Econometric Theory, **40**, 2024, 1391–1421.
doi:10.1017/S0266466623000129

NUCLEAR NORM REGULARIZED QUANTILE REGRESSION WITH INTERACTIVE FIXED EFFECTS

JUNLONG FENG
Hong Kong University of Science and Technology

This paper studies large N and large T conditional quantile panel data models with interactive fixed effects. We propose a nuclear norm penalized estimator of the coefficients on the covariates and the low-rank matrix formed by the interactive fixed effects. The estimator solves a convex minimization problem, not requiring pre-estimation of the (number of) interactive fixed effects. It also allows the number of covariates to grow slowly with N and T. We derive an error bound on the estimator that holds uniformly in the quantile level. The order of the bound implies uniform consistency of the estimator and is nearly optimal for the low-rank component. Given the error bound, we also propose a consistent estimator of the number of interactive fixed effects at any quantile level. We demonstrate the performance of the estimator via Monte Carlo simulations.

1. INTRODUCTION

Panel data models are widely applied in economics and finance. Allowing for rich heterogeneity, interactive fixed effects are important components in such models in many applications. In applications such as asset pricing, it could be desirable to explain or forecast an outcome variable at certain quantile levels. However, the well-studied mean regression with interactive fixed effects (e.g., Pesaran, 2006; Bai, 2009; Moon and Weidner, 2015) misses such distributional heterogeneity.

In this paper, we consider a panel data model where the conditional quantile of an outcome variable is linear in the covariates and in the product of time and individual fixed effects. These interactive fixed effects are unobservables that may be correlated with the covariates. The number of covariates is allowed to grow slowly to infinity with N and T. Meanwhile, we allow the coefficients, the set of

This paper is based on the third chapter of my doctoral dissertation at Columbia. An earlier working paper version of it was circulated under the title “Regularized Quantile Regression with Interactive Fixed Effects.” I am very grateful for helpful comments provided by the Editor, the Co-Editor, and three referees. I thank Jushan Bai, Sokbae (Simon) Lee, and Bernard Salanié, who were gracious with their advice, support, and feedback. I have also greatly benefited from comments and discussions with Songnian Chen, Roger Koenker, José Luis Montiel Olea, Roger Moon, Serena Ng, Jörg Stoye, Peng Wang, Martin Weidner, and participants at the Columbia Econometrics Colloquium, 2021 Econometric Society North American, Asian, and China meetings, and 2021 IAAE. All errors are my own. Address correspondence to Junlong Feng, Department of Economics, Hong Kong University of Science and Technology, Kowloon, Hong Kong; e-mail: jlfeng@ust.hk.

the effective fixed effects and the realization of each of them to all be quantile-level dependent, generating large modeling flexibility.

To estimate the model, this paper proposes a *nuclear norm penalized estimator*. The nuclear norm of a matrix is equal to the sum of all its singular values. By deriving the estimator's theoretical error bound, we show that it can consistently estimate the coefficients and the (realizations of) the interactive fixed effects uniformly in quantile level. The estimator solves a convex problem and computes quickly in practice even in large panel datasets based on our proposed augmented Lagrangian multiplier algorithm. Implementing the estimator does not require pre-estimation of the number of interactive fixed effects or the fixed effects themselves.

To illustrate the estimator, let us consider a simple example where only the coefficients are quantile-level dependent: For a panel dataset $\{(Y_{it}, X_{it}) : i = 1, \ldots, N; t = 1, \ldots, T\}$, suppose the uth conditional quantile of Y_{it} given p covariates X_{it} and r time and individual fixed effects (F_t, Λ_i) is $q_{Y_{it}|X_{it}, F_t, \Lambda_i}(u) := X_{it}'\beta_0(u) + F_t'\Lambda_i$, where both F_t and Λ_i are $r \times 1$ vectors. The interactive fixed effects form an $N \times T$ matrix $L_0 := (\Lambda_1, \ldots, \Lambda_N)'(F_1, \ldots, F_T)$, whose rank is at most r. Thus, the dense matrix L_0 is low-rank, when r is small relative to N and T. Exploiting such low-rankness, our estimator, inspired by the seminal work by Candès and Recht (2009), jointly estimates $(\beta_0(u), L_0)$ for a given quantile level $u \in (0,1)$ by solving

$$\min_{\beta \in \mathcal{B}, L \in \mathcal{L}} \frac{1}{NT} \sum_{i,t} \rho_u (Y_{it} - X_{it}'\beta - L_{it}) + \lambda \|L\|_*, \tag{1.1}$$

where ρ_u is the standard check function in the quantile regression literature, λ is a positive penalty coefficient, $\|L\|_*$ is the nuclear norm of the $N \times T$ matrix L, and $\mathcal{B}$ and $\mathcal{L}$ are convex parameter spaces about which we will be specific later.

The key component of the estimator is the convex nuclear norm penalty. Summing up the singular values of a matrix, the nuclear norm is to the rank, counting the nonzero singular values, what the convex ℓ_1-norm is to the nonconvex ℓ_0-norm of the vector of the singular values. Hence, the nuclear norm penalty can be viewed as the matrix counterpart of the LASSO penalty in regression with high-dimensional regressors. Being a convex surrogate of the rank functional, we show that this penalty is effective to deliver consistent estimates under low-rankness of L_0.

Setting up the minimization problem as in (1.1) yields a convex objective function in (β, L), and one does not need to know r before implementation. To highlight these benefits, let us consider a natural *alternative estimator*

$$\min_{\beta \in \mathcal{B}, \Lambda \in \Xi, F \in \mathcal{F}} \frac{1}{NT} \sum_{i,t} \rho_u (Y_{it} - X_{it}'\beta - \Lambda_i' F_t), \tag{1.2}$$

where $\mathcal{B}$, Ξ and $\mathcal{F}$ are parameter spaces for β, Λ, and F, respectively. Ando and Bai (2020) study a similar estimator where the coefficients are i-specific. The objective function in (1.2) is nonconvex in the parameters (β, Λ, F). As the minimization problem needs to be solved iteratively, nonconvexity leads to two potential issues.

First, one may obtain a local minimum which can be arbitrarily far from the global one. Second, solving the minimization problem may be computationally intensive, especially in large panel datasets. In the simulation experiments in the paper, we find that it computes much slower than the penalized estimator we propose. Meanwhile, to make the alternative estimator (1.2) feasible, r needs to be known or pre-estimated. This step results in additional computation burden and a misspecified r may lead to inconsistent estimates.

Besides convexity and the convenience brought by avoiding estimating r, we show that our approach allows for certain low-rank covariates. This finding echoes a similar result in Moon and Weidner (2019) on nuclear norm penalized mean regression. A covariate is low-rank when, for instance, it has a factor structure. Unlike this paper, such covariates are usually excluded in interactive fixed effects regression for identification purposes (Bai, 2009; Moon and Weidner, 2015).

A major drawback of the nuclear norm penalized estimator is that the rate of convergence of the coefficient estimator is slow compared to, for instance, the rate in mean regressions when the interactive fixed effects are estimated separately (e.g., Bai, 2009). We propose a consistent estimator of r based on our penalized estimator. With this rank estimator, treating our consistent penalized estimator as an initial value for estimator (1.2) with a few rounds of iterations might attain a faster rate of convergence (see Chernozhukov et al., 2019; Moon and Weidner, 2019 for mean regressions) and reduce computation time. On the other hand, the error bound on the estimator of L_0 can be nearly optimal in squared Frobenius norm.

With the dense latent component and the nonsmooth objective function involved, deriving the estimator's uniform error bound is challenging. We prove new results on random matrices for this purpose. Moreover, we develop novel theoretical arguments which relax some usual assumptions or replace some high-level technical conditions in the panel data quantile regression literature with primitive ones that are easier to interpret. These arguments are useful to study quantile regression with fixed effects under other setups.

This paper adds to the literature of panel data quantile regression. Since Koenker (2004), panel data quantile regression began to draw increasing attention. Abrevaya and Dahl (2008), Lamarche (2010), Canay (2011), Kato, Galvao, Jr., and Montes-Rojas (2012), Galvao, Lamarche, and Lima (2013), and Galvao and Kato (2016) study quantile regression with one-way or two-way fixed effects. Harding and Lamarche (2014) consider interactive fixed effects with endogenous regressors. They require the factors to be pre-estimated or known. Chen (2022) considers quantile regression with interactive fixed effects. They need to first estimate the time fixed effects, or, the factors, that are assumed to be quantile-level independent. They then estimate the coefficients and the individual fixed effects via smoothed quantile regression. Chen, Dolado, and Gonzalo (2021) propose a quantile factor model without regressors. They estimate the factors and the factor loadings via nonconvex minimization similar to (1.2). Pre-estimation of the number of factors is needed. Ando and Bai (2020) consider quantile regression with heterogeneous coefficients and a factor structure. They propose both a frequentist and a Bayesian estimation procedure. The number of factors also needs to be estimated first, and the minimization problem is noncovex. Both Ando and Bai (2020) and Chen et al. (2021) establish consistency pointwise in quantile level, while we focus on uniform consistency. On the technical side, both impose stronger assumptions on the conditional density of the outcome variable than our paper.[1] In our simulation study, we find that our estimator is computationally more efficient.

Another literature this paper speaks to is on nuclear norm penalized estimation. This literature was initially motivated by low-rank matrix completion or recovery problems in computer science and statistics (e.g., Candès and Recht, 2009; Ganesh et al., 2010; Zhou et al., 2010; Candès et al., 2011; Hsu, Kakade, and Zhang, 2011; Negahban and Wainwright, 2011; Agarwal, Negahban, and Wainwright, 2012; Negahban et al., 2012, among others). In this literature, the outcome matrix is usually modeled as the sum of a low-rank matrix and some other matrices that are, for instance, sparse or Gaussian. The primary goal is to estimate the low-rank or the sparse matrix. This setup is different from our paper. Nuclear norm penalized estimation and matrix completion related topics have also gained interest in econometrics recently. Chernozhukov et al. (2019), Moon and Weidner (2019),[2] and Athey et al. (2021) investigate nuclear norm penalized mean regression with interactive fixed effects. Beyhum and Gautier (2019) also consider mean regression with interactive fixed effects, but they use a square-root nuclear norm penalty. Bai and Feng (2019) propose a nuclear norm regularized median regression for robust principal component analysis for fat tailed data. Bai and Ng (2021) consider imputation of missing data and counterfactuals. Bai and Ng (2019) study penalized estimation for approximate factor models with singular values thresholding. Chao, Härdle, and Yuan (2021) consider penalized multi-task quantile regression where there are multiple outcome variables and the coefficient matrix is low-rank. Ma, Su, and Zhang (2022) apply nuclear norm penalized logistic regression to study an undirected network formation model.

A recent paper done in parallel with ours by Belloni et al. (2023) studies quantile regression with both interactive fixed effects and high-dimensional regressors. Besides the nuclear norm penalty, they have an additional ℓ_1-norm constraint on the coefficients to deal with the high-dimensional regressors. In contrast, we focus on low-dimensional regressors, although we do allow the number of regressors to slowly grow to infinity. On the other hand, unlike our paper that derives a uniform error bound, they focus on convergence rate pointwise in quantile level. The two papers also differ in assumptions and algorithms. We provide a more detailed

[1] We will discuss these differences in detail in Section S.A.2 of the Supplementary Material.

[2] Moon and Weidner (2019) also briefly discuss nuclear norm penalized quantile regression with a single regressor as an extension. Using a different approach than this paper, they focus on pointwise (in quantile level) convergence rate of the coefficient estimator. In this paper, we obtain uniform rates for both the coefficients and the low-rank component. Also, the number of covariates can be more than one and growing to infinity slowly.

discussion in Section S.A.2 of the Supplementary Material. We view these two papers as complementary.

The rest of the paper is organized as follows: Section 2 introduces the model and the estimator. Section 3 previews the main results and provides some preliminary results. Section 4 proves identification on a restricted set. Section 5 derives the uniform error bound. Section 6 shows Monte Carlo simulation results. Section 7 concludes. The algorithm and implementation details are in the Appendix A. The Supplementary Material provides an alternative approach to proving consistency, compares the assumptions in this paper with the most related literature, and collects all the proofs.

1.1. Notation

Besides the nuclear norm $\|\cdot\|_*$, four additional matrix norms are used in the paper: Let $\|\cdot\|$, $\|\cdot\|_F$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$ denote the spectral norm, the Frobenius norm, the ℓ_1-norm, and the maximum norm. When applied to a vector, the Frobenius norm is equal to the euclidean norm. For two arbitrary $N \times T$ matrices A and B, $\langle A, B\rangle := \sum_{i,t} A_{it}B_{it}$ denotes the inner product of A and B. For two real numbers, $a \vee b$ and $a \wedge b$ return the maximum and the minimum of a and b, respectively.

