lemma 1. *It holds that* $\left|P_X(E)\right|_2=O_P(1)$ *and* $\left|P_X(E)\right|_{\mathrm{op}}=O_P\left(\mu_N/\sqrt{NT}\right)$.

Proof. Let $|\cdot|$ denote the ℓ_2 or operator norm. We use that, due to Assumption 1 (ii), w.p.a. 1, $\left|P_X(E)\right|=\left|X(X^\top X)^{-1}X^\top e\right|$ and

$$\left|X(X^\top X)^{-1}X^\top e\right|=\left|\sum_{k=1}^K X_k\left((X^\top X)^{-1}X^\top e\right)_k\right|\leq\sqrt{\sum_{k=1}^K\left|X_k\right|^2}\left|(X^\top X)^{-1}X^\top e\right|_2.$$

Due to Assumption 1 (ii) and (iii), we have

$$\left|(X^\top X)^{-1}X^\top e\right|_2 \le \left|\left(\frac{X^\top X}{NT}\right)^{-1}\right|_{\mathrm{op}} \left|\frac{X^\top e}{NT}\right|_2 = O_P\left(\frac{1}{\sqrt{NT}}\right) \tag{42}$$

and $|X_k|_2 = \sqrt{(X^\top X)_{kk}} = O_P\left(\sqrt{NT}\right)$ hence the result. □

By Lemma 1 and the inverse triangle inequality, we have

$$\left|\frac{|M_X(E)|_2}{\sqrt{NT}} - \frac{|E|_2}{\sqrt{NT}}\right| \le \frac{|P_X(E)|_2}{\sqrt{NT}} \xrightarrow{\mathbb{P}} 0$$

and we conclude by Assumption 1 (i). For the operator norm, we use Assumption 1 (iv) and

$$\left||M_X(E)|_{\mathrm{op}} - |E|_{\mathrm{op}}\right| \le |P_X(E)|_{\mathrm{op}}.$$

Proof of Proposition 5. Let us consider a cone with constant c. We work for any draw of X and Γ^l and consider the matrices fixed. By the computations in the proof of Lemma 1,

$$|P_X(\Delta)|_2 \le \frac{|X|_2}{NT}\left|\left(\frac{X^\top X}{NT}\right)^{-1}\right|_{\mathrm{op}} \left|X^\top \delta\right|_2.$$

Also, for $k \in \{1,\dots,K\}$, using the cone and the trace duality in the third display, we obtain

$$\begin{aligned}
|\langle X_k, \Delta\rangle| &\le |\langle X_k, \mathcal{P}_{\Gamma^l}(\Delta)\rangle| + \left|\left\langle X_k, \mathcal{P}_{\Gamma^l}^{\perp}(\Delta)\right\rangle\right| \\
&= |\langle \mathcal{P}_{\Gamma^l}(X_k), \mathcal{P}_{\Gamma^l}(\Delta)\rangle| + \left|\left\langle \mathcal{P}_{\Gamma^l}^{\perp}(X_k), \mathcal{P}_{\Gamma^l}^{\perp}(\Delta)\right\rangle\right| \\
&\le \min\left(|\mathcal{P}_{\Gamma^l}(X_k)|_{\mathrm{op}}, |X_k|_{\mathrm{op}}\right)|\mathcal{P}_{\Gamma^l}(\Delta)|_* + \left|\mathcal{P}_{\Gamma^l}^{\perp}(X_k)\right|_{\mathrm{op}}\left|\mathcal{P}_{\Gamma^l}^{\perp}(\Delta)\right|_*,
\end{aligned}$$

hence

$$|P_X(\Delta)|_2^2 \le \sum_{k=1}^K \left(b_k |\mathcal{P}_{\Gamma^l}(\Delta)|_* + b_{\perp k}\left|\mathcal{P}_{\Gamma^l}^{\perp}(\Delta)\right|_*\right)^2.$$

Also, by homogeneity, we have

$$\kappa_{\Gamma^l,c}^2 = 2\mathrm{rank}\left(\Gamma^l\right) \inf_{\Delta\in C_{\Gamma^l}:\ \left|\mathcal{P}_{\Gamma^l}(\Delta)\right|_*=1} \left(|\Delta|^2 - |P_X(\Delta)|_2^2\right).$$

