SFB 649 Discussion Paper 2016-057

Factorisable Multi-Task Quantile Regression

Shih-Kang Chao *
Wolfgang K. Härdle *[2]
Ming Yuan *[3]

* Purdue University, United States of America
*[2] Humboldt-Universit¨at zu Berlin, Germany
*[3] University of Wisconsin-Madison, United States of America

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

Factorisable Multi-Task Quantile Regression*

Shih-Kang Chao[†] Wolfgang K. Härdle[†‡] Ming Yuan[§]

July 7, 2016

Abstract

For many applications, analyzing multiple response variables jointly is desirable because of their dependency, and valuable information about the distribution can be retrieved by estimating quantiles. In this paper, we propose a multi-task quantile regression method that exploits the potential factor structure of multivariate conditional quantiles through nuclear norm regularization. We jointly study the theoretical properties and computational aspects of the estimating procedure. In particular, we develop an efficient iterative proximal gradient algorithm for the non-smooth and non-strictly convex optimization problem incurred in our estimating procedure, and derive oracle bounds for the estimation error in a realistic situation where the sample size and number of iterative steps are both finite. The finite iteration analysis is particular useful when the matrix to be estimated is big and the computational cost is high. Merits of the proposed methodology are demonstrated through a Monte Carlo experiment and applications to climatological and financial study. Specifically, our method provides an objective foundation for spatial extreme clustering, and gives a refreshing look on the global financial systemic risk. Supplementary materials for this article are available online.

KEY WORDS: Factor model; Fast iterative shrinkage-thresholding algorithm; Multivariate Regression; Spatial extreme; Financial risk.

---

*Financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Economic Risk", IRTG 1792, Einstein Foundation Berlin via the Berlin Doctoral Program in Economics and Management Science (BDPEMS), and National Science Foundation and National Institute of Health of the US are gratefully acknowledged.

[†]Department of Statistics, Purdue University, West Lafayette, IN 47906. E-mail: `skchao74@purdue.edu`. Tel: +1 (765) 496-9544. Fax: +1 (765) 494-0558. Partially supported by Office of Naval Research (ONR N00014-15-1-2331).

[‡]Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. Email: `haerdle@wiwi.hu-berlin.de`. Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore.

[§]Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, U.S.A. Email: `myuan@stat.wisc.edu`.

1. Introduction

In a variety of applications in economics (Koenker and Hallock (2001)), biology (Briollais and Durrieu (2014)), ecology (Cade and Noon (2003)), and atmospheric sciences (for example, Friederichs and Hense (2007); Bremnes (2004); Reich et al. (2011); Reich (2012)), the interest is in the conditional quantiles of the response variable. For a single response variable, quantile regression (Koenker and Bassett; 1978) is widely acknowledged as a very convenient and efficient method to estimate conditional quantiles. However, we are often required to consider a multi-task framework, in which the responses $\boldsymbol{Y} = (Y_1, ..., Y_m)$ are predicted by a common vector $\boldsymbol{X} = (X_1, ..., X_p)$, where p, m grow with sample size n. Existing literature on the multi-task quantile regression either assumes a particular structure between the response variables and predictors (Fan et al.; 2015), or considers a factor model where the factors do not depend on the quantile levels (Ando and Tsay; 2011; Chen et al.; 2015) with p, m much smaller than n.

To analyze noncanonical and asymmetric data arising from many applications, we consider a flexible quantile factor model that allows the factor to vary with the quantile level, while making no assumption on the association between the response and prediction variables. Given factors $f_k^\tau(\boldsymbol{X})$ for $k = 1, ..., r_\tau$ for a quantile level $0 < \tau < 1$, we assume the conditional quantile $q_j(\tau|\boldsymbol{X}_i)$ for Y_j in $\boldsymbol{Y}$ at τ has a linear expression in terms of $f_k^\tau(\boldsymbol{X})$,

$$q_j(\tau|\boldsymbol{X}) = \sum_{k=1}^{r_\tau} \Psi_{kj,\tau} f_k^\tau(\boldsymbol{X}), \quad j = 1, ..., m, \tag{1.1}$$

where $\Psi_{kj,\tau} \in \mathbb{R}$ is the factor loading, and r_τ is fixed and much less than the sample size n.

The factors $f_k^\tau(\boldsymbol{X})$ are flexible for analyzing Y_j, which possibly depends on $\boldsymbol{X}$ in a very irregular way. An important special example is the two-piece normal distribution, which is a combination of two centered normal distributions with different variances at the origin. The two-piece normal distribution is especially suitable for modeling the *asymmetric* *likelihood* of upward and downward movement, which is exploited by the Bank of England for making inflation rate prediction intervals (Wallis; 1999, 2014). However, if Y_j follows a two-piece normal distribution whose variances for the left and right part of the distribution are two distinct functions of $\boldsymbol{X}$, traditional approaches such as principal component analysis (PCA) fail to correctly estimate the factors for $\boldsymbol{Y}$, since PCA ignores the fact that they are asymmetric and non-Gaussian. Consequently, the resulting factors are misleading.

Because the factors $f_k^\tau(\boldsymbol{X})$ are latent, direct estimation of the parameters $\Psi_{kj,\tau}$ for $k = 1, ..., r_\tau$ and $j = 1, ..., m$ is not feasible. Therefore, we need additional assumptions. If the transformations $f_k^\tau(\boldsymbol{X}_i)$ are *linear* in $\boldsymbol{X}$, that is, $f_k^\tau(\boldsymbol{X}_i) \stackrel{\text{def}}{=} \boldsymbol{\varphi}_{k,\tau}^\top \boldsymbol{X}_i$, where $\boldsymbol{\varphi}_{k,\tau} = (\varphi_{k1,\tau}, ..., \varphi_{kp,\tau})^\top \in \mathbb{R}^p$, we can rewrite the model (1.1) as

$$q_j(\tau|\boldsymbol{X}_i) = (\boldsymbol{\Gamma}_\tau)_{*j}^\top \boldsymbol{X}_i, \quad i = 1, ..., n, \tag{1.2}$$

where $\boldsymbol{\Gamma}_\tau$ is defined in an obvious manner, and $(\boldsymbol{\Gamma}_\tau)_{*j}$ is the jth column of matrix $\boldsymbol{\Gamma}_\tau$. We note that factors $f_k^\tau(\boldsymbol{X})$ are frequently assumed linear in $\boldsymbol{X}$ in applied statistics and financial econometrics; see, for example, Section 2.2 and Chapter 8 of Reinsel and Velu (1998) for practical examples.

The main focus of this paper is on estimating the matrix $\boldsymbol{\Gamma}_\tau$ in (1.2). After a factorization of the estimated matrix, we obtain the estimated factors and loadings simultaneously; see Section 2.2 for further detail. We may identify $\boldsymbol{\Gamma}_\tau \in \arg\min_{\mathbf{S} \in \mathbb{R}^{p\times m}} Q_\tau(\mathbf{S})$, where $Q_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \mathsf{E}[\widehat{Q}_\tau(\mathbf{S})]$ and

$$\widehat{Q}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big). \tag{1.3}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}\{u \le 0\})$ is the "check function" that forces $\boldsymbol{X}_i^\top \mathbf{S}_{*j}$ to be close to the τ quantile of Y_j as argued in the seminal paper of Koenker and Bassett (1978). $\widehat{Q}_\tau$ is similar to the loss function used in Koenker and Portnoy (1990).

The number of unknown parameters mp may be larger than n in our model, which makes the direct estimation of (1.3) infeasible. We make a key observation that $\mathbf{\Gamma}_\tau$ in (1.2) is of rank r_τ, which is assumed much less than p, m. This observation motivates us to the estimator

$$\widehat{\mathbf{\Gamma}}_\tau \stackrel{\text{def}}{=} \arg \min_{\mathbf{S}\in\mathbb{R}^{p\times m}} \big\{ L_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \widehat{Q}_\tau(\mathbf{S}) + \lambda_\tau \|\mathbf{S}\|_* \big\}, \tag{1.4}$$

where $\|\mathbf{S}\|_*$ is the nuclear norm (sum of singular values) and λ_τ is a user supplied tuning parameter. Nuclear norm encourages the sparsity in the rank of the solution $\widehat{\mathbf{\Gamma}}_\tau$, see Yuan et al. (2007); Bunea et al. (2011); Negahban and Wainwright (2011); Negahban et al. (2012) for the application of nuclear norm penalty in a multivariate mean regression framework.

Despite of theoretical properties of $\widehat{\mathbf{\Gamma}}_\tau$ (see appendix), solving (1.4) exactly for the matrix $\widehat{\mathbf{\Gamma}}_\tau$ is difficult in practice because the first term on the right of (1.4) is neither smooth nor strictly convex. Our first contribution is an efficient algorithm that generates a sequence of matrices $\mathbf{\Gamma}_{\tau,t}$, which converges to $\widehat{\mathbf{\Gamma}}_\tau$ as the number of iterations $t \to \infty$. The algorithm combines the popular smoothing procedure of Nesterov (2005) and the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) of Beck and Teboulle (2009). A convergence analysis shows that it requires $\mathcal{O}(1/\epsilon)$ iterations for the difference in loss function in (1.4) evaluated at the two neighboring steps to be less than ϵ, which is more efficient than $\mathcal{O}(1/\epsilon^2)$ iterations required by the general subgradient method.