2. THE MODEL AND THE ESTIMATOR

We consider a panel dataset $\{(Y_{it}, X_{it}) : i = 1,\ldots,N; t = 1,\ldots,T\}$, where Y_{it} is a scalar outcome and X_{it} is a $p \times 1$ vector of covariates. Let $Y = (Y_{it})$ and $X_j = (X_{j,it})$ $(j = 1,\ldots,p)$ be $N \times T$ matrices of the outcome and the jth covariate. Let $\mathcal{U}$ be a compact subset of $(0,1)$. Throughout the paper, we assume that $\bar{r}$ is fixed, not changing with N and T. For any $u \in \mathcal{U}$, there are $\bar{r}$ possibly u-dependent time and individual fixed effects. For $k = 1,\ldots,\bar{r}$, let $F_k(u) = (F_{1k}(u),\ldots,F_{Tk}(u))'$ be the kth time fixed effect. Let $\Lambda_k(u) = (\Lambda_{1k}(u),\ldots,\Lambda_{Nk}(u))'$ be the kth individual fixed effects. Let $W_X = (X_1,\ldots,X_p)$, $W_L = (\{F_k(u)\}_{k=1,\ldots,\bar{r},u\in\mathcal{U}}, \{\Lambda_k(u)\}_{k=1,\ldots,\bar{r},u\in\mathcal{U}})$, and $W = (W_X, W_L)$. Assume, for all $u \in \mathcal{U}$, the conditional quantile of outcome Y_{it} in matrix notation satisfies the following model with probability one:

$$q_{Y|W}(u) = \sum_{j=1}^{p} X_j\beta_{0,j}(u) + \sum_{k=1}^{\bar{r}} \mathbb{1}_k(u)\Lambda_k(u)F_k(u)' \tag{2.1}$$

$$=: \sum_{j=1}^{p} X_j\beta_{0,j}(u) + L_0(u), \tag{2.2}$$

where $\mathbb{1}_k(u) \in \{0,1\}$ determines whether the kth interactive fixed effect $F_k(u)$ or $\Lambda_k(u)$ affects the uth conditional quantile of Y at all. Model (2.1) allows both the effective fixed effects and the realizations of them to depend on u. Throughout, we allow the fixed effects to be either random or deterministic. When they are random, the covariates can be correlated with them, and all the stochastic statements in this paper are conditional on their realization. Similar setups can be found in Ando and Bai (2020) and Chen et al. (2021).

The fixed effects in equation (2.1) form an $N \times T$ matrix $L_0(u) := \sum_{k=1}^{\bar{r}} \mathbb{1}_k(u) \Lambda_k(u) F_k(u)'$. The rank of the matrix $L_0(u)$ is at most $r(u) := \sum_{k=1}^{\bar{r}} \mathbb{1}_k(u) \leq \bar{r}$ by construction. Since we assume that $\bar{r}$ is fixed, $L_0(u)$ is low-rank, when N and T are large.

Let $\beta_0(u) = (\beta_{0,j}(u))_{j=1,\ldots,p}$. This paper focuses on consistently estimating $(\beta_0(u), L_0(u))$ uniformly in $u \in \mathcal{U}$. When $L_0(u)$ is random, consistency is in terms of its realization.

Remark 1. When $\mathcal{U}$ is a singleton containing u, the conditioning variables W only contain the covariates and the fixed effects at u. The model is then in line with the models in the literature on panel data quantile regression that focus on a fixed $u \in (0,1)$, for example, Harding and Lamarche (2014), Ando and Bai (2020), and Chen et al. (2021).

Now, let us present a few models which admit the conditional quantile function (2.1).

Example 1 (A location shift model with one-way fixed effects only). Suppose the outcome matrix Y is determined by the following linear model with only individual fixed effects $\Lambda^o = (\Lambda_1^o, \ldots, \Lambda_N^o)'$ (similarly, one can also consider a model with time fixed effects only):

$$Y = \beta^o \cdot \mathbf{1}_{N\times T} + \sum_{j=1}^{p} X_j \beta_{0,j} + \Lambda^o \cdot \mathbf{1}_{1\times T} + \epsilon,$$

where ϵ is an $N \times T$ error matrix, $\mathbf{1}_{N\times T}$ and $\mathbf{1}_{1\times T}$ are $N \times T$ and $1 \times T$ matrices of ones. Assume $\{X_j\}_j \perp\!\!\!\perp \epsilon | \Lambda^o$, whereas $\{X_j\}_j$ and Λ^o, as well as Λ^o and ϵ, can be correlated. Assume the ϵ_{it}s, conditional on Λ^o, are identically distributed on $\mathbb{R}$ across t. Let $q_{\epsilon_i|\Lambda^0}(\cdot)$ denote the conditional quantile function of ϵ_{it}. Then the uth conditional quantile of Y is $q_{Y|W}(u) = q_{Y|\{X_j\}_j,\, \Lambda(u)}(u) = \sum_{j=1}^{p} X_j \beta_{0,j} + \Lambda(u) F'$ with probability one for all $u \in (0,1)$, where $\Lambda(u) = \Lambda^o + q_{\epsilon|\Lambda^o}(u) + \beta^o \cdot \mathbf{1}_{N\times 1}$, $q_{\varepsilon|\Lambda^o}(u) = (q_{\epsilon_1|\Lambda^o}(u), \ldots, q_{\epsilon_N|\Lambda^o}(u))'$, and $F = \mathbf{1}_{T\times 1}$. The fixed effects form an $N \times T$ matrix with identical columns and $\bar{r} = 1$.

Example 2 (A location-scale model with interactive fixed effects). Suppose

$$Y = \sum_{j=1}^{p} X_j \beta_{0,j}^a + \sum_{k=1}^{\bar{r}_1} \Lambda_k^a F_k^{a'} + \left(\sum_{j=1}^{p} X_j \beta_{0,j}^b + \sum_{m=1}^{\bar{r}_2} \Lambda_m^b F_m^{b'} \right) \circ \epsilon,$$

where $\circ$ denotes the Hadamard product of matrices. Assume that $(\{X_j\}_j, \{F_k^a, \Lambda_k^a\}_k, \{F_m^b, \Lambda_m^b\}_m) \perp\!\!\!\perp \epsilon$, and let $q_\epsilon(\cdot)$ be the quantile function of the identically distributed ϵ_{it}s. If, for all i, t, and m, and all x, f_m^b, and λ_m^b in the support sets of X_{it}, F_{tm}^b, and Λ_{im}^b, the inequality $x' \beta_0^b + \sum_{m=1}^{\bar{r}_2} f_m^b \lambda_m^b > 0$ holds, then for any $u \in (0,1)$, by letting $\beta_0(u) = \beta_0^a + \beta_0^b q_\epsilon(u)$, $\Lambda(u) = (\Lambda_1^a, \ldots, \Lambda_{\bar{r}_1}^a, \Lambda_1^b q_\epsilon(u), \ldots, \Lambda_{\bar{r}_2}^b q_\epsilon(u))$

and $F = (F^a_1, \ldots, F^a_{\bar{r}_1}, F^b_1, \ldots, F^b_{\bar{r}_2})$, the uth conditional quantile of Y is $q_{Y|W}(u) = q_{Y|\{X_j\}_j, \{F_l, \Lambda_l(u)\}_l}(u) = \sum_{j=1}^{p} X_j \beta_{0,j}(u) + \sum_{l=1}^{\bar{r}_1+\bar{r}_2} \Lambda_l(u) F_l'$ with probability one. In this example, the coefficients and the individual fixed effects are u-dependent.

Example 3 (A random coefficient model with quantile-dependent fixed effects). Let $U_{it} \sim \text{Unif}[0,1]$ and

$$Y_{it} = X_{it}'\beta_0(U_{it}) + \sum_{k=1}^{\bar{r}-1} \mathbb{1}_k(U_{it}) F^o_{tk}(U_{it}) \Lambda^o_{ik}(U_{it}) + \epsilon_{it}.$$

Assume that $X_{it}'\beta_0(u) + \sum_{k=1}^{\bar{r}-1} \mathbb{1}_k(u) F^o_{tk}(u) \Lambda^o_{ik}(u)$ is strictly increasing in u for all realizations of X_{it} and the fixed effects. For instance, suppose $\mathbb{1}_1(u) = 1$, $\mathbb{1}_2(u) = \mathbb{1}(u > 0.3)$, etc. and the product of the fixed effects are positive and increasing in u. Or, the indicator function may not increase in u, for instance, $\bar{r} - 1 = 2$ and $\mathbb{1}_1(u) = 1$ if $u \leq 0.5$ while $\mathbb{1}_2(u) = 1$ if $u > 0.5$, then the assumption is satisfied if both $F^o_{t1}(u)\Lambda^o_{i1}(u)$ and $F^o_{t2}(u)\Lambda^o_{i2}(u)$ are strictly increasing in u with $F^o_{t1}(0.5)\Lambda^o_{i1}(0.5) \leq F^o_{t2}(0.5)\Lambda^o_{i2}(0.5)$. Suppose ϵ_{it} is generated by some strictly increasing function of U_{it}, $G^{-1}(U_{it})$. Finally, assume that $\{U_{it}\}_{i,t}$ are independent of $(\{X_{it}, \{F^o_{tk}(u)\}_{k,u}, \{\Lambda^o_{ik}(u)\}_{k,u}\}_{i,t})$. Then, by strict monotonicity and independence, the uth conditional quantile of Y is $q_{Y|W}(u) = \sum_{j=1}^{p} X_j \beta_{0,j}(u) + \sum_{k=1}^{\bar{r}} \mathbb{1}_k(u) \Lambda_k(u) F_k(u)'$ with probability one for all $u \in (0,1)$, where, for $k < \bar{r}$, $F_k(u) = F^o_k(u)$ and $\Lambda_k(u) = \Lambda^o_k(u)$. For $k = \bar{r}$, $\Lambda_{\bar{r}}(u) = G^{-1}(u)\mathbf{1}_{N\times 1}$ while $\mathbb{1}_{\bar{r}}(u) = 1$ and $F_{\bar{r}}(u) = \mathbf{1}_{T\times 1}$ for all $u \in (0,1)$. In this example, the coefficients, fixed effects, and the set of the effective fixed effects all depend on u. We will revisit this example in our Monte Carlo experiment in Section 6.

Now, we introduce our estimator of $(\beta_0(u), L_0(u))$. For an arbitrary $N \times T$ matrix Z, define $\boldsymbol{\rho}_u(Z) := \sum_{i,t} \rho_u(Z_{it}) \equiv \sum_{i,t} Z_{it}(u - \mathbb{1}(Z_{it} \leq 0))$. By exploiting the linearity of the conditional quantile function (2.2) in $(\beta_0(u), L_0(u))$ and the low-rankness of $L_0(u)$, this paper proposes the following *nuclear norm penalized quantile regression estimator* to jointly estimate $\beta_0(u)$ and $L_0(u)$, for any $u \in \mathcal{U}$:

$$(\hat{\beta}(u), \hat{L}(u)) := \arg\min_{\beta \in \mathbb{R}^p, L \in \mathcal{L}} \frac{1}{NT} \boldsymbol{\rho}_u \left(Y - \sum_{j=1}^{p} X_j \beta_j - L \right) + \lambda \|L\|_*, \tag{2.3}$$

where $\lambda > 0$. The parameter space for the matrix component $\mathcal{L} := \{L \in \mathbb{R}^{N\times T} : \|L\|_\infty \leq \alpha_{NT}\}$ is convex and compact and $\alpha_{NT} \geq 1$ can be (N,T)-dependent. In particular, we allow α_{NT} to *grow to infinity* with N and T. We need α_{NT} for technical reasons to be discussed in Section 4. In Section S.A.1 of the Supplementary Material, we show that we can drop α_{NT} to make $\mathcal{L} = \mathbb{R}^{N\times T}$ under a different set of assumptions.

Two remarks on the estimator are in order.

Remark 2. The estimator does not directly penalize or constrain the rank of the estimated interactive fixed effect matrix to avoid nonconvexity. Instead, it seeks an $\hat{L}(u)$ that has a small nuclear norm. Intuitively, this is reasonable because the rank-$r(u)$ matrix $L_0(u)$ itself typically has a small nuclear norm by low-rankness. To achieve a small nuclear norm, the penalty would shrink some of $\hat{L}(u)$'s singular values even though the rank of $\hat{L}(u)$ may still remain high since rank may increase dramatically even by a very small perturbation to $L_0(u)$. For instance, the $(r(u)+1)$th to the $(N \wedge T)$th singular values in $\hat{L}(u)$ can be nonzero, but they may have a smaller order than the first $r(u)$ singular values. This is shown in Section 5 and is helpful to develop the estimator of $r(u)$ we propose.

Remark 3. The penalty coefficient λ balances how small $\|\hat{L}(u)\|_*$ is and how well the estimator fits the data. When $\lambda = 0$, the trivial solution to (2.3) is $(\hat{\beta}(u), \hat{L}(u)) = (0, Y)$, provided that $Y \in \mathcal{L}$. This estimator fits the data perfectly but is inconsistent as long as the true coefficients $\beta_0(u) \neq 0$. On the other hand, when λ is infinity, $\hat{L}(u)$ would be 0 to set the nuclear norm penalty equal to 0. This could again be implausible because it is equivalent to ignoring $L_0(u)$ and estimating $\beta_0(u)$ simply by pooled quantile regression. As a result, the estimator $\hat{\beta}(u)$ would be inconsistent when the covariates are correlated with the fixed effects. In the next section, we will be precise about the appropriate order of λ that guarantees uniform consistency of the estimator.

3. PRELIMINARY RESULTS

In this section, we first introduce the setups and summarize the main results of the paper. We then show that the estimator lies in a restricted set with probability approaching one (w.p.a.1), so that all the subsequent analysis can be conducted within that set.