Denote by $\{\sigma_k\}$ and $\{\sigma_{\perp k}\}$ the singular values of $\mathcal{P}_{\Gamma^l}(\Delta)$ and $\mathcal{P}_{\Gamma^l}^{\perp}(\Delta)$. The rank of the first (resp. the second) matrix is at most $2\mathrm{rank}\left(\Gamma^l\right)$ (resp. p_N) so, by the Pythagorean theorem,

$$\kappa_{\Gamma^l,c}^2 \ge 2\mathrm{rank}\left(\Gamma^l\right) \inf_{\substack{\sum_k \sigma_k = 1 \\ |\sigma|_0 \le 2\mathrm{rank}(\Gamma^l) \\ \sum_k \sigma_{\perp k}\le c \\ |\sigma_\perp|_0 \le p_N \\ \sigma\ge 0, \sigma_\perp \ge 0}} \left(\sum_k \sigma_k^2 + \sum_k \sigma_{\perp k}^2 - \sum_{k=1}^K \left(b_k + b_{\perp k}\left(\sum_k \sigma_{\perp k}\right)\right)^2\right)$$

$$= 1 + 2\mathrm{rank}\left(\Gamma^l\right) \inf_{\substack{\sum_k \sigma_{\perp k} \le c \\ |\sigma_\perp|_0 \le p_N \\ \sigma_\perp \ge 0}} \left(\sum_k \sigma_{\perp k}^2 - \sum_{k=1}^K \left(b_k + b_{\perp k} \left(\sum_k \sigma_{\perp k} \right) \right)^2 \right) \tag{43}$$

$$= 1 + 2\mathrm{rank}\left(\Gamma^l\right) \inf_{0 \le u \le c} \inf_{\substack{\sum_k \sigma_{\perp k} = u \\ |\sigma_\perp|_0 \le p_N \\ \sigma_\perp \ge 0}} \left(\sum_k \sigma_{\perp k}^2 - \sum_{k=1}^K (b_k + b_{\perp k} u)^2 \right)$$

$$= 1 + 2\mathrm{rank}\left(\Gamma^l\right) \min_{0 \le u \le c} \left(\frac{u^2}{p_N} - \sum_{k=1}^K (b_k + b_{\perp k} u)^2 \right). \tag{44}$$

The degree 2 polynomial in the bracket has a minimum at u_* given by $u_* \left(1 - p_N |b_\perp|_2^2\right) = p_N \langle b_\perp, b \rangle$. If $p_N |b_\perp|_2^2 \ge 1$ then the minimum is at 0 in which case $\kappa^2_{\Gamma^l,c} \ge 1 - 2\mathrm{rank}\left(\Gamma^l\right) |b|_2^2$, else, if $p_N \langle b_\perp, b \rangle < c\left(1 - p_N |b_\perp|_2^2\right)$ the minimum is at u_* and

$$\kappa^2_{\Gamma^l,c} \ge 1 - 2\mathrm{rank}\left(\Gamma^l\right) \left(\left| b + b_\perp \frac{p_N \langle b_\perp, b \rangle}{1 - p_N |b_\perp|_2^2} \right|_2^2 - \frac{p_N \langle b_\perp, b \rangle^2}{\left(1 - p_N |b_\perp|_2^2\right)^2} \right),$$

else, the minimum is at c and

$$\kappa^2_{\Gamma^l,c} \ge 1 - 2\mathrm{rank}\left(\Gamma^l\right) \left(|b + b_\perp c|_2^2 - \frac{c^2}{p_N} \right).$$

Remark 5. *Denoting by $\mathcal{P}^\perp_{A,U\times V}$ the operator defined like $\mathcal{P}_A$ using annihilators which project onto the orthogonal of the vector space spanned by the columns of A and U (resp. A and V) for U and V such that the vector spaces have common dimension $r(A, U \times V)$, noting that to obtain* (31) *it is enough that $\widetilde{\Gamma}\mathcal{P}^\perp_{\widetilde{\Gamma},U\times V}(\Delta)^\top = 0$ and $\widetilde{\Gamma}^\top \mathcal{P}^\perp_{\widetilde{\Gamma},U\times V}(\Delta) = 0$, the result of Theorem 1 holds replacing $\kappa_{\widetilde{\Gamma},c(\rho,\widetilde{\rho})}$ by a compatibility constant replacing $\mathcal{P}^\perp_{\widetilde{\Gamma}}$ by $\mathcal{P}^\perp_{\widetilde{\Gamma},U\times V}$, $\mathcal{P}_{\widetilde{\Gamma}}$ by $\mathcal{P}_{\widetilde{\Gamma},U\times V}$, everywhere* rank $\left(\widetilde{\Gamma}\right)$ *by $r\left(\widetilde{\Gamma}, U \times V\right)$, and with an infimum over U and V after the infimum over $\widetilde{\Gamma}$. The freedom over U and V allows to annihilate low-rank components of X_k if it has an approximate factor structure and deliver constants $b_{\perp k}$ which are $O_P\left(\sqrt{\max(N,T)}\right)$.*