The property of the approximating sequence $\mathbf{\Gamma}_{\tau,t}$ is further characterized by a novel error bound for the Frobenius norm $\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}$ under *finite sample* and *finite iterative steps*. We are interested in finite iteration because when p, m are large, one iteration may take a lot of time as a singular value decomposition is required in each step. Hence, in practice one cannot compute too many iterations. Our theoretical results provide a rule for determining the number of iterations that ensures the oracle rate of the resulting estimator. The proof is founded on one of our intermediate results that the difference $\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}_\tau$ lies in a star-shaped set rather than a cone. This result shares a similar flavor to the estimation for high-dimensional matrix, which is not exactly sparse in rank; see Negahban et al. (2012). In the bulk of the proof of our main theorem, we apply modern random matrix theory which gives a very sharp bound on the spectral norm of a sum of random matrices. Finally, under the realistic situation of finite sample and finite iteration, we derive realistic bounds for the estimation error for factors and loadings, using a state-of-the-art bound of Yu et al. (2015) on the distance between subspaces spanned by the eigenvectors of two matrices.

We demonstrate the performance of our estimator by a Monte Carlo experiment, with data generated from a two-piece normal distribution; see (4.1) for the data generating model. In order to show how our estimator performs for asymmetric data, we consider both high and low asymmetry. We compare our estimator with an oracle estimator, which is estimated under the knowledge of the true rank of $\mathbf{\Gamma}_\tau$. The simulation results show that the difference between $\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}_\tau\|_\mathrm{F}$ and the oracle difference $\|\mathbf{\Gamma}^{or}_{\tau,t} - \mathbf{\Gamma}_\tau\|_\mathrm{F}$ is around 5-10% of the oracle difference. The number of iterations required is generally below 40. Both the error and the required number of iteration increases when τ is close to 0 and 1.

We remark that the our computational method and theoretical tool may be interesting for other multi-task learning problems with non-smooth loss functions that are not strictly convex, such as the support vector machine.

We show that some modern scientific challenges in climatology and finance may be addressed with our method. In climatology, the study of inference methods for spatial extreme is a highly active research area (Davison et al.; 2012). We quantify spatial dependence of extreme temperature across China with our method, which provides an objective rule for spatial extreme clustering. Spatial clustering based on extreme behavior of atmospheric variables has attracted much interest recently (Bernard et al.; 2013; Bador et al.; 2015), because summarizing the data originally observed at a large collection of locations by very few spatial clusters is essential for avoiding the hefty computational cost (Castruccio et al.; 2015) required by the statistical inference of spatial extremes. For financial study, we show via global stock price data that the stock price of firms with large market value and high leverage (the ratio of short and long term debt over common equity) tend to be more vulnerable to systemic risk. Our finding is consistent with the finding of White et al. (2015), but our computational method is scalable to a higher dimension.

The rest of this paper is organized as follows. Section 2 is devoted to the algorithm for finding a good approximating sequence $\mathbf{\Gamma}_{\tau,t}$ approximating $\widehat{\mathbf{\Gamma}}_{\tau}$ defined in (1.4), the estimation of factors and loadings, the choice of λ_τ and the analysis of the convergence properties of the algorithm. In Section 3, the oracle properties of $\mathbf{\Gamma}_{\tau,t}$ and the estimator for factors and loadings are investigated. In Section 4, a Monte Carlo experiment is presented. In Section 5, we analyze challenging scientific questions using our method. Proofs are shifted to the supplementary material.

Notations. In the rest of the paper, we sometimes suppress "τ" in $\mathbf{\Gamma}_\tau$, $\widehat{\mathbf{\Gamma}}_\tau$, λ_τ etc. for brevity, when it does not cause confusion. Given two scalars x and y, $x \wedge y \stackrel{\text{def}}{=} \min\{x, y\}$ and $x \vee y \stackrel{\text{def}}{=} \max\{x, y\}$. $\mathbf{1}(x \leq 0)$ is an index function, which is equal to 1 when $x \leq 0$ and 0 when $x > 0$. For a vector $\mathbf{v} \in \mathbb{R}^p$, let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$ and $\|\mathbf{v}\|_\infty$ be the vector ℓ_1, ℓ_2 and ℓ_∞ norm. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{p\times m}$, denote the singular values of $\mathbf{A}$: $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq ... \geq \sigma_{p\wedge m}(\mathbf{A})$, and we usually write the singular value decomposition (abbreviated as SVD henceforth) $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. We sometimes also write $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ for the largest and smallest singular values of $\mathbf{A}$. Let $\|\mathbf{A}\| = \sigma_{\max}(\mathbf{A})$, $\|\mathbf{A}\|_*$ and $\|\mathbf{A}\|_\mathrm{F}$ be the spectral, nuclear and Frobenius norm of a matrix $\mathbf{A}$. If $\mathbf{A} \in \mathbb{R}^{p\times m}$, for a probability distribution P_X for $\boldsymbol{X} \in \mathbb{R}^p$, define

$$\|\mathbf{A}\|^2_{L_2(P_X)} \stackrel{\text{def}}{=} m^{-1}\mathsf{E}_{P_X}\|\mathbf{A}^\top \boldsymbol{X}_i\|_2^2. \tag{1.5}$$

Denote $\mathbf{A}_{*j}$ and $\mathbf{A}_{i*}$ as the jth column vector and the ith row vector of $\mathbf{A}$. $\mathbf{I}_p$ denotes the $p\times p$ identity matrix. For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times m}$, $\langle\cdot,\cdot\rangle : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \to \mathbb{R}$

denotes the trace inner product given by $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^\top)$. Define the empirical measure of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ by $\mathbb{P}_n$, and the true underlying measure by P with the corresponding expectation as E. For a function $f : \mathbb{R}^p \to \mathbb{R}$, and $\boldsymbol{Z}_i \in \mathbb{R}^p$, define the *empirical process* $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n \{f(\boldsymbol{Z}_i) - \mathsf{E}[f(\boldsymbol{Z}_i)]\}$. Define the "check" function and its subgradient by

$$\rho_\tau(u) \stackrel{\text{def}}{=} u(\tau - \mathbf{1}\{u \leq 0\}), \quad \psi_\tau(u) \stackrel{\text{def}}{=} \tau - \mathbf{1}(u \leq 0).$$

For vectors $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ in $\mathbb{R}^p$, denote $[\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ ... \ \boldsymbol{a}_m] \in \mathbb{R}^{p \times m}$ a matrix with $\boldsymbol{a}_j$ being its jth column. Let $\mathbf{0}_p$ be a p-vector of zeros.

Definition 1.1 (Sub-Gaussian variable and sub-Gaussian norm)**.** *A random variable X is called* sub-Gaussian *if there exists some positive constant K_2 such that* $\mathrm{P}(|X| > t) \leq \exp(1 - t^2/K_2^2)$ *for all $t \geq 0$. The* sub-Gaussian norm *of X is defined as* $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathsf{E}|X|^p)^{1/p}$.

2. Computation

In this section, we discuss an efficient algorithm that generates a sequence to approximate the solution of (1.4), which we call "QISTA". Section 2.1 describes the ideas of the algorithm, which is stated formally in Algorithm 1. Section 2.2 explains the computation of factors and loadings. Section 2.3 discusses the choice of tuning parameter λ. Section 2.4 gives an algorithmic convergence result in Theorem 2.3, whose proof is in the supplementary material.

2.1. A Generalization of FISTA to Non-smooth Loss Function

Obtaining the exact solution for (1.4) is difficult because $\widehat{Q}_\tau(\mathbf{S})$ defined in (1.3) is neither smooth nor strictly convex. In this section we describe an algorithm that generates a sequence of $\mathbf{\Gamma}_{\tau,t}$ which approximates $\widehat{\mathbf{\Gamma}}$. The major challenge is that the subgradient of $\widehat{Q}_\tau(\mathbf{S})$

is not Lipschitz, so the FISTA algorithm of Beck and Teboulle (2009) cannot be applied straightforwardly. To resolve this problem, we need to find a "nice" surrogate for $\widehat{Q}_\tau(\mathbf{S})$.

To develop the ideas, recall from (1.4) that the objective function to be minimized is

$$L_\tau(\mathbf{S}) = (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big) + \lambda\|\mathbf{S}\|_* = \widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*, \tag{2.1}$$

where $\widehat{Q}_\tau(\mathbf{S})$ is neither smooth nor strictly convex. To handle this problem, we introduce the dual variables Θ_{ij}:

$$\widehat{Q}_\tau(\mathbf{S}) = \max_{\Theta_{ij}\in[\tau-1,\tau]}(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\Theta_{ij}\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big). \tag{2.2}$$

See Section S.1.1 in the supplementary material for a proof of (2.2). To smooth this function, denote the matrix $\boldsymbol{\Theta} = (\Theta_{ij})$ for $i = 1, ..., n,\ j = 1, ..., m$, we consider a smooth approximation to $\widehat{Q}_\tau(\mathbf{S})$ as in equation (2.5) of Nesterov (2005):

$$\widehat{Q}_{\tau,\kappa}(\mathbf{S}) \overset{\text{def}}{=} \max_{\Theta_{ij}\in[\tau-1,\tau]}\Big\{(mn)^{-1}\widetilde{Q}_\tau(\mathbf{S},\boldsymbol{\Theta}) - \frac{\kappa}{2}\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2\Big\}, \tag{2.3}$$

where $\widetilde{Q}_\tau(\mathbf{S},\boldsymbol{\Theta}) \overset{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{m}\Theta_{ij}\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big)$, and $\kappa > 0$ is a smoothing regularization constant depending on m, n and the desired accuracy. When $\kappa \to 0$, the approximation is getting closer to the function before smoothing, as shown in Figure 2.1. $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ defined in (2.3) has Lipschitz gradient

$$\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{S}) \overset{\text{def}}{=} -(mn)^{-1}\mathbf{X}^\top[[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{S})]]_\tau, \tag{2.4}$$

where $\mathbf{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ ...\ \boldsymbol{X}_n]^\top$, $[[\mathbf{A}]]_\tau = ([[A_{ij}]]_\tau)$ performs component-wise truncation on a real matrix $\mathbf{A}$ to the interval $[\tau - 1, \tau]$; in particular,

$$[[A_{ij}]]_\tau = \begin{cases} \tau, & \text{if } A_{ij} \geq \tau; \\ A_{ij}, & \text{if } \tau - 1 < A_{ij} < \tau; \\ \tau - 1, & \text{if } A_{ij} \leq \tau - 1. \end{cases}$$