3.1. Preliminaries and Preview of the Main Results

We take the fixed effect approach by treating all the (realized) individual and time fixed effects, and thus $L_0(u)$, for all $u \in \mathcal{U}$, as parameters to be identified and estimated (Moon and Weidner, 2015). Recall that $W_L = (\{F_k(u)\}_{k=1,\ldots,\bar{r}, u\in\mathcal{U}}, \{\Lambda_k(u)\}_{k=1,\ldots,\bar{r}, u\in\mathcal{U}})$. All the subsequent analysis and stochastic statements, including (conditional) probabilities and expectations, are implicitly conditional on (W_L, Ω_L), where Ω_L is the following event:

$$\|L_0(u)\|_\infty \leq \alpha_{NT}, \forall u \in \mathcal{U} \text{ and} \tag{3.1}$$

$$\frac{1}{\sqrt{NT}}\|L_0(u') - L_0(u)\|_F \leq \zeta_L |u' - u|, \forall u, u' \in \mathcal{U} \tag{3.2}$$

for some constant $\zeta_L > 0$. For the ease of notation, we omit conditioning on (W_L, Ω_L) henceforth. Throughout the paper, we maintain the assumption that $\mathbb{P}(\Omega_L) \to 1$. All the error bounds and uniform consistency also hold unconditionally under this assumption since all the probability bounds obtained in this paper are independent of the realization of $L_0(u)$.

Remark 4. We allow α_{NT} to grow with N and T. If α_{NT} has order $\log(NT)$, $L_0(u)$ in all the models in Examples 1–3 satisfies equation (3.1) w.p.a.1 if all the individual and time fixed effects are sub-Gaussian. Note that this *does not* rule out the case where Y_{it} itself has a heavy tail.

Remark 5. Equation (3.2) is needed to prove uniform consistency. One can verify that $L_0(\cdot)$ in Example 1 satisfies equation (3.2) if the conditional quantile function of ϵ, $q_{\epsilon_i|\Lambda^o}(\cdot)$, is Lipschitz continuous almost surely with a Lipschitz constant uniform in i. For Example 2, $L_0(\cdot)$ satisfies condition (3.2) w.p.a.1 if $q_\epsilon(\cdot)$ is Lipschitz continuous and if $\sum_{i,t}(\sum_{m=1}^{\bar{r}_2} F^b_{tm}\Lambda^b_{im})^2/NT$ converges in probability to a constant. Note that condition (3.2) rules out *some* cases where the set of the effective fixed effects changes on $\mathcal{U}$. To see this, suppose in Example 3, there exists a jumping point $u_0 \in \mathcal{U}$ such that $r(u) < r(u')$ for any $u < u_0 \leq u'$. Suppose the individual and time fixed effects do not depend on u and the first $r(u)$ of them at u' are the same as those at u. Then

$$\frac{1}{\sqrt{NT}}\|L_0(u') - L_0(u)\|_F = \sqrt{\frac{1}{NT}\sum_{i,t}\left(\sum_{k=r(u)+1}^{r(u')} F_{tk}\Lambda_{ik}\right)^2},$$

which may converge in probability to a positive constant if the law of large numbers holds for it. Nevertheless, this assumption is not restrictive even in this situation when there are only a finite number of such jumping points in $(0,1)$: Let $\{u_{0,k} : k = 1,2,\ldots,K\}(K < \infty)$ be the set of such jumping points with $u_{0,k} < u_{0,k+1}$, for all $k = 1,\ldots,K-1$. If equation (3.2) holds (w.p.a.1) for compact interval $\mathcal{U}_k \subset (u_{0,k}, u_{0,k+1})$ for each k, we can then establish uniform error bound over each $\mathcal{U}_k$ and uniform bound over $\bigcup_{k=1}^K \mathcal{U}_k$ is immediately obtained.

Under some other assumptions, this paper shows that (i) $(\beta_0(u), L_0(u))$ are identified for all $u \in \mathcal{U}$ over *a restricted set*, a subset of the parameter space where the estimator $(\hat{\beta}(u), \hat{L}(u))$ defined by (2.3) lies w.p.a.1, and (ii) for some constant $C_{error} > 0$, the estimator satisfies the following inequality w.p.a.1:

$$\sup_{u\in\mathcal{U}}\left(\|\hat{\beta}(u) - \beta_0(u)\|_F^2 + \frac{1}{NT}\|\hat{L}(u) - L_0(u)\|_F^2\right) \leq \gamma^2, \tag{3.3}$$

where

$$\gamma = C_{error}\alpha_{NT}^2\left((1+C_\lambda) \vee \sqrt{\log(NT)}\right)\left(\sqrt{\frac{p\log((p+1)NT)}{NT}} \vee \sqrt{\frac{\bar{r}}{N \wedge T}}\right), \tag{3.4}$$

where C_λ is some positive number or sequence of (N,T) that will be introduced in Section 5 and both $\alpha_{NT} \geq 1$ and $C_\lambda > 0$ are allowed to grow to infinity with N and T. The error bound implies uniform consistency of the estimator given a fixed $\bar{r}$ and a fixed or slowly growing p, α_{NT}, and C_λ. Based on this result, we also propose a consistent estimator of the number of the effective interactive fixed effects $r(u)$, for each $u \in \mathcal{U}$.

Now, we derive the restricted set for identification. Properties of the set are crucial to the derivation of the uniform error bound as well.

3.2. The Restricted Set

We first exploit implications from the definition of the estimator to show that the estimator lies in a restricted set that is smaller than the full parameter space. All the subsequent analysis can be conducted in it. Recall that $W_X := (X_1, \ldots, X_p)$. We make the following assumption. Note that the statements are also implicitly conditional on W_L as mentioned earlier.

Assumption 1. (i) Let $V(u) := Y - \sum_{j=1}^{p} X_j \beta_{0,j}(u) - L_0(u)$. For all $u \in \mathcal{U}$, entries in matrix $V(u)$ are independent conditional on W_X and strictly decreasing in u almost surely. (ii) There exists a constant $C_X > 0$ such that $\max_{1 \le j \le p} \|X_j\|_F^2 \le C_X NT$ w.p.a.1.

Entries in $V(u)$ being strictly decreasing in u almost surely implies that for almost all realizations of the $X_{j,it}$s, Y_{it}s are continuously distributed. The independence requirement in part (i) is for simplicity so that some inequalities for random matrices can be easily applied in the proof. The same assumption when $\mathcal{U}$ is a singleton can be found in Ando and Bai (2020) and Chen et al. (2021) as well. Moderate serial correlation in $V_{it}(u)$ can be allowed at a cost of more technical conditions. On the other hand, certain serial correlation in the covariates is allowed, and p is allowed to grow with N and T. For instance, one can verify that Assumption 1(ii) holds if $\max_{j=1,\ldots,p;\, i,t} \mathbb{E}(X_{j,it}^4) < c$ for some $c > 0$ with $X_{j,it}$ independent across i for all j and $p = o(N)$. This rules out some nonstationary regressors with increasing moments such as a random walk $X_{j,it} \sim N(0,t)$.[3]

Under Assumption 1, we can show that the estimation error $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) := (\hat{\beta}(u) - \beta_0(u), \hat{L}(u) - L_0(u))$ lies in a cone uniformly in $u \in \mathcal{U}$ w.p.a.1. The cone has useful properties for later use. To characterize the cone, let us introduce some notations. Let $R(u)\Sigma(u)S(u)'$ be a singular value decomposition of $L_0(u)$. Following Candès and Recht (2009), let $\Phi(u) := \{M \in \mathbb{R}^{N\times T} : \exists A \in \mathbb{R}^{r(u)\times T}$ and $B \in \mathbb{R}^{N\times r(u)}$ $s.t.$ $M = R(u)A + BS(u)'\}$. Denote the orthogonal projection of an arbitrary $N \times T$ matrix W onto this space and its orthogonal complement by $\mathcal{P}_{\Phi(u)} W$ and by $\mathcal{P}_{\Phi(u)^\perp} W$, respectively, then

$$\mathcal{P}_{\Phi(u)} W = R(u)R(u)'W + WS(u)S(u)' - R(u)R(u)'WS(u)S(u)', \tag{3.5}$$

$$\mathcal{P}_{\Phi(u)^\perp} W = \left(I_N - R(u)R(u)'\right) W \left(I_T - S(u)S(u)'\right), \tag{3.6}$$

where I_N and I_T are the $N \times N$ and $T \times T$ identity matrices, respectively.

Remark 6. Let the $T \times r(u)$ matrix $F_0(u)$ and the $N \times r(u)$ matrix $\Lambda_0(u)$ be the true effective fixed effects at u. Define the orthogonal projection

[3] We thank an anonymous referee for suggesting this example.

matrices as $M_{\Lambda_0}(u) := I_N - \Lambda_0(u)[\Lambda_0(u)'\Lambda_0(u)]^{-1}\Lambda_0(u)'$ and $M_{F_0}(u) = I_T - F_0(u)[F_0(u)'F_0(u)]^{-1}F_0(u)'$. One can verify that $\mathcal{P}_{\Phi(u)^\perp}W = M_{\Lambda_0}(u)WM_{F_0}(u)$ and $\mathcal{P}_{\Phi(u)}W = W - M_{\Lambda_0}(u)WM_{F_0}(u)$. Therefore, we can interpret $\mathcal{P}_{\Phi(u)^\perp}W$ as the part of W that can be (linearly) explained *neither* by $F_0(u)$ *nor* by $\Lambda_0(u)$, while $\mathcal{P}_{\Phi(u)}W$ as the part of W that can be explained *either* by $F_0(u)$ *or* by $\Lambda_0(u)$.

LEMMA 1. *Under Assumption 1, for C_X defined in Assumption 1, there exists a constant $C_{op} > 0$ such that for $\lambda = (1+C_\lambda)C_{op}\sqrt{N \vee T}/NT$ where $C_\lambda > 0$ and can be either a constant or can grow to infinity, by letting*

$$\kappa_1(\lambda) := 1 + 2/C_\lambda \text{ and } \kappa_2(\lambda) := 5\sqrt{2C_X p(N \wedge T)\log((p+1)NT)}/(C_\lambda C_{op}), \tag{3.7}$$

we have

$$\sup_{u\in\mathcal{U}}\left(\left\|\mathcal{P}_{\Phi(u)^\perp}\hat{\Delta}_L(u)\right\|_* - \kappa_1(\lambda)\left\|\mathcal{P}_{\Phi(u)}\hat{\Delta}_L(u)\right\|_* - \kappa_2(\lambda)\left\|\hat{\Delta}_\beta(u)\right\|_F\right) \le 0, w.p.a.1. \tag{3.8}$$

Proof. See Section S.B.1 of the Supplementary Material. □

By Lemma 1, define cone $\mathcal{R}_u$ by

$$\mathcal{R}_u := \left\{(\Delta_\beta, \Delta_L) \in \mathbb{R}^p \times \mathbb{R}^{N\times T} : \left\|\mathcal{P}_{\Phi(u)^\perp}\Delta_L\right\|_* \le \kappa_1(\lambda)\left\|\mathcal{P}_{\Phi(u)}\Delta_L\right\|_* + \kappa_2(\lambda)\left\|\Delta_\beta\right\|_F\right\}.$$

Let $\mathcal{D} := \mathbb{R}^p \times \{\Delta_L \in \mathbb{R}^{N\times T} : \|\Delta_L\|_\infty \le 2\alpha_{NT}\}$. Recall that $\mathcal{L} = \{L \in \mathbb{R}^{N\times T} : \|L\|_\infty \le \alpha_{NT}\}$. Under Ω_L, by $\hat{L}(u) \in \mathcal{L}$ and by Lemma 1, we have $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u)) \in \mathcal{R}_u \cap \mathcal{D}$ for all $u \in \mathcal{U}$ w.p.a.1. Let $\Theta_0 := \{(\beta(u), L(u)) \in \mathbb{R}^p \times \mathcal{L} : (\beta(u) - \beta_0(u), L(u) - L_0(u)) \in \mathcal{R}_u \cap \mathcal{D}\}$. Therefore, $(\hat{\beta}(u), \hat{L}(u))$ lies in the restricted set Θ_0 for all $u \in \mathcal{U}$ w.p.a.1.

The key property of $\mathcal{R}_u$ is that for any element $(\Delta_\beta, \Delta_L) \in \mathcal{R}_u$ and for any $u \in \mathcal{U}$, $\|\Delta_L\|_*$ and $\|\Delta_L\|_F$ can be of the same order, a property that low-rank matrices also share[4]. To see why this is true, note that by equation (3.5), the rank of $\mathcal{P}_{\Phi(u)}\Delta_L$ is at most $3r(u)$ for all u. Hence, for all $u \in \mathcal{U}$,

$$\|\mathcal{P}_{\Phi(u)}\Delta_L\|_* \le \sqrt{3r(u)}\|\mathcal{P}_{\Phi(u)}\Delta_L\|_F \le \sqrt{3r(u)}\|\Delta_L\|_F \le \sqrt{3r(u)}\|\Delta_L\|_*,$$

where the first and the last inequalities are by the relationship between the nuclear norm and the Frobenius norm. The second inequality is due to $\left\langle\mathcal{P}_{\Phi(u)}\Delta_L, \mathcal{P}_{\Phi(u)^\perp}\Delta_L\right\rangle = 0$ and by the Pythagoras formula. As a consequence, supposing, for instance, $C_\lambda = 1$ and thus $\lambda = 2C_{op}\sqrt{N \vee T}/NT$, elements in $\mathcal{R}_u$ then satisfy

$$\frac{1}{4\sqrt{3\bar{r}}}\left(\|\Delta_L\|_* - 5\sqrt{2C_X p(N \wedge T)\log((p+1)NT)}\|\Delta_\beta\|_F/C_{op}\right) \le \|\Delta_L\|_F \le \|\Delta_L\|_*, \tag{3.9}$$

[4] Note that this property is nontrivial because in general the nuclear norm $\|\Delta_L\|_*$ can be as large as $\sqrt{N \wedge T}\|\Delta_L\|_F$.

which implies that if $\sqrt{p(N\wedge T)\log((p+1)NT)}\|\Delta_\beta\|_F/\|\Delta_L\|_* = o(1)$, $\|\Delta_L\|_*$ and $\|\Delta_L\|_F$ are of the same order. This property is useful both to show identification and to derive the uniform rate of convergence.