Proof of Theorem 2. The first inequalities follow from Theorem 1 so we only prove (20). Due to Assumption 1 (ii), w.p.a. 1, $\widehat{\beta} - \beta = \left(X^\top X\right)^{-1} X^\top (\gamma - \widehat{\gamma}) + \left(X^\top X\right)^{-1} X^\top e$, also

$$\left| X^\top (\gamma - \widehat{\gamma}) \right|_2^2 = \sum_{k=1}^K \left\langle X_k, \widehat{\Gamma} - \Gamma \right\rangle^2 \le \sum_{k=1}^K |X_k|_{\mathrm{op}}^2 \left| \widehat{\Gamma} - \Gamma \right|_*^2 \quad \text{(by trace duality)},$$

$$\left| \left(X^\top X\right)^{-1} X^\top (\gamma - \widehat{\gamma}) \right|_2 \le \frac{1}{NT} \left| \left(\frac{X^\top X}{NT} \right)^{-1} \right|_{\mathrm{op}} \sqrt{\sum_{k=1}^K |X_k|_{\mathrm{op}}^2} \left| \widehat{\Gamma} - \Gamma \right|_*.$$

By Assumption 1 and (18), we obtain $\left|\left(X^\top X\right)^{-1} X^\top(\gamma-\widehat{\gamma})\right|_2 = O_P\left(\lambda_N r_N \mu_N/(NT)\right)$. Next, by (42), we have $\left|\left(X^\top X\right)^{-1} X^\top e\right|_2 = O_P(1/\sqrt{NT})$. This yields the result.

Proof of Proposition 6. The proof techniques are similar to those in [17]. We make use of the fact that if $Z \in \partial|\cdot|_*\left(\widetilde{\Gamma}\right)$, *i.e.*, is of the form

$$Z = \sum_{k=1}^{\operatorname{rank}\left(\widetilde{\Gamma}\right)} u_k\left(\widetilde{\Gamma}\right) v_k\left(\widetilde{\Gamma}\right)^\top + M_{u\left(\widetilde{\Gamma}\right)} W M_{v\left(\widetilde{\Gamma}\right)},$$

for W such that $|W|_{\mathrm{op}} \le 1$, then

$$\left\langle \widehat{Z} - Z, \widehat{\Gamma} - \widetilde{\Gamma} \right\rangle \ge 0 \tag{45}$$

and, for a well chosen matrix W (see [17] page 2308),

$$\left\langle M_{u\left(\widetilde{\Gamma}\right)} W M_{v\left(\widetilde{\Gamma}\right)}, \widetilde{\Gamma} - \widehat{\Gamma} \right\rangle = -\left| M_{u\left(\widetilde{\Gamma}\right)} \widehat{\Gamma} M_{v\left(\widetilde{\Gamma}\right)} \right|_* = -\left| \mathcal{P}^{\perp}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_*.$$

Now, by (9) and (45), we obtain

$$\begin{aligned} &\left\langle M_X\left(\Gamma - \widehat{\Gamma}\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle \\ &\le \lambda\widehat{\sigma}\left\langle Z, \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle - \left\langle M_X\left(E\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle \\ &\le \lambda\widehat{\sigma}\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* \wedge \left|P_{u\left(\widetilde{\Gamma}\right)}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)P_{v\left(\widetilde{\Gamma}\right)}\right|_* - \lambda\widehat{\sigma}\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* - \left\langle M_X\left(E\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle. \end{aligned} \tag{46}$$

We now use

$$2\left\langle M_X\left(\Gamma - \widehat{\Gamma}\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle = \left|M_X\left(\Gamma - \widehat{\Gamma}\right)\right|_2^2 + \left|M_X\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_2^2 - \left|M_X\left(\Gamma - \widetilde{\Gamma}\right)\right|_2^2 \tag{47}$$

and consider cases (1) $\left\langle M_X\left(\Gamma - \widehat{\Gamma}\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle \le 0$ and (2) $\left\langle M_X\left(\Gamma - \widehat{\Gamma}\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle > 0$.
In case (1), due to (47), we have $\left|M_X\left(\Gamma - \widehat{\Gamma}\right)\right|_2^2 \le \left|M_X\left(\Gamma - \widetilde{\Gamma}\right)\right|_2^2$, hence the result.
In case (2), we have

$$\lambda\widehat{\sigma}\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* \le \lambda\widehat{\sigma}\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* - \left\langle M_X\left(E\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle,$$

thus, because $\rho\lambda\widehat{\sigma} \ge |M_X(E)|_{\mathrm{op}}$, $\widetilde{\Gamma} - \widehat{\Gamma} \in C_{\widetilde{\Gamma}}$. Moreover, by (47) and (46), we have

$$\begin{aligned} &\left|M_X\left(\Gamma - \widehat{\Gamma}\right)\right|_2^2 + \left|M_X\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_2^2 + 2\lambda\widehat{\sigma}\left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* \\ &\le \left|M_X\left(\Gamma - \widetilde{\Gamma}\right)\right|_2^2 + 2\lambda\widehat{\sigma}\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* - 2\left\langle M_X\left(E\right), \widetilde{\Gamma} - \widehat{\Gamma}\right\rangle \\ &\le \left|M_X\left(\Gamma - \widetilde{\Gamma}\right)\right|_2^2 + 2\lambda\widehat{\sigma}\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* + 2\rho\lambda\widehat{\sigma}\left(\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_* + \left|\mathcal{P}^{\perp}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_*\right) \end{aligned}$$