Observe that (2.4) is similar to the subgradient $-\mathbf{X}\{\tau - \mathbf{1}(\mathbf{Y} - \mathbf{XS} \leq 0)\}$ of $\widehat{Q}_\tau(\mathbf{S})$, where the operator $\tau - \mathbf{1}(\cdot \leq 0)$ applies component-wise to the matrix $\mathbf{Y} - \mathbf{XS}$ with a slight abuse of notation. The major difference lies in the fact that (2.4) replaces the discrete non-Lipschitz $\tau - \mathbf{1}(\mathbf{Y} - \mathbf{XS} \leq 0)$ with a Lipschitz function $[[\kappa^{-1}(\mathbf{Y} - \mathbf{XS})]]_\tau$. Figure 2.1 illustrates this in a univariate framework with $m = n = 1$ and $\mathbf{X} = 1$.

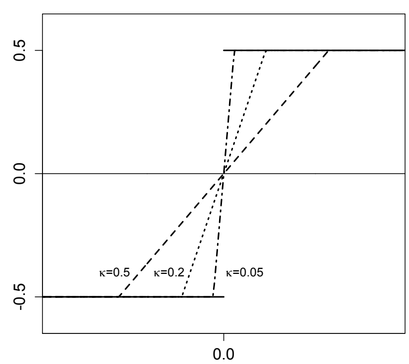

Figure 2.1: The solid line is the function $\psi_\tau(u) = \tau - \mathbf{1}(u \leq 0)$ with $\tau = 0.5$, which has a jump at the origin. The dashed line corresponds to the smoothing gradient $[[\kappa^{-1}(\mathbf{Y} - \mathbf{XS})]]_\tau$ associated with $\kappa = 0.5$. As κ decreases to 0.05, we observe that the smoothing approximation function is closer to $\psi_\tau(u)$.

Now, we replace the optimization problem involving $L_\tau(\mathbf{S})$ in (2.1) by the one involving

$$\widetilde{L}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \widehat{Q}_{\tau,\kappa}(\mathbf{S}) + \lambda \|\mathbf{S}\|_*, \tag{2.5}$$

where we recall the definition of $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ in (2.3). Since the gradient of $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ is Lipschitz, we may apply FISTA of Beck and Teboulle (2009) for minimizing (2.5). Define $S_\lambda(\cdot)$ to be the proximity operator on $\mathbb{R}^{p\times m}$:

$$S_\lambda(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{U}(\mathbf{D}-\lambda\mathbf{I}_{p\times m})_+\mathbf{V}^\top, \tag{2.6}$$

where $\mathbf{I}_{p\times m}$ is the $p\times m$ rectangular identity matrix with the main diagonal elements equal to 1, and the SVD $\mathbf{S}=\mathbf{U}\mathbf{D}\mathbf{V}^\top$. See Theorem S.4.2 in the supplementary material for more detail for the proximity operator. We are now ready to state Algorithm 1 for the optimization problem (1.4). The name of the algorithm reflects the fact that it is an ISTA algorithm for regression quantiles.

Algorithm 1: Quantile Iterative Shrinkage-Thresholding Algorithm (QISTA)

Input: $\mathbf{Y}$, $\mathbf{X}$, $0<\tau<1$, λ, $\epsilon=10^{-6}$, T(chosen as (2.12)) $\kappa=\frac{\epsilon}{2mn}$, $M=\frac{1}{\kappa m^2n^2}\|\mathbf{X}\|^2$;
Initialization: $\mathbf{\Gamma}_{\tau,0}=0$, $\mathbf{\Omega}_{\tau,1}=0$, step size $\delta_1=1$;
for $t=1,2,...,T$ **do**
  $\mathbf{\Gamma}_{\tau,t}=S_{\lambda/M}\big(\mathbf{\Omega}_{\tau,t}-\frac{1}{M}\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{\Omega}_{\tau,t})\big)$;
  $\delta_{t+1}=\frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\mathbf{\Omega}_{\tau,t+1}=\mathbf{\Gamma}_{\tau,t}+\frac{\delta_t-1}{\delta_{t+1}}(\mathbf{\Gamma}_{\tau,t}-\mathbf{\Gamma}_{\tau,t-1})$;
end
Output: $\mathbf{\Gamma}_{\tau,T}$

2.2. Computing Factors and Loadings

To obtain the factors $f_k^\tau(\boldsymbol{X})=\boldsymbol{\varphi}_{k,\tau}^\top\boldsymbol{X}_i$ and loadings $\Psi_{kj,\tau}$ for $j=1,...,m$ and $k=1,...,r_\tau$ which are related to $\mathbf{\Gamma}_\tau$ as in (1.1), by matrix factorization, we may decompose $\mathbf{\Gamma}_\tau=\mathbf{\Phi}_\tau\mathbf{\Psi}_\tau$, where $\mathbf{\Phi}_\tau\in\mathbb{R}^{p\times r}$ and $\mathbf{\Psi}_\tau\in\mathbb{R}^{r\times m}$, and identify $\boldsymbol{\varphi}_{k,\tau}$ as kth column of $\mathbf{\Phi}_\tau$ and $\Psi_{kj,\tau}$ as kj entry of $\mathbf{\Psi}_\tau$. However, decomposition $\mathbf{\Gamma}_\tau=\mathbf{\Phi}_\tau\mathbf{\Psi}_\tau$ is not unique, since for any invertible matrix $\mathbf{P}\in\mathbb{R}^{r\times r}$, we have $\mathbf{\Phi}_\tau\mathbf{\Psi}_\tau=\mathbf{\Phi}_\tau\mathbf{P}\mathbf{P}^{-1}\mathbf{\Psi}_\tau$. Therefore, we need extra r_τ^2 restrictions to fix a matrix $\mathbf{P}$.

We apply the constraint in equation (2.14) on page 28 of Reinsel and Velu (1998): if singular value decomposition $\mathbf{\Gamma}_\tau = \mathbf{U}_\tau \mathbf{D}_\tau \mathbf{V}_\tau^\top$, then we set

$$\mathbf{\Psi}_\tau = \mathbf{V}_\tau \text{ and } \mathbf{\Phi}_\tau = \mathbf{D}_\tau^\top \mathbf{U}_\tau^\top. \tag{2.7}$$

We also allow for other choices.

For any t, given $\mathbf{\Gamma}_{\tau,t}$ at t iteration from Algorithm 1, we can estimate the factors and loadings using (2.7):

$$\begin{aligned} \widehat{f}_k^\tau(\boldsymbol{X}_i) &= (\mathbf{\Phi}_{\tau,t})_{*k}^\top \boldsymbol{X}_i = \sigma_{k,t}(\mathbf{U}_{\tau,t})_{*k}^\top \boldsymbol{X}_i, \\ \widehat{\mathbf{\Psi}}_\tau &= \mathbf{V}_{\tau,t}, \end{aligned} \tag{2.8}$$

where $\mathbf{V}_{\tau,t} \in \mathbb{R}^{m\times m}$, $\mathbf{D}_{\tau,t} \in \mathbb{R}^{p\times m}$ and $\mathbf{U}_{\tau,t} \in \mathbb{R}^{p\times p}$ are from the singular value decomposition $\mathbf{\Gamma}_{\tau,t} = \mathbf{U}_{\tau,t}\mathbf{D}_{\tau,t}\mathbf{V}_{\tau,t}^\top$, and $\sigma_{k,t}$ is the kth largest singular value of $\mathbf{\Gamma}_{\tau,t}$.