Remark 7. Note that since the estimation error $(\hat{\Delta}_\beta(u), \hat{\Delta}_L(u))$ is in $\mathcal{R}_u$ w.p.a.1, $\|\hat{\Delta}_L(u)\|_F$ and $\|\hat{\Delta}_L(u)\|_*$ can also be of the same order w.p.a.1. Let the singular values of $\hat{\Delta}_L(u)$ be $\hat{\sigma}_k(u), k = 1,\ldots,N\wedge T$. By the definition of the Frobenius and the nuclear norms, it thus says $\sum_k \hat{\sigma}_k^2(u)$ and $(\sum_k \hat{\sigma}_k(u))^2$ are of the same order w.p.a.1. Therefore, although it is unclear whether the nuclear norm penalty makes some singular values of $\hat{\Delta}_L(u)$ to be exact zero so that $\hat{\Delta}_L(u)$, and in turn, $\hat{L}(u)$, is low-rank, the singular values, in descending order, must decay in order. For instance, if the leading singular values have order $\sqrt{NT}$, then the number of such singular values must be $O(1)$ and the order of the remaining singular values must be smaller than it.

The cone $\mathcal{R}_u$ in Lemma 1 is similar to those obtained in the broad literature of nuclear norm penalized estimation under different objective functions (see, e.g., Agarwal et al., 2012; Negahban and Wainwright, 2012; Chernozhukov et al., 2019; Athey et al., 2021). When there are no regressors, these results essentially say the following: The estimation error, when orthogonally projected onto the space where the true low-rank component lies, dominates its projection onto the orthogonal complement in order. Different from the mentioned literature, we establish uniformity in Lemma 1 under a nonsmooth objective function.

4. IDENTIFICATION OVER THE RESTRICTED SET

In this section, we establish identification of $(\beta_0(u), L_0(u))$ for all $u \in \mathcal{U}$ over the restricted set Θ_0. Identification is in the sense that $(\beta_0(u), L_0(u))$ is uniquely determined over Θ_0 uniformly over $u \in \mathcal{U}$ by the joint distribution of $(Y, X_1, \ldots, X_p)$ implicitly conditional on the true fixed effects. This notion of identification is often adopted in the literature of panel data models with interactive fixed effects, for instance, Moon and Weidner (2015, 2019).

Identification on Θ_0, a subset of the whole parameter space $\mathbb{R}^p \times \mathcal{L}$, is sufficient for consistency (similar to Belloni and Chernozhukov, 2011). Under the event $(\hat{\beta}(u), \hat{L}(u)) \in \Theta_0$, which happens w.p.a.1 by Lemma 1, the estimator is equivalent to the one under the same objective function as (2.3) with the parameter space replaced by Θ_0. Consistency of the latter only requires identification on Θ_0, so identification on Θ_0 is sufficient for our estimator as well.

4.1. Assumptions

Assumption 2. There exists a $\delta > 0$ such that the conditional density $f_{V_{it}(u)|W_X}$ satisfies

$$\underline{f} := \inf_{\substack{s\in[-\delta,\delta],\, u\in\mathcal{U} \\ 1\le i\le N,\, 1\le t\le T}} f_{V_{it}(u)|W_X}(s) > 0 \; a.s.$$

Note that δ in Assumption 2 can be arbitrarily small. A sufficient condition for Assumption 2 to hold is that $f_{V_{it}(u)|W_X}(0) > 0$ uniformly in i,t and $u \in \mathcal{U}$ and the functions $\{f_{V_{it}(u)|W_X}\}_{i,t,u}$ are equicontinuous at 0 for all realizations of W_X. This assumption can be shown to be weaker than the assumptions on the conditional density in Ando and Bai (2020), Chen et al. (2021), and Belloni et al. (2023). See Section S.A.2 of the Supplementary Material for a detailed discussion. Allowing the conditional density to reach zero and/or to be nondifferentiable in some regions is useful in economic and financial applications when, for instance, the number of interactive fixed effects changes across quantile levels. For example, Ando and Bai (2020) find that the numbers of common factors for stock returns are different at the 0.05th and the 0.95th quantiles. Example 4 illustrates why a changing number of interactive fixed effects may lead to a zero or nondifferentiable density function. Moreover, not requiring the density to have a bounded derivative entertains some applications where the conditional density functions are wiggly.

Example 4. Let $Y_{it} = \Lambda_{1i}F_{1t} + \Lambda_{2i}F_{2t}\mathbb{1}(U_{it} > 0.5) + G(U_{it})$, where G is some strictly increasing and differentiable function. Suppose $\Lambda_{2i}F_{2t} > 0$. By definition, $V_{it}(u) = G(U_{it}) + \Lambda_{2i}F_{2t}\mathbb{1}(U_{it} > 0.5) - G(u) - \Lambda_{2i}F_{2t}\mathbb{1}(u > 0.5)$. Note that, for any $u \in (0,1)$, $V_{it}(u)$ is not supported on the interval $(G(0.5) - G(u) - \Lambda_{2i}F_{2t}\mathbb{1}(u > 0.5), G(0.5) + \Lambda_{2i}F_{2t} - G(u) - \Lambda_{2i}F_{2t}\mathbb{1}(u > 0.5)]$. This interval never contains 0 for all u, so our assumption is satisfied for all u except for $u = 0.5$. However, for any u, the density of $V_{it}(u)$ is equal to 0 in this interval and is thus not bounded away from zero on all compact intervals, and it is nondifferentiable at the boundaries of the interval.

Assumption 3. Parameter C_λ in Lemma 1 is $O\left(\sqrt{\log(NT)}\right)$. The α_{NT} in the parameter space $\mathcal{L}$ in equation (2.3) is no smaller than 1, can grow to infinity with N and T, and makes $1/\gamma \to \infty$ where γ is defined in equation (3.4). Besides, either one of the following two conditions holds: (i) $\max_{j=1,\ldots,p;\,i,t} |X_{j,it}| \leq \sqrt{C_X}$ *a.s.* where C_X is the same as in Assumption 1, and $p = o((N \wedge T)/(\log(NT)\alpha_{NT}^2))$. (ii) There exists a positive constant C_φ and an increasing continuous function $\varphi : [0,\infty) \to [0,\infty)$ such that $\varphi(x) \geq x^{2+\eta}$ for some $\eta > 0$ for all $x \geq 0$ and

$$\max_{j=1,\ldots,p;\,i,t} \mathbb{E}\varphi(|X_{j,it}|) < C_\varphi \text{ and } p \cdot \left(\frac{p}{\varphi\left(\frac{\sqrt{N \wedge T}}{\sqrt{(p+1)\log(NT)\alpha_{NT}}} \right)} \right)^{\frac{\eta}{\eta+2}} = o(1). \qquad \textbf{(4.1)}$$

Assumption 3 restricts $C_\lambda = O\left(\sqrt{\log(NT)}\right)$ so that the penalty is not unnecessarily large that slows down the rate of convergence (see γ in equation (3.4)). Assumption 3 describes two cases for the number and the tails of the regressors. Case (ii) allows for unbounded regressors whose moments that are at least slightly higher than the second-order exist. Note that a necessary condition for the second part in equation (4.1) to hold is still $p = o((N \wedge T)/(\log(NT)\alpha_{NT}^2))$, same as Case (i). The actual number of regressors allowed depends on how thin the tails of the regressors are; a large p is allowed under a large η. When φ is the exponential function, for instance, the $X_{j,it}$s are sub-Gaussian, one can pick a sufficiently large η so that p can be as large as $O((N\wedge T)^{1-\epsilon}/(\log(NT)\alpha_{NT}^2))$ for an arbitrarily small $\epsilon>0$. When $X_{j,it}$ only has, for instance, finite fourth moment, by letting $\eta=2$ and $\varphi(x)=x^4$, we have $p=o\left(\left[(N\wedge T)/(\alpha_{NT}^2\log(NT))\right]^{2/5}\right)=o(N)$, so Assumption 1(ii) holds if $X_{j,it}$ is independent across i for all j.

Assumption 4. (i) For any vector $\tau\in\mathbb{R}^p$, there exists a constant $C_{\Phi X}>1$ such that for some arbitrarily small $\varepsilon>0$,

$$\inf_{u\in\mathcal{U}}\left(\mathbb{E}\left\|\mathcal{P}_{\Phi(u)^\perp}\left(\sum_{j=1}^p X_j\tau_j\right)\right\|_F^2-(C_{\Phi X}+\varepsilon_0)^2\,3r(u)\kappa_1(\lambda)^2\mathbb{E}\left\|\mathcal{P}_{\Phi(u)}\left(\sum_{j=1}^p X_j\tau_j\right)\right\|_F^2\right)\geq 0. \tag{4.2}$$

(ii) The smallest eigenvalue of the $p\times p$ matrix $\mathbb{E}\left(X_{it}X_{it}'\right)$ is no smaller than some $\sigma_{min}^2>0$ for all i and t.

The purpose of Assumption 4 is as follows: We can show that Assumptions 2 and 3 imply that the expected difference in the objective function $\mathbb{E}[\boldsymbol{\rho}_u(V(u)-\sum_{j=1}^p X_j\Delta_{\beta,j}-\Delta_L)-\boldsymbol{\rho}_u(V(u))]$ can be lower bounded by a linear function of $\mathbb{E}\|\sum_{j=1}^p X_j\Delta_{\beta,j}+\Delta_L\|_F^2$. Assumption 4(i) further guarantees that, under some restrictions,

$$\mathbb{E}\left\|\sum_{j=1}^p X_j\Delta_{\beta,j}+\Delta_L\right\|_F^2\geq C_{RSC}\left(\mathbb{E}\left\|\sum_{j=1}^p X_j\Delta_{\beta,j}\right\|_F^2+\|\Delta_L\|_F^2\right) \tag{4.3}$$

for $(\Delta_L,\Delta_\beta)\in\mathcal{R}_u\cap\mathcal{D}$ for some $C_{RSC}>0$. Finally, the usual no-perfect multicollinearity condition Assumption 4(ii) lower bounds $\mathbb{E}\|\sum_{j=1}^p X_j\Delta_{\beta,j}\|_F^2$ by $\sigma_{min}^2 NT\|\Delta_\beta\|_F^2$.

Equation (4.3) is a version of the *restricted strong convexity* in the machine learning and statistics literature on low-rank matrix recovery or nuclear norm penalized estimation (e.g., Negahban and Wainwright, 2011, 2012; Agarwal et al., 2012; Negahban et al., 2012; Chernozhukov et al., 2019; Belloni et al., 2023; etc.). Our Assumption 4(i) provides a lower-level and more interpretable sufficient condition for it under our setup. It essentially says that for any linear combination of the regressors, the part that can not be explained by the true fixed effects must be sufficiently larger than the part that can be explained either by the true individual or the time fixed effects. Indeed, we can show that under Assumption 4(i), any linear combination of the regressors lies away from the cone where Δ_L lies in. Hence, it is not possible for $\mathbb{E}\|\sum_{j=1}^p X_j\Delta_{\beta,j}+\Delta_L\|_F^2$ to be zero unless both Δ_β and Δ_L are both zero. See Lemma S.B.3 in the Supplementary Material for details.

Remark 8. The bound on the smallest eigenvalue σ_{min}^2 is assumed to be a constant for simplicity. This is the case when p is fixed or p is increasing with

(N,T) but, for instance, $\mathbb{E}(X_{it}X_{it}')$ is diagonal. See also the examples in Belloni and Chernozhukov (2011). In general, if p grows to infinity, σ_{min}^2 can diminish to zero. Since the error bound γ decreases in σ_{min}^2 (see the proof of Theorem 2 for details), the rate of convergence of our estimator will be slower in that case.

4.1.1. *On Low-Rank Regressors.* In the literature of interactive fixed effects where the individual and time fixed effects $(\Lambda_0(u), F_0(u))$ are treated as separate parameters, a commonly used identification condition is that all the regressors are high-rank (e.g., Bai, 2009; Moon and Weidner, 2015; see also Ahn, Lee, and Schmidt (2001) for an example of nonidentification when there are both a low-rank common component and a time-invariant regressor with time-varying coefficients).

Moon and Weidner (2019) find that nuclear norm penalized estimation allows for low-rank regressors[5] in mean regressions under an identification condition similar to Assumption 4. We show that this insight carries over to quantile regression.