and, by definition of $\kappa_{\widetilde{\Gamma},c(\rho)}$,

$$\left|M_X\left(\Gamma - \widehat{\Gamma}\right)\right|_2^2 + \left|M_X\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_2^2 \le \left|M_X\left(\Gamma - \widetilde{\Gamma}\right)\right|_2^2 + 2\lambda(1+\rho)\widehat{\sigma}\left|\mathcal{P}_{\widetilde{\Gamma}}\left(\widetilde{\Gamma} - \widehat{\Gamma}\right)\right|_*$$

$$\leq \left|M_X\left(\Gamma-\widetilde{\Gamma}\right)\right|_2^2+2\lambda(1+\rho)\widehat{\sigma}\frac{\sqrt{2\mathrm{rank}\left(\widetilde{\Gamma}\right)}}{\kappa_{\widetilde{\Gamma},c(\rho)}}\left|M_X\left(\widetilde{\Gamma}-\widehat{\Gamma}\right)\right|_2,$$

hence

$$\frac{1}{NT}\left|M_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2\leq\frac{1}{NT}\left|M_X\left(\Gamma-\widetilde{\Gamma}\right)\right|_2^2+\frac{2(\lambda(1+\rho)\widehat{\sigma})^2}{NT}\frac{\mathrm{rank}\left(\widetilde{\Gamma}\right)}{\kappa^2_{\widetilde{\Gamma},c(\rho)}}.$$

Proof of Proposition 7. (9) yields, for all $k=1,\dots,\mathrm{rank}\left(\widehat{\Gamma}\right)$,

$$\begin{aligned}u_k\left(\widehat{\Gamma}\right)^\top M_X\left(\Gamma^l-\widehat{\Gamma}\right)v_k\left(\widehat{\Gamma}\right)&=\lambda\widehat{\sigma}-u_k\left(\widehat{\Gamma}\right)^\top M_X\left(\Gamma^d+E\right)v_k\left(\widehat{\Gamma}\right)\\&=\lambda\widehat{\sigma}-\left\langle M_X\left(\Gamma^d+E\right),u_k\left(\widehat{\Gamma}\right)v_k\left(\widehat{\Gamma}\right)^\top\right\rangle,\\&\geq\lambda(1-\rho)\widehat{\sigma}-\left|\Gamma^d\right|_{\mathrm{op}},\end{aligned}$$

and, by summing the inequalities,

$$\left\langle\sum_{k=1}^{\mathrm{rank}\left(\widehat{\Gamma}\right)}u\left(\widehat{\Gamma}\right)_k v\left(\widehat{\Gamma}\right)_k^\top,P_{u\left(\widehat{\Gamma}\right)}M_X\left(\Gamma^l-\widehat{\Gamma}\right)P_{v\left(\widehat{\Gamma}\right)}\right\rangle\geq\left(\lambda(1-\rho)\widehat{\sigma}-\left|\Gamma^d\right|_{\mathrm{op}}\right)\mathrm{rank}\left(\widehat{\Gamma}\right),\tag{48}$$

thus

$$\left|P_{u\left(\widehat{\Gamma}\right)}M_X\left(\Gamma^l-\widehat{\Gamma}\right)P_{v\left(\widehat{\Gamma}\right)}\right|_2\geq\left(\lambda(1-\rho)\widehat{\sigma}-\left|\Gamma^d\right|_{\mathrm{op}}\right)\sqrt{\mathrm{rank}\left(\widehat{\Gamma}\right)}.$$

Proposition 11.

Proposition 11. *Let* $m=\left(\frac{|X|_{\mathrm{op}}}{NT}\left|\left(\frac{X^\top X}{NT}\right)^{-1}\right|_{\mathrm{op}}\right)^2\left(\sum_{k=1}^K|X_k|_{\mathrm{op}}^2\right)\left(\mathrm{rank}\left(\Gamma\right)+\mathrm{rank}\left(\widehat{\Gamma}\right)\right)$, *we have*

$$\left|P_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2\leq\frac{m}{(1-m)_+}\left|M_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2,\ \left|\Gamma-\widehat{\Gamma}\right|_2^2\leq\left(1+\frac{m}{(1-m)_+}\right)\left|M_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2.$$