Remark 2.1 (Sign identifiability)**.** *The sign in* (2.7) *is in general indeterminable. Nonetheless, this issue can often be addressed in practice based on the first factor* $f_1^{\tau_1}(\boldsymbol{X}_i) \geq f_1^{\tau_2}(\boldsymbol{X}_i)$ *for* $\tau_1 > \tau_2$. *For implementation, we suggest estimate both* $\widehat{f}_1^{\tau_1}(\boldsymbol{X}_i)$ *and* $\widehat{f}_1^{\tau_2}(\boldsymbol{X}_i)$ *(say* $\tau_1 = 0.9, \tau_2 = 0.1$*), and determine the sign so that* $\widehat{f}_1^{\tau_1}(\boldsymbol{X}_i) \geq \widehat{f}_1^{\tau_2}(\boldsymbol{X}_i)$. *This approach works well in our empirical analysis. Though the monotonicity of empirical quantile curves can be violated (Chernozhukov et al.; 2010; Dette and Volgushev; 2008) and the factors* $\widehat{f}_1^{\tau_1}(\boldsymbol{X}_i) \geq \widehat{f}_1^{\tau_2}(\boldsymbol{X}_i)$ *for* $\tau_1 \geq \tau_2$ *may cross, working with more extreme quantiles (e.g.,* $\tau_1 = 0.9, \tau_2 = 0.1$*) can often resolve the problem.*

2.3. Tuning

For the implementation of Algorithm 1, it is crucial to appropriately select λ. We propose to select λ based on the "pivotal principle". We define the random variable

$$\Lambda_\tau = (nm)^{-1}\|\mathbf{X}^\top \widetilde{\mathbf{W}}_\tau\|, \tag{2.9}$$

where $(\widetilde{W}_\tau)_{ij} = \mathbf{1}(U_{ij} \le 0) - \tau$, $\{U_{ij}\}$ are i.i.d. uniform (0,1) random variables for $i = 1, ..., n$ and $j = 1, ..., m$, independent from $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. The random variable Λ_τ is pivotal conditioning on design $\mathbf{X}$, as it does not depend on unknown $\mathbf{\Gamma}_\tau$. Notice that $(nm)^{-1}\mathbf{X}^\top \widetilde{\mathbf{W}}_\tau = \nabla \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)$, which is the subgradient of $\widehat{Q}_\tau(\mathbf{\Gamma}_\tau)$ defined in (3.1) evaluated at the true matrix $\mathbf{\Gamma}_\tau$. Set

$$\lambda_\tau = 2 \cdot \Lambda_\tau(1 - \eta|\mathbf{X}), \tag{2.10}$$

where $\Lambda_\tau(1-\eta|\mathbf{X}) \stackrel{\text{def}}{=} (1-\eta)$-quantile of Λ_τ conditional on $\mathbf{X}$, for $0 < \eta < 1$ close to 1, for instant $\eta = 0.9$. The choice of λ_τ will be justified theoretically in Section 3.

Remark 2.2. *Using the theory we develop in Section 3, in principle one can select λ based on* (3.7)*, but this does not adapt to the data $\boldsymbol{X}_i$.* (2.10) *is inspired by the high-dimensional quantile regression estimation in Belloni and Chernozhukov (2011).*

2.4. Algorithmic Convergence Analysis

An analysis of the performance of Algorithm 1 is given by the following theorem.

Theorem 2.3 (Convergence analysis of Algorithm 1)**.** *Let $\{\mathbf{\Gamma}_{\tau,t}\}_{t=0}^T$ be the sequence generated by Algorithm 1, $\widehat{\mathbf{\Gamma}}_\tau$ be the optimal solution for minimizing* (2.1) *and $\mathbf{\Gamma}_{\tau,\infty} = \lim_{t\to\infty} \mathbf{\Gamma}_{\tau,t}$ be a minimizer of $\widetilde{L}_\tau(\mathbf{S})$ defined in* (2.5)*. Then for any t and $\epsilon > 0$,*

$$\left|L_\tau(\mathbf{\Gamma}_{\tau,t}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau)\right| \le \frac{3\epsilon(\tau \vee \{1-\tau\})^2}{4} + \frac{4\|\mathbf{\Gamma}_{\tau,0} - \mathbf{\Gamma}_{\tau,\infty}\|_{\mathrm{F}}^2 \|\mathbf{X}\|^2}{(t+1)^2 \epsilon mn}. \tag{2.11}$$

On the other hand, if we require $L_\tau(\mathbf{\Gamma}_{\tau,t}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau) \le \epsilon$, then

$$t_\tau \ge 2\frac{\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_{\tau,0}\|_{\mathrm{F}}\|\mathbf{X}\|}{\epsilon\sqrt{mn}\sqrt{1 - \frac{3(\tau\vee\{1-\tau\})^2}{4}}}. \tag{2.12}$$

See Section S.1.2 in the supplementary material for a proof for Theorem 2.3. The first term on the right-hand side of (2.11) is related to the smoothing error, which cannot be made small by increasing the number of iterations, but can only be reduced by choosing a smaller smoothing parameter κ. The second term is related to the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009).

Remark 2.4 (Convergence Speed)**.** *The algorithm of Beck and Teboulle (2009) yields the convergence rate $\mathcal{O}(1/\sqrt{\epsilon})$. In our case, the smoothing error deteriorates the convergence rate and at best we have $\mathcal{O}(1/\epsilon)$, which is comparable to the rate from a smoothing optimization method of Nesterov (2005). Our rate is an improvement from $\mathcal{O}(1/\epsilon^2)$ of the general subgradient method.*

Remark 2.5 (Effect of τ)**.** *The quantile level τ enters the numerical bound* (2.11) *by* $\big(1-(\tau\vee\{1-\tau\})^2/2\big)^{-1/2}$*, which increases when τ is getting close to the boundary of the interval* $(0,1)$.

Remark 2.6. *Algorithm 1 requires SVD in each iteration, and may be computationally expensive when p, m are very large. Hence, we will derive the bounds for $\mathbf{\Gamma}_{\tau,t}$ under finite t in Section 3. An alternative approach is to formulate the optimization problem* (1.4) *into a semidefinite program and then apply available solvers. See, for example, Jaggi and Sulovský (2010). This approach avoids performing SVD in each step, but in general it requires $\mathcal{O}(1/\epsilon)$ steps to reach an ϵ-accurate solution.*

3. Oracle Properties

In this section we investigate the theoretical properties of the estimator generated by Algorithm 1. Section 3.1 focuses on the estimator $\mathbf{\Gamma}_{\tau,t}$ from the tth iteration of Algorithm 1, and develops a oracle bound for this matrix. Section 3.2 is concerned with the estimation of the factors and loadings, which are defined in Section 2.2.

3.1. Oracle Properties of $\mathbf{\Gamma}_{\tau,t}$

In this section, we present the non-asymptotic oracle bounds of the estimator $\mathbf{\Gamma}_{\tau,t}$ generated by Algorithm 1, which shows that our estimator approximates the true matrix $\mathbf{\Gamma}$ well without knowing the *support* (defined later) of the true matrix. The main result is Theorem 3.6.

In order to develop ideas, we introduce some useful notations. The subgradient for $\widehat{Q}_\tau(\mathbf{S})$ is the matrix

$$\nabla\widehat{Q}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} (nm)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{W}_{\tau,i}(\mathbf{S})^\top = (nm)^{-1}\mathbf{X}^\top\mathbf{W}_\tau(\mathbf{S}) \in \mathbb{R}^{p\times m}, \tag{3.1}$$

where $\mathbf{X} = [\boldsymbol{X}_1 \dots \boldsymbol{X}_n]^\top \in \mathbb{R}^{n\times p}$ is the design matrix and

$$\boldsymbol{W}_{\tau,i}(\mathbf{S}) \stackrel{\text{def}}{=} \big(\mathbf{1}(Y_{ij} - \boldsymbol{X}_i^\top\mathbf{S}_{*j} \le 0) - \tau\big)_{1\le j\le m}, \quad \mathbf{W}_\tau(\mathbf{S}) = [\boldsymbol{W}_{\tau,1}(\mathbf{S}) \dots \boldsymbol{W}_{\tau,n}(\mathbf{S})]^\top \in \mathbb{R}^{n\times m}.$$

We write $\boldsymbol{W}_{\tau,i}(\mathbf{\Gamma}) \stackrel{\text{def}}{=} \boldsymbol{W}_{\tau,i}$ and $\mathbf{W}_\tau \stackrel{\text{def}}{=} \mathbf{W}_\tau(\mathbf{\Gamma})$. For developing the error bounds, we make the following assumptions:

(A1) (Sampling setting) Samples $(\boldsymbol{X}_1,\boldsymbol{Y}_1),\dots,(\boldsymbol{X}_n,\boldsymbol{Y}_n)$ are i.i.d. copies of $(\boldsymbol{X},\boldsymbol{Y})$ random vectors in $\mathbb{R}^{p+m}$. $F^{-1}_{Y_{ij}|\boldsymbol{X}_i}(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top\mathbf{\Gamma}_{*j}(\tau)$.

(A2) (Covariates) Let $\boldsymbol{X} \sim (0,\mathbf{\Sigma}_X)$ whose density exists. Suppose $0 < \sigma_{\min}(\mathbf{\Sigma}_X) < \sigma_{\max}(\mathbf{\Sigma}_X) < \infty$, and there exist constants $B_p, c_1, c_2 > 0$ such that $\|\boldsymbol{X}_i\|$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\boldsymbol{X}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ satisfies

$$\mathrm{P}\big\{\sigma_{\min}(\widehat{\mathbf{\Sigma}}_X) \ge c_1\sigma_{\min}(\mathbf{\Sigma}_X), \sigma_{\max}(\widehat{\mathbf{\Sigma}}_X) \le c_2\sigma_{\max}(\mathbf{\Sigma}_X), \|\boldsymbol{X}_i\| \le B_p\big\} \ge 1-\gamma_n, \tag{3.2}$$

for a sequence $\gamma_n \to 0$.

(A3) (Conditional densities) There exist constants $\bar{f} > 0$, $\underline{f} > 0$ and $\bar{f}' < \infty$ such that

$$\max_{j \le m} \sup_{\boldsymbol{x}, y} \left| f_{Y_j|\boldsymbol{X}}(y|\boldsymbol{x}) \right| \le \bar{f}, \quad \max_{j \le m} \sup_{\boldsymbol{x}, y} \left| \frac{\partial}{\partial y_j} f_{Y_j|\boldsymbol{X}}(y|\boldsymbol{x}) \right| \le \bar{f}', \quad \min_{j \le m} \inf_{\boldsymbol{x}} f_{Y_j|\boldsymbol{X}}(\boldsymbol{x}^\top \boldsymbol{\Gamma}_{*j}|\boldsymbol{x}) \ge \underline{f},$$

where $f_{Y_j|\boldsymbol{X}}$ is the conditional density function of Y_j on $\boldsymbol{X}$.