To see it, let x be an $NT \times p$ matrix, where the jth column is the vectorized X_j. We show in Section S.B.2 of the Supplementary Material that Assumption 4(i) is *equivalent* to that for all $u \in \mathcal{U}$, the following holds:

$$\mathbb{E}\left(x'\left(M_{F_0}(u) \otimes M_{\Lambda_0}(u)\right)x - \frac{(C_{\Phi X}+\varepsilon_0)^2 3r(u)\kappa_1(\lambda)^2}{1+(C_{\Phi X}+\varepsilon_0)^2 3r(u)\kappa_1(\lambda)^2}x'x\right) \text{ is positive semidefinite.} \tag{4.4}$$

Now, we can compare equation (4.4) with an identification condition that only allows for high-rank regressors, Assumption ID in Moon and Weidner (2015). Assumption ID assumed that $\mathbb{E}[x'\left(M_F \otimes M_{\Lambda_0}\right)x]$ is positive definite *for all* $F \in \mathbb{R}^{T\times r}$ for some $r \geq r(u)$. Since the requirement is imposed on all F instead of only on the true time fixed effects F_0, it requires that the regressors still have variation after projecting onto the *true* individual fixed effects and *arbitrary* time fixed effects. Therefore, low-rank regressors are not allowed. In contrast, in our assumption, $\mathbb{E}[x'\left(M_{F_0}(u) \otimes M_{\Lambda_0}(u)\right)x]$ only looks at projections onto the *true* time and individual fixed effects. Therefore, low-rank regressors can be allowed as long as they cannot be fully explained by the true fixed effects, for instance, a regressor that also has a factor structure and the factors and factor loadings are orthogonal to the true fixed effects. On the other hand, we require that after projecting out the true fixed effects, the regressors still have *sufficiently large* variation while the condition in Moon and Weidner (2015) only requires that there is variation remaining. Therefore, Assumption 4(i) is neither stronger nor weaker than the

[5] Here, a low-rank regressor refers to the situation where the $N \times T$ data matrix of this regressor is low-rank. Across regressors, perfect multicollinearity is not allowed so full rank of the variance matrix $\mathbb{E}X_{it}X_{it}'$ is still required by Assumption 4(ii).

identification condition in the literature where the interactive fixed effects are estimated separately.

4.2. Identification

THEOREM 1. *Under Assumptions 1–4, there exists a constant $C_{RSC} > 0$ such that for all fixed and sufficiently large N and T,*

$$\inf_{\substack{u\in\mathcal{U}\\(\Delta_\beta,\Delta_L)\in\mathcal{D}\cap\mathcal{R}_u}}\left(\mathbb{E}\left[\boldsymbol{\rho}_u\left(V(u)-\sum_{j=1}^{p}X_j\Delta_\beta-\Delta_L\right)-\boldsymbol{\rho}_u\left(V(u)\right)\right]\right.$$
$$\left.-\frac{(1\wedge\delta)^2\underline{f}C_{RSC}\left(NT\|\Delta_\beta\|_F^2+\|\Delta_L\|_F^2\right)}{4\left(2\alpha_{NT}+(\alpha_{NT}\|\Delta_\beta\|_F/\gamma\vee 1)\right)^2}\right)\geq 0. \qquad \textbf{(4.5)}$$

Therefore, $(\beta_0(u), L_0(u))$ are identified over Θ_0 for all $u\in\mathcal{U}$. The number of interactive fixed effects $r(u)$ is also identified for all $u\in\mathcal{U}$. Constant C_{RSC} increases in $C_{\Phi X}$ and C_λ but is upper bounded. Its exact form is in the proof.

Proof. See a proof sketch and the formal proof in Section S.B.2 of the Supplementary Material. □

Theorem 1 proves identification of the true parameters over the restricted set Θ_0. Identification in this subset is sufficient for the purpose of consistent estimation because we have shown in Lemma 1 that the estimator falls into it w.p.a.1. Once $L_0(u)$ and thus $r(u)$ are identified, the fixed effects $F_0(u)$ and $\Lambda_0(u)$ are also identified up to rotation or normalization. For instance, Bai (2009) shows that $F_0(u)$ and $\Lambda_0(u)$ are uniquely determined (up to a columnwise sign change) given $L_0(u)$ under the normalization that $F_0(u)'F_0(u)/T = I_r$ and $\Lambda_0(u)'\Lambda_0(u)$ is diagonal, where I_r is the $r\times r$ identity matrix.

Two remarks are in order.

Remark 9. The theorem requires sufficiently large N and T. This will be satisfied automatically when proving consistency as we will then send N and T to infinity.

Remark 10. The constant C_{RSC} increases in $C_{\Phi X}$ because the latter controls the distance from $(\Delta_\beta, \sum_{j=1}^{p}X_j\Delta_{\beta,j})$ to cone $\mathcal{R}_u$. When the regressors lie further away from the cone, we obtain a larger lower bound. Meanwhile, C_{RSC} also increases in C_λ because a larger C_λ leads to smaller $\kappa_1(\lambda)$ and $\kappa_2(\lambda)$ defined in equation (3.7), and the cone then becomes smaller and is thus easier to be separated from $(\Delta_\beta, \sum_{j=1}^{p}X_j\Delta_{\beta,j})$.

Before we end this section, we discuss the major theoretical difficulty in proving Theorem 1. In the proof, we first lower bound the expectation in (4.5) by a linear function of $\mathbb{E}\|\sum_{j=1}^{p}X_j\Delta_{\beta,j}+\Delta_L\|_F^2$, then invoke Assumption 4 to further bound it by the lower bound in the theorem. The difficulty in the first step arises from the high-dimensionality of Δ_L. For illustration, consider a simple case without covariates. The expectation under consideration can then be simplified as $\mathbb{E}\left[\boldsymbol{\rho}_u(V(u)-\Delta_L)-\boldsymbol{\rho}_u(V(u))\right]$. By Knight's identity (Knight, 1998) and by the definition of $V_{it}(u)$, it can be rewritten as

$$\sum_{i,t}\int_0^{\Delta_{L,it}}\left(F_{V_{it}(u)}(s)-F_{V_{it}(u)}(0)\right)ds, \tag{4.6}$$

where $F_{V_{it}(u)}$ is the cumulative distribution function of $V_{it}(u)$. The key problem is as follows. The magnitude of some $\Delta_{L,it}$s, even under our constraint $\|\Delta_{L,it}\|_\infty \leq 2\alpha_{NT}$ and even when $\|\Delta_L\|_F^2/NT \to 0$, can be as large as $2\alpha_{NT}$ which is allowed to grow to infinity. Hence, if one adopts the standard argument in quantile regression to first-order Taylor expand $F_{V_{it}(u)}(s)$ around 0 (supposing $F_{V_{it}(u)}(\cdot)$ is differentiable on $(0,s)$), we get $F_{V_{it}(u)}(s)-F_{V_{it}(u)}(0)=sf_{V_{it}(u)}(\tilde{s}(s))$, where $f_{V_{it}(u)}$ is the density function of $V_{it}(u)$ and the mean value $\tilde{s}(s)$ lies between 0 and $\Delta_{L,it}$. Now that $\tilde{s}(s)$ can be also as large as $2\alpha_{NT}, f_{V_{it}(u)}(\tilde{s}(s))$ may approach to zero as $2\alpha_{NT}$ grows if one only assumes the density is continuous at zero and $f_{V_{it}(u)}(0)>0$. So its infimum could be 0 and a strictly positive lower bound does not obtain.

We resolve the issue by restricting $\|\Delta_L\|_\infty \leq 2\alpha_{NT}$, achieved by adopting a compact parameter space $\mathcal{L}$ for the matrix component as in equation (2.3). Then as we show in Section S.B.4 of the Supplementary Material that each integral in the summation (4.6) is decreasing in the absolute upper limit, we can lower bound the integral by $\int_0^{(1\wedge\delta)\Delta_{L,it}/2\alpha_{NT}}\left(F_{V_{it}(u)}(s)-F_{V_{it}(u)}(0)\right)ds$, where δ is the constant in Assumption 2. For this integral, even if $\Delta_{L,it}$ is diverging, $|\Delta_{L,it}|/2\alpha_{NT}\leq 1$ so $(1\wedge\delta)\Delta_{L,it}/2\alpha_{NT}$ must lie in the region where the conditional density is positive. Then, first-order Taylor expanding $F_{V_{it}(u)}(s)$ around 0 yields a positive quadratic lower bound.

We also develop an alternative approach (see Section S.A.1 of the Supplementary Material) which relaxes the parameter space $\mathcal{L}$ to be $\mathbb{R}^{N\times T}$ at the cost of more technical conditions. We provide a detailed comparison in Section S.A.2 of the Supplementary Material to compare the two approaches with the literature.

5. ASYMPTOTIC RESULTS

In this section, we derive the uniform rate of convergence of our estimator.

Assumption 5. There exists a nonnegative $\zeta_X = O(\sqrt{p})$ such that

$$\|\beta_0(u')-\beta_0(u)\|_F \leq \zeta_X|u'-u|, \forall u,u'\in\mathcal{U}. \tag{5.1}$$

This assumption is the same as in Belloni and Chernozhukov (2011), and is made to obtain uniformity; it trivially holds when $\mathcal{U}$ is a singleton. The coefficients in Example 1 automatically satisfy Assumption 5 with $\zeta_X=0$, whereas in Example 2, the assumption holds if the quantile function $q_\epsilon(\cdot)$ is Lipschitz continuous on $\mathcal{U}$. Note that since $\beta_0(u)$ contains p components, we need to allow ζ_X to have order $\sqrt{p}$.

THEOREM 2. *Under Assumptions 1–5 and the condition for* λ *in Lemma 1, there exists a constant* $C_{error} > 0$ *such that the following holds w.p.a.1:*

$$\sup_{u\in\mathcal{U}} \|\hat{\beta}(u)-\beta_0(u)\|_F^2+\frac{1}{NT}\|\hat{L}(u)-L_0(u)\|_F^2 \leq \gamma^2$$
$$:= C_{error}^2\alpha_{NT}^4\left[(1+C_\lambda)^2\vee\log(NT)\right]\left(\frac{p\log((p+1)NT)}{NT}\vee\frac{\bar{r}}{N\wedge T}\right). \tag{5.2}$$

The constant C_{error} *decreases in* σ_{min}^2, $\underline{f}$, $C_{\Phi X}$, *and* C_λ, *and is lower bounded by a positive constant. Its exact formula is in the proof.*

Proof. See Section S.B.3 of the Supplementary Material. □

Theorem 2 implies uniform consistency of our estimator of $\beta_0(u)$ and $L_0(u)$ over $\mathcal{U}$ for a fixed $\bar{r}$ and slowly growing α_{NT} and p. The key determinants of the rate of convergence are $p\log((p+1)NT)/NT$ and $1/(N\wedge T)$. This part is similar to the nuclear norm penalized mean regression literature (Chernozhukov et al., 2019; Moon and Weidner, 2019; Athey et al., 2021). We will discuss this part in further detail in this section. Other determinants of the error bound include the number of interactive fixed effects ($\bar{r}$), the size of $\mathcal{R}_u\cap\mathcal{D}$ (i.e., the magnitude of α_{NT} and C_λ), and the strength of identification ($\underline{f}$, σ_{min}^2 and $C_{\Phi X}$). The way that these parameters affect the error bound is expected: A larger $\bar{r}$ results in a relatively higher-rank common component, making the estimation problem more difficult. Stronger identification (a larger $C_{\Phi X}$, a larger σ_{min}^2 and/or a larger $\underline{f}$) and a smaller relevant parameter space (a smaller α_{NT} and a larger C_λ) makes it easier to separate the estimation errors $\hat{\Delta}_{\beta,j}(u)$s and $\hat{\Delta}_L(u)$.

The error bound also implies that the optimal rate for C_λ is $\sqrt{\log(NT)}$. This is because it will then make $(1+C_\lambda)^2$ and $\log(NT)$ equal in order while minimizing C_{error} compared to a constant C_λ because C_{error} decreases as C_λ increases. With this C_λ, λ then has order $\sqrt{\log(NT)(N\vee T)}/NT$.

5.1. On the Rate of $\hat{L}(u)$ and a Rank Estimator

Under our restriction on the order of p in Assumption 3, $p\log((p+1)NT)/(NT) = o(1/(N\wedge T))$. Then, if $\alpha_{NT} = O(\log(NT))$ and $C_\lambda = O(\sqrt{\log(NT)})$, the error bound on $\hat{L}(u)$ in Theorem 2 is nearly optimal implied by Agarwal et al. (2012) in the following sense. Agarwal et al. (2012) study a model where an observable $N\times T$ matrix Y is the sum of a low-rank matrix, a sparse matrix and a noise matrix of i.i.d. Gaussian entries. To apply their result to our model, consider a special case, where $u = 0.5$ and $Y = L_0 + V$, where V is an $N\times T$ matrix of i.i.d. $N(0,v^2)$ entries. This model both satisfies the conditional quantile model (2.2) studied in this paper at $u = 0.5$ with $\beta(0.5) = 0$ and $q_{Y|L_0}(0.5) = L_0$, and also satisfies their setup with the sparse component being exactly zero. Their Theorem 2 (p. 1195) shows that the lower bound on the minimax risk in the squared Frobenius norm over the family

$\{L_0 : \text{rank}(L_0) \le \bar{r}, \|L_0\|_\infty \le \alpha_{NT}\}$ has the order $1/(N \wedge T)$. The order of this lower bound and our upper bound are equal up to a factor of $\alpha_{NT}^4 \log(NT)$.

From the error bound on $\hat{L}(u)$, we can obtain the order of the singular values of the estimation error $\hat{\Delta}_L(u)$ by Weyl's theorem. Let $\sigma_1(u) \ge \cdots \ge \sigma_{r(u)}(u) > 0$ be the nonzero singular values of $L_0(u)$, and $\hat{\sigma}_1(u) \ge \cdots \ge \hat{\sigma}_{N\wedge T}(u)$ be the singular values of $\hat{L}(u)$. We have the following corollary.