Proof. By Theorem C.5 in [12], the definition of P_X, and the computations in the proof of Theorem 2, we have, w.p.a. 1,

$$\left|P_X\left(\Gamma-\widehat{\Gamma}\right)\right|_2^2\leq\left(\frac{|X|_{\mathrm{op}}}{NT}\left|\left(\frac{X^\top X}{NT}\right)^{-1}\right|_{\mathrm{op}}\right)^2\left(\sum_{k=1}^K|X_k|_{\mathrm{op}}^2\right)\mathrm{rank}\left(\Gamma-\widehat{\Gamma}\right)\left|\widehat{\Gamma}-\Gamma\right|_2^2\leq m\left|\widehat{\Gamma}-\Gamma\right|_2^2.$$

We conclude by the Pythagorean theorem. □

Proof of Proposition 8. By (9), we have $\Gamma^l-\widehat{\Gamma}=\sum_{k=1}^K\left(\widehat{\beta}_k-\beta_k\right)X_k-\Gamma^d-E+\lambda_N\widehat{\sigma}\widehat{Z}$, hence

$$\left|\Gamma-\widehat{\Gamma}\right|_{\mathrm{op}}\leq\left|\widehat{\beta}-\beta\right|_2\sqrt{\sum_{k=1}^K|X_k|_{\mathrm{op}}^2}+\left|\Gamma^d\right|_{\mathrm{op}}+|E|_{\mathrm{op}}+\lambda_N\widehat{\sigma}$$

and we conclude using Theorem 2 and Assumption 2 (ii).

Proof of Proposition 9. The Weyl's inequality, yields, for $k \in \{1, \dots, \min(N,T)\}$,

$$\left|\sigma_k\left(\Gamma^l\right) - \sigma_k\left(\widehat{\Gamma}\right)\right| \le \left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}}.$$

This implies, for $k \le \text{rank}\left(\Gamma^l\right)$,

$$\sigma_k\left(\widehat{\Gamma}\right) \ge \sigma_k\left(\Gamma^l\right) - \left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}} \tag{49}$$

and, for $k > \text{rank}\left(\Gamma^l\right)$,

$$\sigma_k\left(\widehat{\Gamma}\right) \le \left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}}. \tag{50}$$

By Assumption 5 (i) and Proposition 8, we have $\mathbb{P}\left(\left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}} \le (\rho+1)\lambda_N h\sigma\right) \to 1$. By Theorem 2 and $\lambda_N^2 r_N = o(NT)$, we obtain $\mathbb{P}\left((\rho+1)\lambda_N h\sigma < t\right) \to 1$ and, by (50),

$$\mathbb{P}\left(\forall k > \text{rank}\left(\Gamma^l\right),\ t > \sigma_k\left(\widehat{\Gamma}\right)\right) \to 1. \tag{51}$$

By Assumption 5 (ii), we have $\mathbb{P}\left(\sigma_k\left(\Gamma^l\right) - \left|\Gamma^l - \widehat{\Gamma}\right|_{\text{op}} \le (\rho+1)\lambda_N h^3\sigma\right) \to 1$. By Theorem 2 and $\lambda_N^2 r_N = o(NT)$, we obtain $\mathbb{P}\left(t < (\rho+1)\lambda_N h^3\sigma\right) \to 1$ and, by (49),

$$\mathbb{P}\left(\forall k \le \text{rank}\left(\Gamma^l\right),\ t < \sigma_k\left(\widehat{\Gamma}\right)\right) \to 1. \tag{52}$$

Combining (51) and (52), we obtain the first result. The other results are obtained similarly.

Proof of Proposition 10. Because

$$\left|M_{v\left(\widehat{\Gamma}^t\right)} - M_{v\left(\Gamma^l\right)}\right|_2^2 = \left|P_{v\left(\widehat{\Gamma}^t\right)} - P_{v\left(\Gamma^l\right)}\right|_2^2 = \text{rank}\left(\widehat{\Gamma}^t\right) + \text{rank}\left(\Gamma^l\right) - 2\sum_{k=1}^{\text{rank}\left(\Gamma^l\right)} v_k\left(\Gamma^l\right)^\top P_{v\left(\widehat{\Gamma}^t\right)} v_k\left(\Gamma^l\right),$$

$$= \text{rank}\left(\widehat{\Gamma}^t\right) - \text{rank}\left(\Gamma^l\right) + 2\sum_{k=1}^{\text{rank}\left(\Gamma^l\right)} v_k\left(\Gamma^l\right)^\top M_{v\left(\widehat{\Gamma}^t\right)} v_k\left(\Gamma^l\right)$$

$$\left|\Gamma^l M_{v\left(\widehat{\Gamma}^t\right)}\right|_2^2 = \sum_{k=1}^{\text{rank}\left(\Gamma^l\right)} \sigma_k\left(\Gamma^l\right)^2 v_k\left(\Gamma^l\right)^\top M_{v\left(\widehat{\Gamma}^t\right)} v_k\left(\Gamma^l\right),$$