Assumption (A1) allows us to compute with ease the second moment and the tail probability of some empirical processes (see Remark S.3.4). (A1) may be replaced by m-dependent or weak dependent conditions, but we would need a modified random matrix theory (see the proof for the detail of Theorem 3.6). We leave this for future study. In Assumption (A2), we assume $\mathsf{E}[\boldsymbol{X}] = 0$ for simplicity and it can be easily generalized. B_p is usually assumed uniformly bounded by a constant independent of p in multitask learning literature (for example, p.2 of Maurer and Pontil (2013) and Theorem 1 of Yousefi et al. (2016)). For the condition (3.2), when the $\boldsymbol{X}$ is from a p-Gaussian distribution $N(0, \boldsymbol{\Sigma}_X)$, Lemma 9 in Wainwright (2009) shows that (3.2) holds with $c_1 = 1/9$, $c_2 = 9$ and $\gamma_n = 4\exp(-n/2)$. Vershynin (2012b) discusses the condition (3.2) for a more general class of random vector $\boldsymbol{X}$. (A3) is common in quantile regression literature, see for example Belloni and Chernozhukov (2011); Belloni et al. (2011).

In what follows, we define the "support" of matrices by projections.

Definition 3.1. *For $\mathbf{A} \in \mathbb{R}^{p \times m}$ with rank r, the singular value decomposition of $\mathbf{A}$ is $\mathbf{A} = \sum_{j=1}^{r} \sigma(\mathbf{A}) \mathbf{u}_j \mathbf{v}_j^\top$. The* support *of $\mathbf{A}$ is defined by (S_1, S_2) in which $S_1 = span\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ and $S_2 = span\{\mathbf{v}_1, ..., \mathbf{v}_r\}$. Define the projection matrix on S_1: $\mathbf{P}_1 \overset{\text{def}}{=} \mathbf{U}_{[1:r]} \mathbf{U}_{[1:r]}^\top$, in which $\mathbf{U}_{[1:r]} = [\mathbf{u}_1 \dots \mathbf{u}_r] \in \mathbb{R}^{p \times r}$; $\mathbf{P}_2 \overset{\text{def}}{=} \mathbf{V}_{[1:r]} \mathbf{V}_{[1:r]}^\top$, where $\mathbf{V}_{[1:r]} = [\mathbf{v}_1 \dots \mathbf{v}_r] \in \mathbb{R}^{m \times r}$. Denote $\mathbf{P}_1^\perp = \mathbf{I}_{p \times r} - \mathbf{P}_1$ and $\mathbf{P}_2^\perp = \mathbf{I}_{m \times r} - \mathbf{P}_2$. For any matrix $\mathbf{S} \in \mathbb{R}^{p \times m}$, define*

$$\mathcal{P}_{\mathbf{A}}(\mathbf{S}) \overset{\text{def}}{=} \mathbf{P}_1 \mathbf{S} \mathbf{P}_2; \quad \mathcal{P}_{\mathbf{A}}^\perp(\mathbf{S}) \overset{\text{def}}{=} \mathbf{P}_1^\perp \mathbf{S} \mathbf{P}_2^\perp.$$

Define for any $a \geq 0$,

$$\mathcal{K}(\boldsymbol{\Gamma}; a) \stackrel{\text{def}}{=} \left\{\mathbf{S} \in \mathbb{R}^{p\times m} : \|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\mathbf{S})\|_* \leq 3\|\mathcal{P}_{\boldsymbol{\Gamma}}(\mathbf{S})\|_* + a\right\}. \tag{3.3}$$

See Remark 3.2 for more discussion of the set $\mathcal{K}(\boldsymbol{\Gamma}; a)$.

An important equality we will use repeatedly in the proofs is that for any $\mathbf{S}, \mathbf{A} \in \mathbb{R}^{p\times m}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathbf{A}}(\mathbf{S})\|_* + \|\mathcal{P}_{\mathbf{A}}^{\perp}(\mathbf{S})\|_*$, which essentially corresponds to the decomposability of nuclear norm. See Definition 1 on page 541 of Negahban et al. (2012). Moreover, the rank of $\mathcal{P}_{\mathbf{A}}(\mathbf{S})$ is at most $\text{rank}(\mathbf{A})$.

We remind the readers that singular vectors corresponding to nonzero distinct singular values are uniquely defined, and unique up to a unitary transformation for those corresponding to repeated nonzero singular values. The singular vectors corresponding to 0 singular values are not unique. However, in Definition 3.1 we do not require a unique choice of singular vectors as the nuclear norm is invariant to unitary transformations.

Remark 3.2 (Shape of $\mathcal{K}(\boldsymbol{\Gamma}; a)$)**.** *The shape of $\mathcal{K}(\boldsymbol{\Gamma}; a)$ is not a cone when $a > 0$, but is still a star-shaped set. This set has a similar shape as the set defined in equation (17) on page 544 in Negahban et al. (2012). The reader is referred to their Figure 1 on page 544 for an illustration of that set.*

Remark 3.3. *For any $\boldsymbol{\Delta} \in \mathbb{R}^{p\times m}$, from (A2),*

$$\|\boldsymbol{\Delta}\|_{L_2(P_X)}^2 = m^{-1}\mathsf{E}\big[\|\boldsymbol{\Delta}^\top \boldsymbol{X}_i\|_2^2\big] = m^{-1}\sum_{j=1}^{m} \boldsymbol{\Delta}_{*j}^\top \mathsf{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\top]\boldsymbol{\Delta}_{*j} \geq m^{-1}\sigma_{\min}(\boldsymbol{\Sigma}_X)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2. \tag{3.4}$$

Moreover, by $\|\mathcal{P}_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta})\|_{\mathrm{F}} \leq \|\boldsymbol{\Delta}\|_{\mathrm{F}}$, we have a bound

$$\|\boldsymbol{\Delta}\|_{L_2(P_X)} \geq \left(\frac{\sigma_{\min}(\boldsymbol{\Sigma}_X)}{m}\right)^{1/2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \geq \left(\frac{\sigma_{\min}(\boldsymbol{\Sigma}_X)}{m}\right)^{1/2}\|\mathcal{P}_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta})\|_{\mathrm{F}}. \tag{3.5}$$

We first present some preliminary results. The next lemma gives the bound for $n^{-1}\|\mathbf{X}^\top\mathbf{W}\|$, which leads to a bound for $\|\nabla\widehat{Q}(\mathbf{\Gamma})\|$. The detailed proof can be found in the supplementary material.

Lemma 3.4. *Under assumptions (A1) and (A2),*

$$\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq C^*\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)\{\tau\vee(1-\tau)\}}\sqrt{\frac{p+m}{n}},\ \text{where } C^*=4\sqrt{2\frac{c_2}{C'}\log 8} \tag{3.6}$$

with probability greater than $1-3e^{-(p+m)\log 8}-\gamma_n$, where C' and c_2 are absolute constants given by Lemma S.4.3 in the supplementary material and Assumption (A2).

Please see Section S.2.1 for a proof of Lemma 3.4. We will take

$$\lambda = 2\frac{C^*}{m}\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)\{\tau\vee(1-\tau)\}}\sqrt{\frac{p+m}{n}}. \tag{3.7}$$

Define for any $\kappa>0$,

$$g_n(\kappa) \stackrel{\text{def}}{=} \kappa(\tau\vee\{1-\tau\})^2\frac{nm}{2}. \tag{3.8}$$

Sometimes we write $g_n(\kappa)=g_n$. The constant $g_n(\kappa)$ is the smoothing error, and κ controls the level of smoothing, as explained in Section 2.1. In Algorithm 1 we recommend $\kappa=\epsilon/(2mn)$, but we allow for other choices. Define

$$\widetilde{\nu}_\tau(a) \stackrel{\text{def}}{=} \frac{3}{8}\frac{\underline{f}}{\overline{f}'}\inf_{\substack{\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},a)\\ \mathbf{\Delta}\neq 0}}\frac{\left(\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^2]\right)^{3/2}}{\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^3]}, \tag{3.9}$$

which controls the strict convexity of $Q_\tau(\mathbf{S})$.

Lemma 3.5. *Under assumptions (A1)-(A3), λ is set as* (3.7). *Let $\mathbf{\Gamma}_{\tau,\infty}$ be the minimizer* *of* $\widetilde{L}_\tau(\mathbf{S})$ *defined in* (2.5). *Under the condition on* r*:*

$$\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{mn}}+\sqrt{C_2(c_3)g_n(\kappa)}<\widetilde{\nu}_\tau(g_n), \quad (3.10)$$

then with probability greater than $1-\gamma_n-16(pm)^{1-c_3^2}-3\exp\{-(p+m)\log 8\}$,

$$\|\mathbf{\Gamma}_{\tau,\infty}-\mathbf{\Gamma}\|_{L_2(P_X)}\le 4\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{mn}}+4\sqrt{C_2(c_3)g_n(\kappa)} \quad (3.11)$$

$\|\mathbf{\Gamma}_{\tau,\infty}-\mathbf{\Gamma}\|_{\mathrm{F}}\le\sqrt{m/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\mathbf{\Gamma}_{\tau,\infty}-\mathbf{\Gamma}\|_{L_2(P_X)}$, *where* $C_\tau(c_3)=16\sqrt{\log 8\{\tau\vee(1-\tau)\}/C'}+32\sqrt{2}c_3$, C' *and* c_2 *are absolute constants given by Lemma S.4.3 in the supplementary material and Assumption (A2);* $C_2(c_3)=(4\underline{f})^{-1}(c_3C_1\sqrt{B_p/\sigma_{\max}(\mathbf{\Sigma}_X)}+3)$ *where* C_1 *is a universal constant.* $r=\mathrm{rank}(\mathbf{\Gamma})$ *and* $g_n(\kappa)$ *is defined in* (3.8).