COROLLARY 1. *Under the conditions in Theorem 2, the following holds w.p.a.1:*

$$\sup_{u\in\mathcal{U}} \left\{ \max \left\{ |\hat{\sigma}_1(u) - \sigma_1(u)|, \ldots, |\hat{\sigma}_{r(u)}(u) - \sigma_{r(u)}(u)|, \hat{\sigma}_{r(u)+1}(u), \ldots, \hat{\sigma}_{N\wedge T}(u) \right\} \right\}$$
$$\le \sqrt{NT}\gamma := C_{error}\alpha_{NT}^2 \left[(1 + C_\lambda) \vee \sqrt{\log(NT)} \right] \sqrt{(p\log((p+1)NT)) \vee (\bar{r}(N \vee T))}, \tag{5.3}$$

where C_{error} is the same as in Theorem 2.

Proof. See Section S.B.3 of the Supplementary Material. □

Note that the nonzero singular values of $L_0(u)$ has order $\sqrt{NT}$ if $L_0(u)$ is formed by strong factors and factor loadings or if elements in $L_0(u)$ are $O(1)$. Although Theorem 2 and Corollary 1 are silent about whether the estimated low-rank component $\hat{L}(u)$ is low-rank or not, Corollary 1 says that for large enough N and T, since $\gamma = o(1)$, there does exist a large gap between the largest $r(u)$ and the remaining $((N \wedge T) - r(u))$ singular values of $\hat{L}(u)$. Specifically, the first $r(u)$ singular values of $\hat{L}(u)$ are of order $\sqrt{NT}$, whereas the other singular values have order $\sqrt{NT}\gamma$. This confirms the intuition in Remark 2.

This implication naturally leads to an estimator of $r(u)$. Let $\hat{r}(u) = \sum_k \mathbb{1}(\hat{\sigma}_k(u) \ge C_r)$ for an (N,T)-dependent C_r such that $\sqrt{NT}\gamma = o(C_r)$ and $C_r = o(\sqrt{NT})$. The following corollary establishes consistency of this estimator.

COROLLARY 2. *Under the conditions in Theorem 2, for any $u \in \mathcal{U}$, suppose all the nonzero singular values of $L_0(u)$ are of order $\sqrt{NT}$, then $\mathbb{P}\left(\hat{r}(u) = r(u)\right) \to 1$.*

Proof. See Section S.B.3 of the Supplementary Material. □

Remark 11. Consistency can be shown to hold uniformly in $u \in \mathcal{U}$, i.e., $\mathbb{P}(\sup_{u\in\mathcal{U}} |\hat{r}(u) - r(u)| = 0) \to 1$, if the required order of the nonzero singular values of $L_0(u)$ holds uniformly in u as well.

The result of Theorem 2 does not imply entrywise convergence of $\hat{L}_0(u)$ to $L_0(u)$. Given consistency in the average Frobenius norm, one may obtain entry-wise consistency or consistency in the sup-norm if the entries in the estimation error matrix $\hat{\Delta}_L$ have similar order, or in other words, the entries are well spread out. Our proof strategy does not provide results on this aspect of the error matrix. Yet from the simulation results in Section 6, $\|\hat{L}(u) - L_0(u)\|_\infty$ does seem to converge to 0 in the experiment.

5.2. On the Rate of $\hat{\beta}(u)$

Theorem 2 implies uniform consistency of $\hat{\beta}(u)$. Suppose $\alpha_{NT} = O(\log(NT))$. When p is fixed, the rate of convergence is slower than $\sqrt{1/NT}$, which is the rate obtained in *mean regressions* when $r(u)$ is known,[6] and the interactive fixed effects are estimated separately (e.g., Bai, 2009). This is perhaps due to penalization. When p is allowed to grow with N and T, recall that Assumption 3 implies $p = o\left((N \wedge T)/(\log(NT)\alpha_{NT}^2)\right)$. So the rate of convergence can be nearly $\sqrt{p/NT}$ if N and T are of the same order.

To obtain a faster rate of convergence for $\hat{\beta}(u)$, one possibility is to use the penalized estimator $\hat{\beta}(u)$ as a first-step estimator; using it as an initialization with the rank estimator proposed in Section 5, one can adopt the iterative estimator (1.2) described in the Introduction without penalization. Although this is back to a nonconvex problem, the initial value is already lying in a small neighborhood of the true parameter by uniform consistency. A few rounds of iterations even before convergence is reached might correct the penalization bias and achieve a faster rate (see Chernozhukov et al., 2019; Moon and Weidner, 2019 for mean regressions). The benefits of adopting such a two-step procedure instead of a fully iterative approach are threefold. First, it provides a consistent initial guess of $\beta_0(u)$ that lies close to the true parameter. Second, as a by-product, the penalized estimation step also provides a rank estimator which is needed for the iterative approach. Third, from our Monte Carlos, we find that the penalized estimator computes very fast compared to the estimator (1.2) under a fully iterative procedure, i.e., iterating until convergence from some non-consistent initial guess. Hence, obtaining the penalized estimator first, using it as an initial guess, and iterating only a few rounds for problem (1.2) may gain overall computation efficiency compared to solving (1.2) by the fully iterative procedure. Exploring these conjectures is left for future research.

6. MONTE CARLO SIMULATIONS

In this section, we illustrate the finite sample performance of our estimator using Monte Carlo simulations.

6.1. Data Generating Process

We consider the following data generating process (DGP), which is a special case of Example 3 and is adapted from Ando and Bai (2020):

$$Y_{it} = \alpha + \sum_{j=1}^{p} X_{j,it}\beta_j(U_{it}) + \sum_{k=1}^{5} \mathbb{1}_k(U_{it})F_{kt}\Lambda_{ki}(U_{it}) + \varepsilon_{it},$$

[6] The same rate can be achieved when $r(u)$ is unknown but consistently estimated. See, for instance, footnote 5 in Bai (2003).

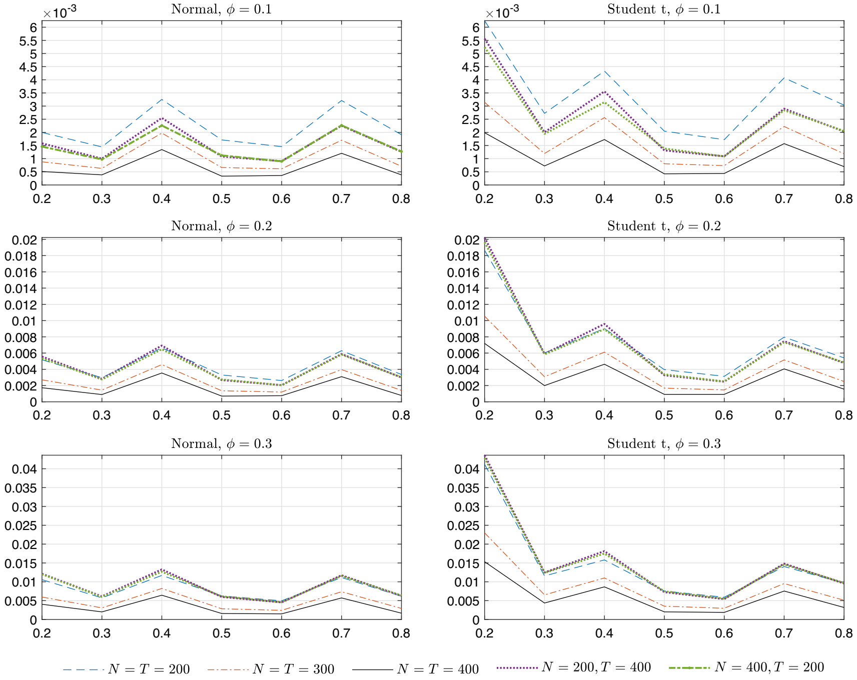

Figure 1. MSE$_\beta$.

where the U_{it}s are independently drawn from Unif[0, 1]. We set $\alpha = 2$, $p = 10$ and the coefficients satisfy

$$\beta_j(U_{it}) = \begin{cases} -1 + 0.1U_{it}, & \text{if } j \text{ is odd,} \\ 1 + 0.1U_{it}, & \text{if } j \text{ is even.} \end{cases}$$

The indicator functions $\mathbb{1}_k(\cdot) : (0,1) \mapsto \{0,1\}, k = 1, \ldots, 5$, satisfy

$$\mathbb{1}_1(u) = \mathbb{1}_2(u) = \mathbb{1}_3(u) = 1, \forall u \in (0,1), \quad \mathbb{1}_4(u) = \mathbb{1}(u > 0.35), \quad \mathbb{1}_5(u) = \mathbb{1}(u > 0.65).$$

We draw the time fixed effects $F_{1t}, \ldots, F_{5t}$ independently from Unif[0, 2]. We generate the individual fixed effects as $\Lambda_{ki}(U_{it}) = \chi_{ki} + 0.1U_{it}$, where the χ_{ki}s are independently drawn from Unif[0, 2] for $k = 1, \ldots, 5$.

For the covariates, we first draw $\eta_{j,it}, j = 1, \ldots, 10$, independently from Unif[0, 2] if j is odd and from Unif[0, 1] if j is even. The covariates are generated by

$$X_{j,it} = \begin{cases} \eta_{j,it} + \phi \cdot \left(F_{jt}^2 + \chi_{ji}^2\right), & j = 1, \ldots, 5, \\ \eta_{j,it} + 0.2 \cdot X_{j-5,it}, & j = 6, \ldots, 10. \end{cases}$$

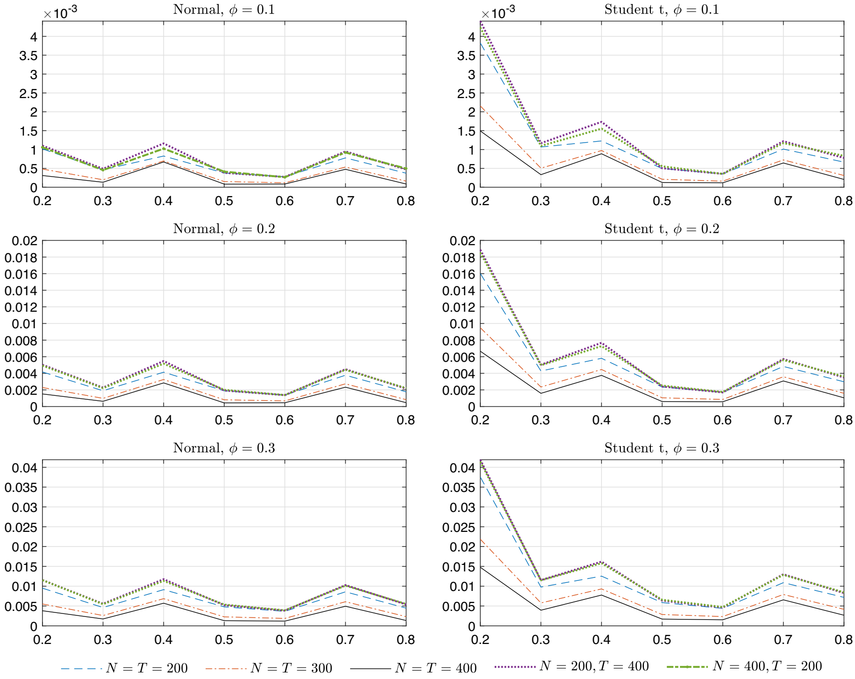

FIGURE 2. Bias^2_β.

The parameter $\phi \in \{0.1, 0.2, 0.3\}$ governs the correlation between the covariates and the fixed effects. Finally, ε_{it} is generated by $G^{-1}(U_{it})$, where G is the cumulative distribution function of either the standard normal distribution or Student's t-distribution with 2 degrees of freedom.

Since $X_{j,it}$ and F_{kt} are positive for all i, t, j, and k almost surely and $\beta_j(\cdot)$, $\Lambda_{ki}(\cdot)$, and $G^{-1}(\cdot)$ are strictly increasing (almost surely) for all i, j, and k, Y_{it} is strictly increasing in U_{it} almost surely. Let $\mathbf{1}_{N\times T}$ be an $N \times T$ matrix of all ones. The uth conditional quantile of Y is thus $q_{Y|W_X}(u) = \sum_{j=1}^{10} X_j \beta_j(u) + L_0(u)$, where

$$L_0(u) = \begin{cases} \left(\alpha + G^{-1}(u)\right)\mathbf{1}_{N\times T} + \sum_{k=1}^{3} \Lambda_k(u) F_k', & \text{if } u \le 0.35, \\ \left(\alpha + G^{-1}(u)\right)\mathbf{1}_{N\times T} + \sum_{k=1}^{4} \Lambda_k(u) F_k', & \text{if } 0.35 < u \le 0.65, \\ \left(\alpha + G^{-1}(u)\right)\mathbf{1}_{N\times T} + \sum_{k=1}^{5} \Lambda_k(u) F_k', & \text{if } u > 0.65. \end{cases}$$

From the model, one can see that the rank of $L_0(u)$ and the set of effective fixed effects vary in u. Also, the model allows the covariates to be correlated with $L_0(u)$. Higher correlation (greater ϕ) would make it more difficult to separately estimate $\beta(u)$s and the common component $L_0(u)$ since it tends to yield a smaller restricted strong convexity constant C_{RSC}. Finally, Y_{it} is allowed to have heavy tails because ε_{it} can be Student's t-distributed.