the result follows from

$$\begin{aligned}
\left|M_{v\left(\widehat{\Pi}_v^t\right)} - M_{v\left(\Pi_v^l\right)}\right|_2^2 &\le \left|\text{rank}\left(\widehat{\Gamma}^t\right) - \text{rank}\left(\Gamma^l\right)\right| + \frac{2}{\sigma_{\text{rank}\left(\Gamma^l\right)}\left(\Gamma^l\right)^2}\left|\Gamma^l M_{v\left(\widehat{\Gamma}^t\right)}\right|_2^2 \\
&\le \left|\text{rank}\left(\widehat{\Gamma}^t\right) - \text{rank}\left(\Gamma^l\right)\right| + \frac{2}{\sigma_{\text{rank}\left(\Gamma^l\right)}\left(\Gamma^l\right)^2}\left|\Gamma^l - \widehat{\Gamma}^t\right|_{\text{op}}^2 \left|M_{v\left(\widehat{\Gamma}^t\right)}\right|_2^2 \\
&\le o_P(1) + 2r_N\left(\frac{(\rho+1)\lambda_N(h^2+1)\left(\sigma + o_P(1)\right)}{\sigma_{\text{rank}\left(\Gamma^l\right)}\left(\Gamma^l\right)}\right)^2.
\end{aligned}$$

Proof of Theorem 3. Using that $M_{u(\widehat{\Pi}_u^t)}$ and $M_{v(\widehat{\Pi}_v^t)}$ are self-adjoint, a solution to (25) satisfies, for $l=1,\dots,K$, $\left\langle M_{u(\widehat{\Pi}_u^t)}X_lM_{v(\widehat{\Pi}_v^t)}, Y-\sum_{k=1}^K\widetilde{\beta}_k^{(1)}X_k\right\rangle=0$, hence

$$\begin{aligned}
&\left\langle M_{u(\Pi_u^l)}X_lM_{v(\Pi_v^l)},\Gamma^d+E+\sum_{k=1}^K\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)X_k\right\rangle\\
&=\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_lM_{v(\Pi_v^l)},\Gamma^d+E+\sum_{k=1}^K\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)X_k\right\rangle\\
&\quad+\left\langle M_{u(\Pi_u^l)}X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),\Gamma^d+E+\sum_{k=1}^K\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)X_k\right\rangle\\
&\quad-\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),\Gamma+E+\sum_{k=1}^K\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)X_k\right\rangle,
\end{aligned}$$

so

$$\begin{aligned}
&\sum_{k=1}^K\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)\Bigg(\left\langle M_{u(\Pi_u^l)}X_lM_{v(\Pi_v^l)},X_k\right\rangle-\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_lM_{v(\Pi_v^l)},X_k\right\rangle\\
&\quad-\left\langle M_{u(\Pi_u^l)}X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),X_k\right\rangle\\
&\quad+\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),X_k\right\rangle\Bigg)\\
&=-\left\langle M_{u(\Pi_u^l)}X_lM_{v(\Pi_v^l)},\Gamma^d+E\right\rangle+\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_lM_{v(\Pi_v^l)},\Gamma^d+E\right\rangle\\
&\quad+\left\langle M_{u(\Pi_u^l)}X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),\Gamma^d+E\right\rangle\\
&\quad-\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_l\left(M_{v(\Pi_v^l)}-M_{v(\widehat{\Pi}_v^t)}\right),\Gamma+E\right\rangle. \qquad (53)
\end{aligned}$$

Let us show that $\left\langle M_{u(\Pi_u^l)}X_lM_{v(\Pi_v^l)},X_k\right\rangle$, which by Assumption 9 (v) diverges like NT, is the high-order term multiplying $\left(\beta_k-\widetilde{\beta}_k^{(1)}\right)$ in (53). This also yields the consistency of the estimator of the covariance matrix. By self-adjointness of the projections, Theorem C.5 in [12], and Proposition 9 with the modifications of Section 4.7 which imply rank $\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)\le 2\overline{r}_N$ w.p.a. 1, denoting, for a matrix M and $r\in\mathbb{N}$ by $|M|_{2,r}^2=\sum_{k=1}^r\sigma_k(M)^2$, we have,

$$\begin{aligned}
&\left|\left\langle\left(M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right)X_lM_{v(\Pi_v^d)},X_k\right\rangle\right|\\
&\le(1+o_P(1))\left|M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right|_2\left|X_lM_{v(\Pi_u^l)}X_k^\top\right|_{2,2r_N}\\
&\le\left(\sqrt{2\overline{r}_N}+o_P(1)\right)\left|M_{u(\Pi_u^l)}-M_{u(\widehat{\Pi}_u^t)}\right|_2\left|X_lM_{v(\Pi_v^l)}\right|_{\mathrm{op}}\left|X_kM_{v(\Pi_v^l)}\right|_{\mathrm{op}},
\end{aligned}$$