See Section S.2.2 for a proof of Lemma 3.5. When the level of smoothness $g_n(\kappa)\to 0$ (or when $\kappa\to 0$), the bound (3.11) converges to the oracle bound of $\widehat{\mathbf{\Gamma}}$ (A.6) in Theorem A.2. The key ingredient in the proof is a new tail probability bound for the empirical process $\mathbb{G}_n\{\widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta})-\widehat{Q}_\tau(\mathbf{\Gamma})\}$, which builds on a sharp bound for the spectral norm of a partial sum of random matrices. See Maurer and Pontil (2013) and Tropp (2011) for more details of such a bound.

Define

$$h_n(\kappa)\stackrel{\text{def}}{=}4\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma^2_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{n}} + 4\sqrt{C_2(c_3)m\sigma_{\min}(\mathbf{\Sigma}_X)^{-1}g_n(\kappa)} \quad (3.12)$$

which is essentially the convergence rate of $\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}\|_{\mathrm{F}}$. Moreover, define

$$a_{n,t}(\kappa,\epsilon) \stackrel{\text{def}}{=} \kappa(\tau \vee \{1-\tau\})^2 mn + \frac{8c_2^2(\|\mathbf{\Gamma}\|_{\mathrm{F}}^2 + h_n^2)\sigma_{\max}^2(\mathbf{\Sigma}_X)}{(t+1)^2\epsilon m}. \tag{3.13}$$

$a_{n,t}(\kappa,\epsilon)$ is related to the algorithmic convergence rate (2.11).

Theorem 3.6. *Under assumptions (A1)-(A3), and λ is set as (3.7). Let $\{\mathbf{\Gamma}_{\tau,t}\}_{t=1}^{T}$ be a sequence generated by Algorithm 1. Under the growth condition of r,*

$$\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p + \log m)}{mn}} + \sqrt{C_2(c_3)a_{n,t}(\kappa,\epsilon)} < \widetilde{\nu}_\tau(a_{n,t}(\kappa,\epsilon)), \tag{3.14}$$

then with probability greater than $1 - 2\gamma_n - 32(pm)^{1-c_3^2} - 6\exp\{-(p+m)\log 8\}$,

$$\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{L_2(P_X)} \le 4\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{mn}} + 4\sqrt{C_2(c_3)a_{n,t}(\kappa,\epsilon)}, \tag{3.15}$$

$\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}} \le \sqrt{m/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{L_2(P_X)}$, *where* $C_\tau(c_3) = 16\sqrt{\log 8\{\tau \vee (1-\tau)\}/C'} + 32\sqrt{2}c_3$, *$C'$ and c_2 are absolute constants given by Lemma S.4.3 in the supplementary material and Assumption (A2); $C_2(c_3) = (4\underline{f})^{-1}(c_3C_1\sqrt{B_p/\sigma_{\max}(\mathbf{\Sigma}_X)}+3)$ where C_1 is a universal constant. $r = \mathrm{rank}(\mathbf{\Gamma})$ and $a_{n,t}(\kappa,\epsilon)$ is defined in* (3.13).

See Section S.2.4 for a proof of Theorem 3.6. In the first term in (3.15), there are three main components in (A.6), which correspond to the rank, covariates $\boldsymbol{X}$ and conditional density of Y given $\boldsymbol{X}$. When p and m are fixed with respect to n, the errors decrease in $n^{-1/2}$. However, the error will diverge to infinity if p or m grows faster than n, which corresponds to the result for the multivariate regression for mean, see Negahban and Wainwright (2011), Koltchinskii et al. (2011) among others. $r(p+m)$ can be interpreted as the true number of unknown parameters. The covariates can influence the bounds (A.6) through the condition number $\sigma_{\max}(\mathbf{\Sigma}_X)/\sigma_{\min}(\mathbf{\Sigma}_X)$ of the covariance matrix $\mathbf{\Sigma}_X$ and B_p. The estimation at τ close to 0 or 1 is difficult as $\tau \vee (1-\tau)$ grows when τ moves away from 0.5. For the second term on the right hand side of (3.15), $a_{n,t}(\kappa, \epsilon)$ can be made small by choosing ϵ, κ small and increasing t, and the bound (3.15) would be close to (A.6).

Remark 3.7 (Comment on $\widetilde{\nu}$)**.** *In Lemma 3.5 and Theorem 3.6, the growth conditions* (3.10) *and* (3.14) *are crucial for guaranteeing the strong convexity of* $Q_\tau(\mathbf{S})$*. It is easy to see that* $\widetilde{\nu}_\tau(a_{n,t}(\kappa,\epsilon)) < \widetilde{\nu}_\tau(g_n)$ *since* $a_{n,t}(\kappa,\epsilon) > g_n$ *and* $\mathcal{K}(\mathbf{\Gamma}, g_n) \subset \mathcal{K}(\mathbf{\Gamma}, a_{n,t}(\kappa,\epsilon))$*. We note that* $\widetilde{\nu}_\tau(0)$ *is related to the "restricted nonlinearity constant" in the Lasso for quantile regression of Belloni and Chernozhukov (2011). In Section S.4.1, we discuss these growth conditions in more detail.*

Remark 3.8 (Not exactly sparse $\mathbf{\Gamma}$)**.** *When* $\mathbf{\Gamma}$ *is not exactly sparse in rank (the number of nonzero singular values is not sparse), we may characterize the error by using the devise of Negahban et al. (2012). Let* $\mathcal{V} \subset \mathbb{R}^m$ *and* $\mathcal{U} \subset \mathbb{R}^p$ *be two subspaces with dimension* r*, let* $\mathcal{M} = \{\mathbf{\Delta} \in \mathbb{R}^{p\times m} : \text{row space of } \mathbf{\Delta} \subset \mathcal{V},\ \text{column space of } \mathbf{\Delta} \subset \mathcal{U}\}$*;* $\overline{\mathcal{M}}^{\perp} = \{\mathbf{\Delta} \in \mathbb{R}^{p\times m} : \text{row space of } \mathbf{\Delta} \subset \mathcal{V}^{\perp},\ \text{column space of } \mathbf{\Delta} \subset \mathcal{U}^{\perp}\}$ *(defined similarly as in Example 3 on page 542 of Negahban et al. (2012)). For any matrix* $\mathbf{S} \in \mathbb{R}^{p\times m}$*,*

$$\mathcal{P}_{\mathcal{M}}(\mathbf{S}) = \mathbf{P}_{\mathcal{U}}\mathbf{S}\mathbf{P}_{\mathcal{V}}, \quad \mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{S}) = \mathbf{P}^{\top}_{\mathcal{U}}\mathbf{S}\mathbf{P}^{\top}_{\mathcal{V}},$$

where $\mathbf{P}_{\mathcal{V}} = \mathbf{V}\mathbf{V}^{\top}$*,* $\mathbf{P}^{\perp}_{\mathcal{V}} = \mathbf{I}_{m\times r} - \mathbf{P}_{\mathcal{V}}$*,* $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_r]$*, and* $\{\mathbf{v}_j\}_{j=1}^r$ *is a set of orthonormal basis for* $\mathcal{V}$*; analogously,* $\mathbf{P}_{\mathcal{U}} = \mathbf{U}\mathbf{U}^{\top}$*,* $\mathbf{P}^{\perp}_{\mathcal{U}} = \mathbf{I}_{p\times r} - \mathbf{P}_{\mathcal{U}}$*,* $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_r]$*, and* $\{\mathbf{u}_j\}_{j=1}^r$ *is a set of orthonormal basis for* $\mathcal{U}$*. Moreover, we have the decomposability: for any matrix* $\mathbf{S}$*,* $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathcal{M}}(\mathbf{S})\|_* + \|\mathcal{P}^{\top}_{\overline{\mathcal{M}}}(\mathbf{S})\|_*$.

It can be shown that when $\lambda \geq 2\|\nabla\widehat{Q}(\mathbf{\Gamma})\|$*, with probability greater than* $1 - \gamma_n -$

$16(pm)^{1-c_3^2} - 3\exp\{-(p+m)\log 8\}$, *the difference* $\mathbf{\Delta}_{\tau,t} = \mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}$ *lies in the set*

$$\begin{aligned}&\mathcal{K}(\overline{\mathcal{M}}, 4\|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Gamma})\| + 2a_{n,t}(\kappa,\epsilon)/\lambda)\\ &\qquad \overset{\text{def}}{=} \left\{\mathbf{\Delta} \in \mathbb{R}^{p\times m} : \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Delta})\| \le 3\|\mathcal{P}_{\overline{\mathcal{M}}}(\mathbf{\Delta})\| + 4\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma})\| + \frac{2b_{n,t}(\kappa,\epsilon)}{\lambda}\right\},\end{aligned} \tag{3.16}$$

where $b_{n,t}(\kappa,\epsilon) > 0$ *is an appropriately adapted version of* $a_{n,t}(\kappa,\epsilon)$ *for* $\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma})\|$. *The oracle property of* $\mathbf{\Gamma}_{\tau,t}$ *can be shown via similar argument as showing Theorem 3.6, and we leave out the detail. The proof for* (3.16) *is in Section S.4.2.*

3.2. Realistic Bounds for Factors and Loadings

In this section we discuss the bounds for the estimated factors and loadings, defined in (2.8). The bounds will be stated in terms of $\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}}$, and then Theorem 3.6 can be applied for finding the explicit rate for the factors and loadings.