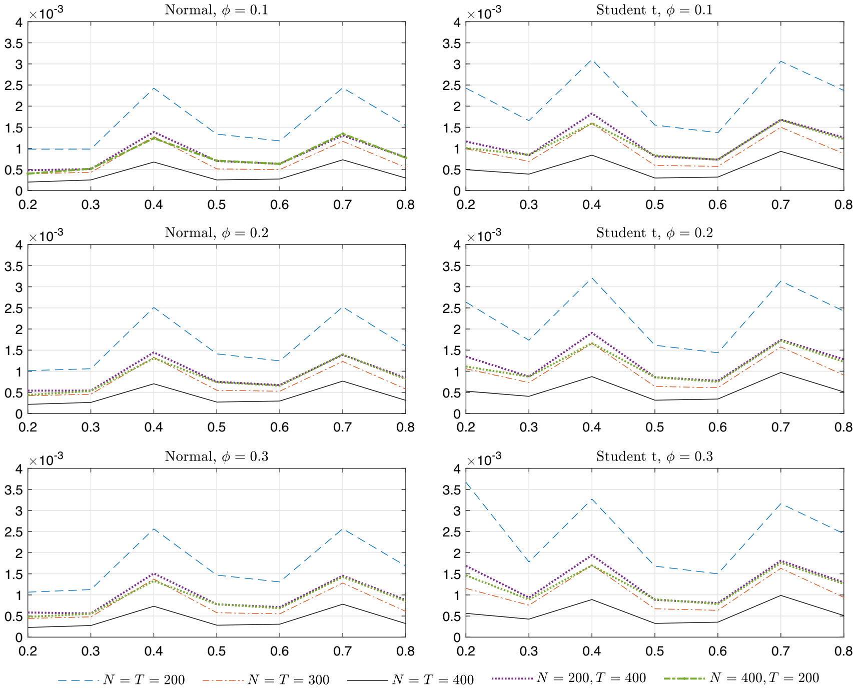

FIGURE 3. Var_β.

6.2. Evaluate the Performance

To illustrate the performance of the estimator, we conduct Monte Carlo simulations with various sample sizes for $\phi \in \{0.1, 0.2, 0.3\}$ and $u \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. For the sample size, $(N, T) \in \{(200, 200), (300, 300), (400, 400), (200, 400), (400, 200)\}$. The first three sample sizes with $N = T$ allow us to see the convergence of the estimator. The other two sample sizes with $N \neq T$ and $N \wedge T = 200$ allow us to test the theory which suggests that the rate of convergence does not depend on $N \vee T$ under a fixed p.

We use multiple measures to evaluate the estimator's performance. For the 10 coefficients $\beta_j(u)$s, we compute their average squared bias Bias^2_β, variance Var_β, and the mean squared error (MSE) MSE_β over 100 simulation replications. For the low-rank component, we compute the average squared Frobenius norm of the estimation error MSL_L and the maximum deviation MaxDev_L. Meanwhile, we also look at the MSE of the estimated conditional quantile function $\hat{q}_{Y|W_X}(u)$ (Ando and Bai, 2020).[7]

[7] Let $\hat{\beta}_{j,b}(u)$ and $\hat{L}_b(u)$ denote the estimator of $\beta_j(u)$ and $L_0(u)$ in the bth simulation replication. These measures are defined as follows: $\text{Bias}^2_\beta := \sum_{j=1}^{p}(\sum_{b=1}^{100}(\hat{\beta}_{j,b}(u) - \beta_j(u))/100)^2/p$. $\text{Var}_\beta := \sum_{j=1}^{p}[\sum_{b=1}^{100}\hat{\beta}_{j,b}(u)^2/$

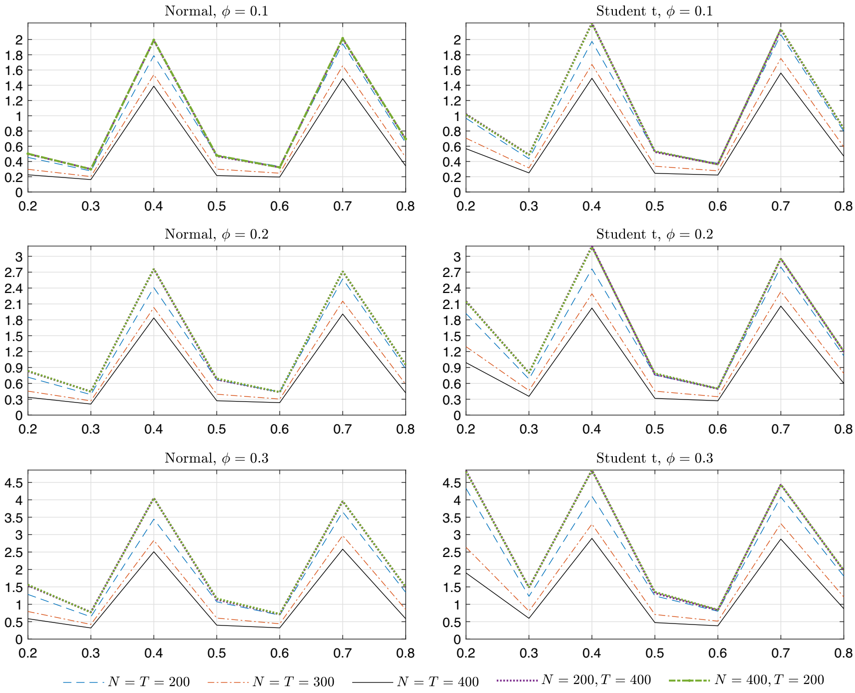

FIGURE 4. MSE_L.

Note that MSE_q can be small even if MSE_L is large; the latter is also affected by C_{RSC}, which in turn, is affected by the correlation between the covariates and $L_0(u)$ (determined by parameter ϕ). Hence, by comparing MSE_q with MSE_L across different values of ϕ, we can see how the strength of identification affects the results.

6.3. Implementation

We present the details about our algorithm in the Appendix A. The algorithm is adapted from the Augmented Lagrangian Multiplier method proposed in Lin, Chen, and Ma (2010), Candès et al. (2011), and Yuan and Yang (2013). The method was originally designed for $u = 0.5$ with no covariates. We extend it to accommodate any $u \in (0,1)$ with covariates. From the simulation results, the new algorithm works fast and well. For λ, here we simply set it to be $\lambda = \sqrt{\log(NT)(N \vee T)}/(4NT)$. This choice of λ corresponds to its optimal rate

$100 - (\sum_{b=1}^{100} \hat{\beta}_{j,b}(u)/100)^2]/p$. $\mathrm{MSE}_\beta = \mathrm{Bias}_\beta^2 + \mathrm{Var}_\beta$. $\mathrm{MSE}_L := \sum_{b=1}^{100}[\sum_{i,t}(\hat{L}_{it,b}(u) - L_{0,it}(u))^2/NT]/100$. $\mathrm{MSE}_q := \sum_{b=1}^{100}[\| \sum_{j=1}^{p} X_j(\hat{\beta}_{j,b}(u) - \beta_j(u)) + (\hat{L}_b(u) - L_0(u))\|_F^2/NT]/100$. $\mathrm{MaxDev}_L = \sum_{b=1}^{100} \|\hat{L}_b(u) - L_0(u)\|_\infty/100$.

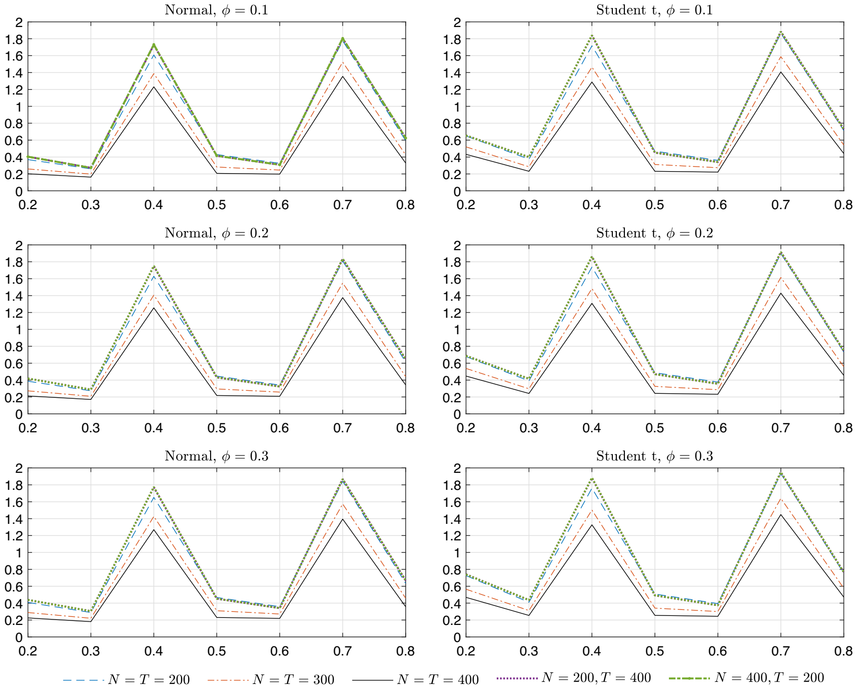

Figure 5. MSE_q.

derived in the theory. Alternatively, one may use cross-validation or adopt the BIC criterion proposed in Belloni et al. (2023) to select λ.

6.4. Results

Figures 1–6 present MSE_β, Bias^2_β, Var_β, MSE_L, MSE_q, and MaxDev_L under different ϕ, u, (N,T) and the distribution of ε_{it}. All experiments were performed in MATLAB 2019b under Windows 10 on a desktop computer with 10-core 2.8GHz Intel i9 processor and 16 GB RAM using parallel computing. From the results, we have the following key observations.

We can see that in all specifications, all these measures shrink toward zero as N and T *both* increase for all u considered. The spikes at $u = 0.4$ and $u = 0.7$ are by construction; due to the jump of the number of interactive fixed effects at quantile levels 0.35 and 0.65 in our DGP, the conditional density of $V_{it}(u)$ hits zero at these two quantiles, resulting in large MSE.

When only N or T increases, we can see that the MSE_L curves in Figures 4 for $(N = T = 200)$, $(N = 200, T = 400)$, and $(N = 400, T = 200)$ almost coincide, confirming our theory which says the rate of convergence of $\hat{L}(u)$ largely depends on the minimum between N and T only. We can see a similar pattern in MSE_β (Figure 1) and Bias^2_β (Figure 2) when ϕ is large $(\phi = 0.2, 0.3)$, even though Var_β

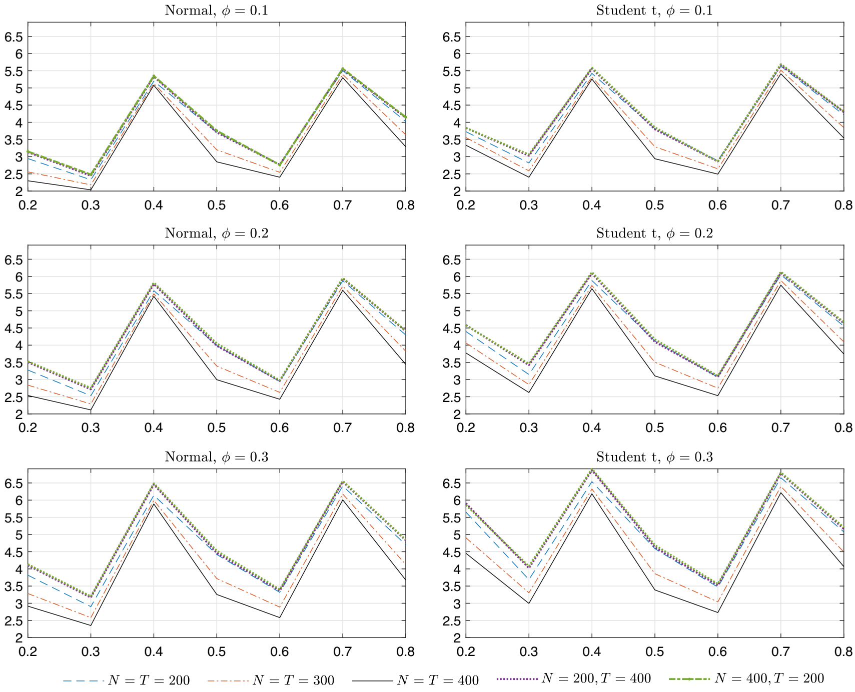

FIGURE 6. MaxDev_L.

(Figure 3) decreases when only N or T increases. This is because Bias_β^2 dominates in MSE_β due to penalization, so, again, the rate of convergence of $\hat{\beta}(u)$ only depends on $N \wedge T$ under these ϕs. Yet, we can also see that when $\phi = 0.1$, MSE_β does shrink when only one of N and T increases. This suggests that the error bound for $\hat{\beta}(u)$ is an upper bound and the actual rate of convergence of $\hat{\beta}(u)$ may be faster when the regressors and the common component only have weak correlation. This is left for future research.

Across different specifications, we can see that convergence is robust to heavy tailed outcome variables, although at a given sample size, MSE_β and MSE_L are all greater than in the case where the outcome variable has a thinner tail. Meanwhile, MSE_β and MSE_L are smaller when ϕ is smaller, whereas MSE_q (Figure 5) stays relatively unchanged across different ϕ. This is expected because with a higher correlation between the covariates and the low-rank common component, the constant C_{RSC} tends to be smaller, making it harder to separate the regressor matrices from the common component, but MSE_q, showing how well the conditional quantile is fitted, does not rely on the quality of such separation.

Finally, although we do not have a theory on the convergence of $\|\hat{L}(u) - L_0(u)\|_\infty$, we can see from Figure 6 that $\|\hat{L}(u) - L_0(u)\|_\infty$ also seems to converge to zero with a rate of convergence only depending on $N \wedge T$.

In terms of computation speed, the computation time in these experiments is from around 2–38 s. We also computed the iterative estimator (1.2) in the Introduction under the true number of interactive fixed effects. Our estimator computes much faster than it.