hence, by Proposition 10 with the modifications of Section 4.7,

$$\begin{aligned}&\left|\left\langle\left(M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right)X_l M_{v\left(\widehat{\Pi}_v^t\right)},X_k\right\rangle\right|\\&\le\frac{2(\rho+1)\overline{r}_N\lambda_N}{\overline{v}_N}\left((h^2+1)\widetilde{\sigma}+o_P\left(1\right)\right)\left|\left(\Pi_l^d+U_l\right)M_{v\left(\Pi_v^l\right)}\right|_{\mathrm{op}}\left|\left(\Pi_k^d+U_k\right)M_{v\left(\Pi_v^l\right)}\right|_{\mathrm{op}}.\end{aligned}$$

We treat similarly $\left|\left\langle M_{u\left(\Pi_u^l\right)}X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right),X_k\right\rangle\right|$, and, for the fourth term, use that

$$\begin{aligned}&\left|\left\langle\left(M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right)X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right),X_k\right\rangle\right|\\&\le\left|\left(M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right)X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right)\right|_*|X_k|_{\mathrm{op}}\\&\le\left(\sqrt{2\overline{r}_N}+o_P(1)\right)\left|\left(M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right)X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right)\right|_2|X_k|_{\mathrm{op}}\\&\le\left(\sqrt{2\overline{r}_N}+o_P(1)\right)\left|M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right|_{\mathrm{op}}\left|X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right)\right|_2|X_k|_{\mathrm{op}}\\&\le\left(\sqrt{2\overline{r}_N}+o_P(1)\right)\left|M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right|_2|X_l|_{\mathrm{op}}\left|M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right|_2|X_k|_{\mathrm{op}}\\&\le\frac{(\rho+1)^2(2\overline{r}_N)^{3/2}\lambda_N^2}{\overline{v}_N^2}\left((h^2+1)^2\widetilde{\sigma}^2+o_P\left(1\right)\right)|X_l|_{\mathrm{op}}|X_k|_{\mathrm{op}},\end{aligned}$$

where we use Proposition 9 in the third display and Proposition 10 (with the modifications of Section 4.7) in the last display. Let us consider now the quantities on the right-hand side in (53). Proceeding like above, we have

$$\begin{aligned}&\left|\left\langle\left(M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right)X_l M_{v\left(\Pi_v^l\right)},\Gamma^d+E\right\rangle\right|\\&\le\left(1+o_P(1)\right)\left|M_{u\left(\Pi_u^l\right)}-M_{u\left(\widehat{\Pi}_u^t\right)}\right|_2\left|X_l M_{v\left(\Pi_v^l\right)}\left(\Gamma^d+E\right)^\top\right|_{2,2r_N}\\&\le\frac{2(\rho+1)\overline{r}_N\lambda_N\left((h^2+1)\widetilde{\sigma}+o_P\left(1\right)\right)}{\overline{v}_N}\left(\rho\lambda_N\sigma+\left|\Gamma^d M_{v\left(\Pi_v^l\right)}\right|_{\mathrm{op}}\right)\left(\rho\lambda_N\sigma_l+\left|\Pi_l^d M_{v\left(\Pi_v^l\right)}\right|_{\mathrm{op}}\right)\end{aligned}$$

and treat similarly $\left\langle M_{u\left(\Pi_u^l\right)}X_l\left(M_{v\left(\Pi_v^l\right)}-M_{v\left(\widehat{\Pi}_v^t\right)}\right),\Gamma^d+E\right\rangle$. With the same arguments, the absolute value of the last term of (53) is smaller than

$$\frac{(\rho+1)\sqrt{2}(2\overline{r}_N)^{3/2}\lambda_N^2\left((h^2+1)^2\widetilde{\sigma}^2+o_P\left(1\right)\right)}{\overline{v}_N^2}|X_l|_{\mathrm{op}}\left(|\Gamma|_{\mathrm{op}}+\rho\lambda_N(h^2+1)\widetilde{\sigma}+o_P(1)\right).$$

Let us now look at the first terms on the left-hand side and on the right-hand side of (53). By (iv), for all $k,l\in\{1,\dots,K\}$,

$$\left\langle M_{u\left(\Pi_u^l\right)}X_l M_{v\left(\Pi_v^l\right)},X_k\right\rangle=\left\langle M_{u\left(\Pi_u^l\right)}U_l M_{v\left(\Pi_v^l\right)},U_k\right\rangle+o_P(NT)$$

so, by (v), $\left\langle M_{u\left(\Pi_u^l\right)}X_l M_{v\left(\Pi_v^l\right)},X_k\right\rangle$ are the high-order terms on the left-hand side of (53). Similarly, by (iv), the high-order terms on the right-hand side of (53) are $\left\langle M_{u\left(\Pi_u^l\right)}U_l M_{v\left(\Pi_v^l\right)},E\right\rangle$.