First we observe that by Mirsky's theorem, the singular values can be consistently estimated.

Lemma 3.9. *Let* $\{\mathbf{\Gamma}_{\tau,t}\}_{t=1}^{T}$ *be a sequence generated by Algorithm 1, then for any* t,

$$\sum_{j=1}^{p\wedge m}\left\{\sigma_j(\mathbf{\Gamma}_{\tau,t}) - \sigma_j(\mathbf{\Gamma})\right\}^2 \le \|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}}^2. \tag{3.17}$$

The proof of Lemma 3.9 is a straightforward application of Mirsky's theorem (see, e.g., Theorem 4.11 on page 204 of Stewart and Sun (1990)). The detail is omitted.

Theorem 3.10. *If the nonzero singular values of matrix* $\mathbf{\Gamma}_\tau$ *are distinct, then with the choice of* $\widehat{\mathbf{\Psi}}_\tau$ *and* $\widehat{f}_k^{\tau}(\boldsymbol{X}_i)$ *in* (2.8) *for a given* t,

$$1 - |(\widehat{\mathbf{\Psi}}_\tau)_{*j}^{\top}(\mathbf{\Psi}_\tau)_{*j}| \le \frac{2(2\|\mathbf{\Gamma}\| + \|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}})\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}}}{\min\{\sigma_{j-1}^2(\mathbf{\Gamma}) - \sigma_j^2(\mathbf{\Gamma}), \sigma_j^2(\mathbf{\Gamma}) - \sigma_{j+1}^2(\mathbf{\Gamma})\}} \tag{3.18}$$

If, in addition, let the SVDs $\mathbf{\Gamma}_\tau = \mathbf{U}_\tau \mathbf{D}_\tau \mathbf{V}_\tau^\top$ *and* $\mathbf{\Gamma}_{\tau,t} = \widehat{\mathbf{U}}_\tau \widehat{\mathbf{D}}_\tau \widehat{\mathbf{V}}_\tau^\top$*, suppose* $(\widehat{\mathbf{U}}_\tau)_{*j}^\top (\mathbf{U}_\tau)_{*j} \geq 0$*, then*

$$\left|\widehat{f}_k^\tau(\boldsymbol{X}_i) - f_k^\tau(\boldsymbol{X}_i)\right| \leq \|\boldsymbol{X}_i\| \left(\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}} + 2\sigma_k(\mathbf{\Gamma}) \sqrt{\frac{(2\|\mathbf{\Gamma}\| + \|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}})\|\mathbf{\Gamma}_{\tau,t} - \mathbf{\Gamma}\|_{\mathrm{F}}}{\min\{\sigma_{k-1}^2(\mathbf{\Gamma}) - \sigma_k^2(\mathbf{\Gamma}), \sigma_k^2(\mathbf{\Gamma}) - \sigma_{k+1}^2(\mathbf{\Gamma})\}}} \right) \tag{3.19}$$

See Section S.2.6 for a proof for Theorem 3.10. The oracle inequalities in Theorem 3.6 can then be applied to find the exact rate for the loadings and factors.

Remark 3.11. *The condition* $(\widehat{\mathbf{U}}_\tau)_{*j}^\top (\mathbf{U}_\tau)_{*j} \geq 0$ *essentially says that the sign of* $(\widehat{\mathbf{U}}_\tau)_{*j}$ *is correctly chosen, which can usually be done in practice. See Remark 2.1 for more discussion.*

Remark 3.12 (Repeated singular values)**.** *Theorem 3.10 is under the condition that the singular values for* $\mathbf{\Gamma}$ *are distinct. If there are repeated singular values, then the corresponding singular vectors are not uniquely defined, and we can only obtain a bound for the "canonical angle" (see, e.g., Yu et al. (2015)) of the subspaces generated by the singular vectors associated with the repeated singular values.*

4. Simulation

In this section, we check the performance of the proposed method via Monte Carlo experiments, and compare with an oracle estimator computed under the knowledge of the true rank.

Given two distinct matrices $\mathbf{S}_1, \mathbf{S}_2$ with *nonnegative* entries, $\mathrm{rank}(\mathbf{S}_1) = r_1$ and $\mathrm{rank}(\mathbf{S}_2) = r_2$, we simulate data from the two-piece normal model (Wallis; 2014)

$$Y_{ij} = \Phi_\sigma^{-1}(U_{ij}) \boldsymbol{X}_i^\top \big((\mathbf{S}_1)_{*j} \mathbf{1}\{U_{ij} \leq 0.5\} + (\mathbf{S}_2)_{*j} \mathbf{1}\{U_{ij} > 0.5\} \big), \tag{4.1}$$

$$i = 1, ..., n = 500; \; j = 1, ..., m = 300,$$

U_{ij} are i.i.d. $U(0,1)$ independent of $\boldsymbol{X}_i$. $\boldsymbol{X}_i \in \mathbb{R}^p$ follows a multivariate $U([0,1])$ distribution for $p = 300$ with covariance matrix $\boldsymbol{\Sigma}$ in which $\boldsymbol{\Sigma}_{ij} = 0.1 * 0.8^{|i-j|}$ for $j = 1, ..., p$. See Falk (1999) for more details on simulating $\boldsymbol{X}_i$. The conditional quantile function $q_j(\tau|\boldsymbol{x})$ of Y_{ij} on $\boldsymbol{x}$ for the distribution of Y_{ij} is

$$q_j^l(\tau|\boldsymbol{x}) = \Phi^{-1}(\tau)\boldsymbol{x}^\top\big(\mathbf{S}_1\mathbf{1}\{\tau \le 0.5\} + \mathbf{S}_2\mathbf{1}\{\tau > 0.5\}\big) \stackrel{\text{def}}{=} \boldsymbol{x}^\top(\boldsymbol{\Gamma}_\tau)_{*j}, \tag{4.2}$$

where $\boldsymbol{\Gamma}_\tau$ is defined in an obvious manner. The number of repetitions is 500.

In our simulation study, we fix $\mathbf{S}_1$ with $\text{rank}(\mathbf{S}_1) = 2$. However, we consider two models for $\mathbf{S}_2$:

I. Model ES (equally sparse): $\mathbf{S}_2^{ES}$ with $\text{rank}(\mathbf{S}_2^{ES}) = 2$;

II. Model AS (asymmetrically sparse): $\mathbf{S}_2^{AS}$ with $\text{rank}(\mathbf{S}_2^{AS}) = 6$.

The entries of $\mathbf{S}_1$, $\mathbf{S}_2^{ES}$ and $\mathbf{S}_2^{AS}$ will be randomly selected. The specific steps for generating these matrices are detailed in Section S.4.3. We only note here that the singular values of matrices $\mathbf{S}_1$ and $\mathbf{S}_2^l$ for $l \in \{ES, AS\}$ are randomly selected and are all distinct.

We apply Algorithm 1 with $\tau = 5\%, 10\%, 20\%, 80\%, 90\%, 95\%$ to compute the estimator $\widehat{\boldsymbol{\Gamma}}_\tau^l$ for $\boldsymbol{\Gamma}_\tau^l$, defined in (4.2), where $l \in \{ES, AS\}$. The tuning parameter λ is selected as described in Section 2.3. We stop the algorithm when the change in the loss function $L_\tau(\mathbf{S})$ (defined in (2.1)) from two consecutive iterations is less than 10^{-6}. The performance of $\widehat{\boldsymbol{\Gamma}}_\tau^l$ is measured by the Frobenius error: $\|\boldsymbol{\Gamma}_{\tau,\lambda}^l - \widehat{\boldsymbol{\Gamma}}_\tau^l\|$, for $l \in \{ES, AS\}$. The results for prediction error have similar pattern as the Frobenius error, so we do not report them here. We also report the average number of iterations for running Algorithm 1. The error of $\widehat{\boldsymbol{\Gamma}}_\tau^l$ is compared with that of an oracle estimator computed using the knowledge of true rank r_1 or $r_2^l \stackrel{\text{def}}{=} \text{rank}(\mathbf{S}_2^l)$ depending on τ (or $l \in \{ES, AS\}$). The oracle estimator is computed in a similar way as Algorithm 1, while we replace the soft thresholding operator S_λ by a hard thresholding operator, which truncates all but the first r_1 or r_2^l singular values to 0. The iteration stops when the change in the function $\widehat{Q}_\tau(\mathbf{S})$ is less than 10^{-6}.

The mean and standard deviation of the Frobenius errors is in Table 4.2. When the variance is larger ($\sigma = 1$), we have greater errors as expected. The errors vary with τ, which is almost 2 times higher when τ is close to 0.05 and 0.95 than when τ is 0.2 and 0.8. If we compare the error of $\widehat{\mathbf{\Gamma}}_\tau^l$, for $l \in \{ES, AS\}$ to that of the the oracle estimator, the oracle estimators always have smaller errors for all τ. However, their difference is at most around 5-10% of the oracle error. In addition, the standard deviation of the oracle Frobenius error is also less than that of $\widehat{\mathbf{\Gamma}}_\tau^l$.