7. CONCLUDING REMARKS

In this paper, we study a conditional quantile panel data model with interactive fixed effects. By exploiting the low-rankness of the matrix formed by the interactive fixed effects, we propose a nuclear norm penalized estimator. The estimator jointly estimates the coefficients and the low-rank matrix by solving a convex problem. We derive a uniform error bound on the estimator and in turn, establish uniform consistency. Based on the error bound, we also construct a consistent estimator of the number of interactive fixed effects at any given quantile level. From the Monte Carlo simulations, our estimator performs well and computes quickly.

We conjecture that after the penalized estimator and the rank estimator are obtained, by using them as the initial value, a few rounds of iterations based on the iterative estimator's minimization problem would remove the bias and restore the rate of convergence of the coefficient estimator that is slowed down by penalization. Inference may also be available based on such post-penalization procedures. Moreover, as we find that the penalization allows certain low-rank regressors, it may be interesting to investigate how the performance of the estimator changes when explicitly imposing the assumption that some or all the regressors are low-rank. We leave these questions for future work.

A. Implementation of the Estimator

We adapt an augmented Lagrange multiplier (ALM) algorithm introduced in Lin et al. (2010), Candès et al. (2011), and Yuan and Yang (2013). Rewrite the minimization problem (2.3) as[8]

$$\min_{L,V,\beta} \ \frac{1}{\lambda NT}\boldsymbol{\rho}_u(V)+\|L\|_*,$$
$$s.t. \ \sum_{j=1}^{p} X_j\beta_j + L + V = Y.$$

The ALM method is based on the augmented Lagrangian

$$\mathscr{L}(L,\beta,V,H)=\frac{1}{\lambda NT}\boldsymbol{\rho}_u(V)+\|L\|_*+\left\langle H, Y-\sum_{j=1}^{p}X_j\beta_j-L-V\right\rangle+\frac{\mu}{2}\left\|Y-\sum_{j=1}^{p}X_j\beta_j-L-V\right\|_F^2, \tag{A.1}$$

[8] In theory, the estimator (2.3) proposed in Section 2 solves a constrained minimization problem with $\|L\|_\infty \le \alpha_{NT}$. However, since α_{NT} is allowed to grow to infinity with N and T, in practice, we can set α_{NT} as a large positive number, solve an unconstrained problem using the algorithm in this appendix, and check whether the obtained $\hat{L}(u)$ satisfies the constraint. Moreover, we propose alternative assumptions in Section S.A.1 of the Supplementary Material; under them, the constraint in the minimization problem can be dropped.

where $H \in \mathbb{R}^{N\times T}$ is the Lagrangian multiplier of the linear constraint $\sum_{j=1}^{p} X_j\beta_j + L + V = Y$ and $\mu > 0$ is the penalty parameter for the violation of the constraint. By separability of the parameters in $\mathscr{L}$, the ALM method treats L, (β, V) and H as three blocks of parameters, and iteratively updates each one of them at a time until converged. When updating (β, V), inner-loop iterations are carried out by updating one of β and V at a time by treating the other parameter as well as L and H fixed. Given the kth step $L^{(k)}, (\beta^{(k)}, V^{(k)})$, and $H^{(k)}$ $((\beta^{(k)}, V^{(k)})$ as the accumulation point in the inner-loop iterations), ALM updates L, β, and V by the first-order condition and H as follows:

$$L\text{-minimization: } 0 \in \nabla \|L^{(k+1)}\|_* - \left(H^{(k)} - \mu\left(V^{(k)} + \sum_{j=1}^{p} X_j\beta_j^{(k)} + L^{(k+1)} - Y\right)\right), \tag{A.2}$$

(V,β)-minimization: Given the lth step update $\beta^{(l)}$ and $V^{(l)}$ in the inner loop,

$$V\text{-minimization: } 0 \in \frac{1}{\lambda NT}\nabla \boldsymbol{\rho}_u(V^{(l+1)}) - \left(H^{(k)} - \mu\left(V^{(l+1)} + \sum_{j=1}^{p} X_j\beta_j^{(l)} + L^{(k+1)} - Y\right)\right), \tag{A.3}$$

$$\beta\text{-minimization: } 0 = \left\langle H^{(k)}, X_j\right\rangle + \mu\left\langle Y - \sum_{j=1}^{p} X_j\beta_j^{(l+1)} - L^{(k+1)} - V^{(l+1)}, X_j\right\rangle, \forall j = 1, \dots, p, \tag{A.4}$$

$$H\text{-minimization: } H^{(k+1)} = H^{(k)} - \mu\left(V^{(k+1)} + \sum_{j=1}^{p} X_j\beta_j^{(k+1)} + L^{(k+1)} - Y\right), \tag{A.5}$$

where ∇ denotes the subgradient operator. It can be verified that the three first-order conditions (A.2)–(A.4) have explicit solutions.

For equation (A.2), let $R^{(k)}\text{diag}(\{\sigma_j^{(k)}\}_j)S^{(k)\prime}$ be a singular value decomposition of the matrix $(Y - V^{(k)} - \sum_{j=1}^{p} X_j\beta_j^{(k)} + H^{(k)}/\mu)$. According to Yuan and Yang (2013), the solution to equation (A.2) is

$$L^{(k+1)} = R^{(k)}\text{diag}\left(\max\left\{\sigma_j^{(k)} - \frac{1}{\mu}, 0\right\}\right)S^{(k)\prime}. \tag{A.6}$$

For equation (A.3), let $\Gamma_V^{(l)} = H^{(k)}/\mu - \sum_{j=1}^{p} X_j\beta_j^{(l)} - L^{(k+1)} + Y$. For every $i = 1, \dots, N$ and $t = 1, \dots, T$, $\left(\nabla \boldsymbol{\rho}_u\left(V^{(l+1)}\right)\right)_{it} = u\mathbb{1}(V_{it}^{(l+1)} > 0) + (u-1)\mathbb{1}(V_{it}^{(l+1)} < 0)$. It can be verified that the following is a solution:

$$V_{it}^{(l+1)} = \begin{cases} \max\left\{\Gamma_{V,it}^{(l)} - \frac{u}{\mu\lambda NT}, 0\right\}, & \text{if } \Gamma_{V,it}^{(l)} \geq 0, \\ -\max\left\{-\Gamma_{V,it}^{(l)} - \frac{1-u}{\mu\lambda NT}, 0\right\}, & \text{if } \Gamma_{V,it}^{(l)} < 0. \end{cases} \tag{A.7}$$

For equation (A.4), it is the first-order condition of a least square problem. Let $\Gamma_\beta^{(l+1)} = Y - L^{(k+1)} - V^{(l+1)} + H^{(k)}/\mu$. Define the $NT \times p$ matrix $\boldsymbol{X} = (\text{vec}(X_1), \dots, \text{vec}(X_p))$. Then,

$$\beta^{(l+1)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\text{vec}\left(\Gamma_\beta^{(l+1)}\right)\right). \tag{A.8}$$

Finally, following Yuan and Yang (2013), we set $\mu = 0.25NT/\|Y\|_1$. The termination criterion for the outer loop[9] is set as $\|\beta^{(k+1)} - \beta^{(k)}\|_F^2/p + \|L^{(k+1)} - L^{(k)}\|_F^2/NT \leq 10^{-6}$, and for the inner loop is set as $\|\beta^{(l+1)} - \beta^{(l)}\|_F^2/p \leq 10^{-4}$. The following algorithm summarizes these steps.

Algorithm 1: Nuclear Norm Penalized Quantile Regression by ALM.

```
initialize: $\beta^0 = \mathbf{0}, V^0 = H^0 = \mathbf{0}, \mu = 0.25NT/\|Y\|_1, \lambda =$
            $\sqrt{\log(NT)(N \vee T)}/(4NT)$.
while not converged do
    compute $L^{(k+1)}$ as (A.6);
    while not converged do
        compute $V^{(l+1)}$ as (A.7);
        compute $\beta^{(l+1)}$ as (A.8);
    end
    compute $H^{(k+1)}$ as (A.5);
end
output: $\beta, L$.
```

On the computation side, although Algorithm 1 has two loops, the most computation-intensive part, i.e., computing $L^{(k+1)}$ which involves singular value decomposition, is designed to only show up in the outer-loop. The inner loop only involves elementary computation and solving a least square problem. We find in the simulation study that the algorithm computes very fast.

On the theory side, Section S.B.5 of the Supplementary Material proves that the solution to Algorithm 1 is a solution to the minimization problem (2.3) which defines our estimator.

SUPPLEMENTARY MATERIAL

Feng, J. (2023): Supplement to "Nuclear norm regularized quantile regression with interactive fixed effects," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466623000129.

REFERENCES

Abrevaya, J. & C.M. Dahl (2008) The effects of birth inputs on birthweight: Evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics* 26(4), 379–397.

Agarwal, A., S. Negahban, & M.J. Wainwright (2012) Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics* 40(2), 1171–1197.

[9] We also experimented with different termination criteria for instance by including $\|V^{(k+1)} - V^{(k)}\|_F^2/NT$ and/or $\|H^{(k+1)} - H^{(k)}\|_F^2/NT$ (as in Lin et al., 2010; Candès et al., 2011). The results (including computation time) are almost the same.

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

Ando, T. & J. Bai (2020) Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association* 115(529), 266–279.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, & K. Khosravi (2021) Matrix completion methods for causal panel data models. *Journal of the American Statistical Association* 116(536), 1716–1730.

Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica* 71(1), 135–171.

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

Bai, J. & J. Feng (2019) Robust principal component analysis with non-sparse errors. Preprint, arXiv:1902.08735.

Bai, J. & S. Ng (2019) Rank regularized estimation of approximate factor models. *Journal of Econometrics* 212(1), 78–96.

Bai, J. & S. Ng (2021) Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association* 116(536), 1746–1763.

Belloni, A., M. Chen, O.H.M. Padilla, & Z.K. Wang (2023) High-dimensional latent panel quantile regression with an application to asset pricing. *Annals of Statistics* 51(1), 96–121.

Belloni, A. & V. Chernozhukov (2011) $\ell 1$-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39(1), 82–130.

Beyhum, J. & E. Gautier (2019) Square-root nuclear norm penalized estimator for panel data models with approximately low-rank unobserved heterogeneity. Preprint, arXiv:1904.09192.

Canay, I.A. (2011) A simple approach to quantile regression for panel data. *Econometrics Journal* 14(3), 368–386.

Candès, E.J., X. Li, Y. Ma, & J. Wright (2011) Robust principal component analysis? *Journal of the ACM* 58(3), 1–37.

Candès, E.J. & B. Recht (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* 9(6), 717–772.

Chao, S.-K., W.K. Härdle, & M. Yuan (2021) Factorisable multitask quantile regression. *Econometric Theory* 37(4), 794–816.

Chen, L. (2022) Two-step estimation of quantile panel data models with interactive fixed effects. *Econometric Theory*, 1–28, first view 18 August 2022. doi:10.1017/S0266466622000366.

Chen, L., J.J. Dolado, & J. Gonzalo (2021) Quantile factor models. *Econometrica* 89(2), 875–910.

Chernozhukov, V., C. Hansen, Y. Liao, & Y. Zhu (2019) Inference for heterogeneous effects using low-rank estimations. Preprint, arXiv:1812.08089.

Galvao, A.F. & K. Kato (2016) Smoothed quantile regression for panel data. *Journal of Econometrics* 193(1), 92–112.

Galvao, A.F., C. Lamarche, & L.R. Lima (2013) Estimation of censored quantile regression for panel data with fixed effects. *Journal of the American Statistical Association* 108(503), 1075–1089.

Ganesh, A., J. Wright, X. Li, E. J. Candès, & Y. Ma (2010) Dense error correction for low-rank matrices via principal component pursuit. In *2010 IEEE International Symposium on Information Theory*, pp. 1513–1517. IEEE.

Harding, M. & C. Lamarche (2014) Estimating and testing a quantile regression model with interactive effects. *Journal of Econometrics* 178, 101–113.

Hsu, D., S.M. Kakade, & T. Zhang (2011) Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory* 57(11), 7221–7234.

Kato, K., A.F. Galvao, Jr., & G.V. Montes-Rojas (2012) Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics* 170(1), 76–91.

Knight, K. (1998) Limiting distributions for l_1 regression estimators under general conditions. *Annals of Statistics* 26(2), 755–770.

Koenker, R. (2004) Quantile regression for longitudinal data. *Journal of Multivariate Analysis* 91(1), 74–89.

Lamarche, C. (2010) Robust penalized quantile regression estimation for panel data. *Journal of Econometrics* 157(2), 396–408.

Lin, Z., M. Chen, and Y. Ma (2010). The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. Preprint, arXiv:1009.5055.

Ma, S., L. Su, & Y. Zhang (2022) Detecting latent communities in network formation models. *Journal of Machine Learning Research* 23(1), 13971–14031.

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

Moon, H. R. and M. Weidner (2019) Nuclear norm regularized estimation of panel regression models. Preprint, arXiv:1810.10987.

Negahban, S. & M.J. Wainwright (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics* 39(2), 1069–1097.

Negahban, S. & M.J. Wainwright (2012) Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research* 13(May), 1665–1697.

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

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

Yuan, X. & J. Yang (2013) Sparse and low-rank matrix decomposition via alternating direction method. *Pacific Journal of Optimization* 9(1), 167.

Zhou, Z., X. Li, J. Wright, E. Candès, & Y. Ma (2010) Stable principal component pursuit. In *2010 IEEE International Symposium on Information Theory*, pp. 1518–1522. IEEE.