As a result, $\widetilde{\beta}^{(1)}$ is asymptotically equivalent to the ideal estimator $\overline{\beta}$

$$\overline{\beta} \in \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \left| \mathcal{P}_{\Pi^l}^{\perp}\left(Y - \sum_{k=1}^{K} \beta_k U_k \right) \right|_2^2 . \tag{54}$$

Hence, w.p.a. 1, $\overline{\beta} - \beta = \left(\mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U) \right)^{-1} \mathcal{P}_{\Pi^l}^{\perp}(U)^\top e$ and we conclude by usual arguments. To obtain the first part of the second statement we use that $U^\top U - \mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U)$ is symmetric positive definite. It is clearly symmetric. The positive definiteness follows from the following computations. Let $b \in \mathbb{R}^K$, we have

$$\begin{aligned}
\sum_{k,l} b_k b_l \operatorname{tr}\left(U_k^\top U_l \right) &\geq \sum_{k,l} b_k b_l \operatorname{tr}\left(M_{v(\Pi_v^l)} U_k^\top U_l \right) \\
&= \sum_{k,l} b_k b_l \operatorname{tr}\left(M_{v(\Pi_v^l)} U_k^\top M_{u(\Pi_v^l)} U_l M_{v(\Pi_v^l)} \right) + \sum_{k,l} b_k b_l \operatorname{tr}\left(M_{v(\Pi_v^l)} U_k^\top P_{u(\Pi_v^l)} U_l M_{v(\Pi_v^l)} \right) \\
&\geq \sum_{k,l} b_k b_l \operatorname{tr}\left(\mathcal{P}_{\Pi^l}^{\perp}(U_k)^\top \mathcal{P}_{\Pi^l}^{\perp}(U_l) \right).
\end{aligned}$$

Because $U^\top U$ has a fixed dimension, all norms are equivalent and $\left| U^\top U - \mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U) \right|_{\mathrm{op}} \leq \operatorname{tr}\left(U^\top U - \mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U) \right) = |\mathcal{P}_{\Pi^l}(U)|_2^2 = o_P(|U|_2^2)$. We conclude using that $|U|_2^2 \leq K \left| U^\top U \right|_{\mathrm{op}}$. Also, from the above, $\mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U) = \mathcal{P}_{\Pi^l}^{\perp}(U)^\top \mathcal{P}_{\Pi^l}^{\perp}(U) + M$ where M is a smaller order term by condition (iv). We obtain the last part of the second statement using the next lemma.

Lemma 2. *Assume U and (Π_u^l, Π_v^l) are independent, and $\mathbb{E}\left[\max\left(\operatorname{rank}\left(\Pi_u^l \right), \operatorname{rank}\left(\Pi_v^l \right) \right) \right] = o\left(\sqrt{\min(N,T)} \right)$, then $|\mathcal{P}_{\Pi^l}(U)|_2^2 / (NT) = o_P(1)$, hence $\mathcal{P}_{\Gamma^r}^{\perp}(U)^\top \mathcal{P}_{\Gamma^r}^{\perp}(U)/(NT) \xrightarrow{\mathbb{P}} \mathbb{E}\left[U^\top U \right]$.*

Proof. We prove that, for $k \in \{1, \dots, K\}$, $|\mathcal{P}_{\Pi^l}(U_k)|_2^2 / (NT)$ converges to 0 in L^1. This relies on (26) and the facts that $M_{u(\Pi^l)}$ is a contraction for the Euclidian norm and

$$\begin{aligned}
\mathbb{E}\left[\left| U_k P_{v(\Pi^l)} \right|_2^2 \right] &= \mathbb{E}\left[\mathbb{E}\left[\left| U_k P_{v(\Pi^l)} \right|_2^2 \middle| \Pi_u^l, \Pi_v^l \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} \left| U_{i\cdot} P_{v(\Pi^l)} \right|_2^2 \middle| \Pi_u^l, \Pi_v^l \right] \right] = N \mathbb{E}\left[\operatorname{rank}\left(\Pi_v^l \right) \right] u^2 = o(NT)
\end{aligned}$$

and similarly for $\mathbb{E}\left[\left| P_{u(\Pi^l)} U_k \right|_2^2 \right]$. By the arguments in the previous proof $U^\top U/(NT)$ and $\mathcal{P}_{\Gamma^r}^{\perp}(U)^\top \mathcal{P}_{\Gamma^r}^{\perp}(U)/(NT)$ have same limit, hence the result by the law of large numbers. □

Toulouse School of Economics, Université Toulouse Capitole, 21 allée de Brienne, 31000 Toulouse, France

E-mail address: jad.beyhum@tse-fr.eu, eric.gautier@tse-fr.eu