When we compare the errors of the two models ES and AS, we find that their errors are compatible when τ is less than 0.5. Nonetheless, when τ is greater than 0.5, the errors of the model AS is around $\sqrt{r_2^{AS}/r_2^{ES}} = \sqrt{6/2} \approx 1.732$ times of that of the model ES. The oracle estimator also shows a similar pattern. This is consistent with our error bounds, which predicts that the model with a larger rank would have greater errors.

The mean of number of iterations is reported in Table 4.1. More iterations are required when τ is close to 0 and 1 and when σ is larger. Estimating $\widehat{\mathbf{\Gamma}}_\tau^l$ for $l = AS$ requires more iterations than for $l = ES$, when τ is greater than 0.5. The pattern coincides with the algorithmic convergence analysis in Section 2.4.

Table 4.1: Averaged number of iterations.

τ	0.05	0.1	0.2	0.8	0.9	0.95
			$\sigma = 0.5$			
ES	20.9	18.0	16.0	16.0	18.0	20.3
AS	20.8	18.0	16.0	23.0	25.1	28.7
			$\sigma = 1$			
ES	26.5	23.0	21.0	20.6	23.0	26.0
AS	26.5	23.1	21.0	29.1	32.9	37.1

Table 4.2: Averaged Frobenius errors with standard deviations. "Or." denotes the oracle estimator, which is estimated under the knowledge of true rank. The numbers in parentheses are standard deviations of the errors.

τ	0.05	0.1	0.2	0.8	0.9	0.95
			$\sigma = 0.5$			
ES	60.995	48.746	34.302	33.973	48.375	60.604
	(0.253)	*(0.227)*	*(0.209)*	*(0.202)*	*(0.217)*	*(0.247)*
ES Or.	57.261	44.926	30.006	29.853	44.735	57.007
	(0.191)	*(0.152)*	*(0.116)*	*(0.118)*	*(0.152)*	*(0.184)*
AS	60.978	48.724	34.289	60.487	85.997	108.310
	(0.263)	*(0.220)*	*(0.207)*	*(0.539)*	*(0.567)*	*(0.820)*
AS Or.	57.239	44.911	30.002	54.922	80.583	102.663
	(0.202)	*(0.164)*	*(0.120)*	*(0.744)*	*(0.464)*	*(0.572)*
			$\sigma = 1$			
ES	118.245	93.419	64.289	63.634	92.519	117.365
	(0.570)	*(0.420)*	*(0.387)*	*(0.382)*	*(0.372)*	*(0.438)*
ES Or.	113.636	88.781	58.913	58.593	88.365	113.099
	(0.427)	*(0.338)*	*(0.238)*	*(0.221)*	*(0.301)*	*(0.378)*
AS	118.259	93.434	64.291	120.338	170.904	217.185
	(0.530)	*(0.412)*	*(0.380)*	*(1.151)*	*(1.273)*	*(1.547)*
AS Or.	113.647	88.788	58.911	108.754	161.303	205.371
	(0.387)	*(0.308)*	*(0.224)*	*(0.711)*	*(0.929)*	*(1.188)*

Remark 4.1. *If the true rank is known, an alternative approach to compute the oracle estimator is to apply the classical quantile regression equation with Y_{ij} on $\boldsymbol{X}_i$ to get a primary estimator for $\boldsymbol{\Gamma}$, and then truncate all but r_1 or r_2^l singular values of the primary estimator to attain low rankness. However, this gives huge Frobenius and prediction errors, and we do not report the results here.*

5. Empirical Analysis

In this section, we use our method to study important scientific problems in finance and climatology. Section 5.1 is devoted to spatial clustering based on extreme temperature. In Section 5.2, we analyze global financial risk. To keep our discussion brief, we omit "τ-

quantile" when it does not cause confusion; for example, the expression "τ-quantile of Y_j has high loading in $f_1^\tau(X_i)$" will be shortened to "Y_j has high loading in $f_1^\tau(X_i)$".

5.1. Spatial Clustering with Extreme Temperature

Spatial clustering is particularly crucial for modern climatological modeling in a data-rich environment, where the size of a grid can be very large. In a relevant study, Bador et al. (2015) construct spatial clusters in Europe that visualize the spatial dependence in extreme high temperature in summer. They argue that mean and correlation based methods fail to capture such distributional features of extreme events. In this section, we apply our method to a daily temperature data set of the year 2008 from $m = 159$ weather stations around China, which is downloaded from the website of Research Data Center of CRC 649 of Humboldt-Universität zu Berlin. The ideas and technique we demonstrate in this section can be applied on even larger data with big m.

Let Y_{ij} be the temperature (in Celsius) at j weather station on i day, where $i = 1, ..., n = 365$ and $j = 1, ..., m$. Before applying our method, we remove the common mean of Y_{ij} by fitting a curve with typical smoothing spline, see Section S.4.4 for more details. In Figure 5.1, the lower left subfigure is the fitted mean curve, which shows a seasonal pattern. After removing the mean, the temperature curves of 159 weather stations are shown in the upper left panel of Figure 5.1. We note that the de-trended curves also demonstrate seasonality: the dispersion is larger in winter than in summer.

We apply Algorithm 1 on the de-trended temperature curves. Let b_l, $l = 1, ..., p$ be B-spline basis functions with equally distributed knots on $[0, 1]$ interval, we choose $\boldsymbol{X}_i = (b_1(i/365), ..., b_p(i/365))$ for $i = 1, ..., 365$. The number of basis function is selected as $p = \lceil n^{2/5} \rceil = 11$, which is slightly larger than the rate suggested by the asymptotic theory if we assume the curves are smooth. We take $\tau = 1\%$ and 99%. The tuning parameter λ is selected by the method in Section 2.3, and the estimated value is $\lambda = 0.000156$.

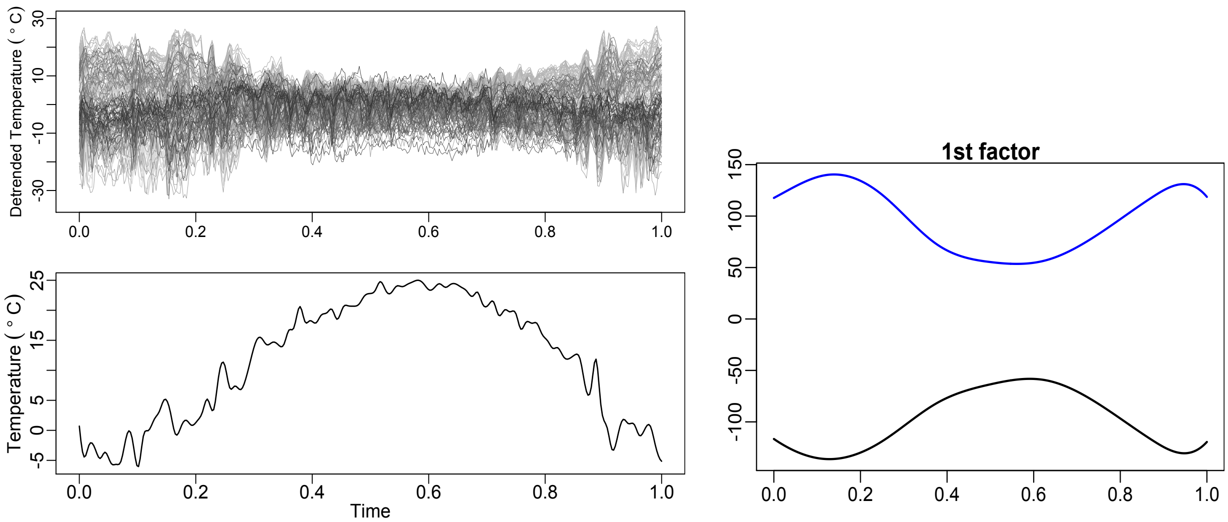

Figure 5.1: Upper left panel: The temperature time series in excess to national mean of the 159 weather stations around China; Lower left panel: the fitted temperature common mean curve estimated by smoothing spline; Right panel: The plot for the first factor, in which the black lines corresponds to 1% quantile factors and the blue lines corresponds to 99% quantile factors.

The right panel in Figure 5.1 presents the first factors $f_1^{0.01}(X_i)$ and $f_1^{0.99}(X_i)$. The two factors enclose a region that is wide in the ends and narrow in the middle. This is related to the fact that the dispersion in temperature among weather stations tends to be higher in winter and lower in summer, as shown in the upper left panel in Figure 5.1. The other factors are rather small in absolute value relative to the first factor, so we do not include them in the analysis for brevity.

The upper left (right) panels in Figure 5.2 show the locations of the weather stations, and the color corresponds to the magnitude of the factor loadings to $f_1^{0.01}(X_i)$ ($f_1^{0.99}(X_i)$). In the upper left panel in Figure 5.2, stations in northeastern China are highly associated with the factor $f_1^{0.01}(X_i)$, while the stations in southern China have zero or even slightly negative association to the factor $f_1^{0.01}(X_i)$. The upper right panel in Figure 5.2 show the opposite pattern to the factor $f_1^{0.99}(X_i)$. These loadings quantify the spatial correlation in extremely high (0.99 quantile) or low (0.01 quantile) temperatures at these weather stations, which provides a foundation for spatial clustering. However, the cutoff points of the loadings for

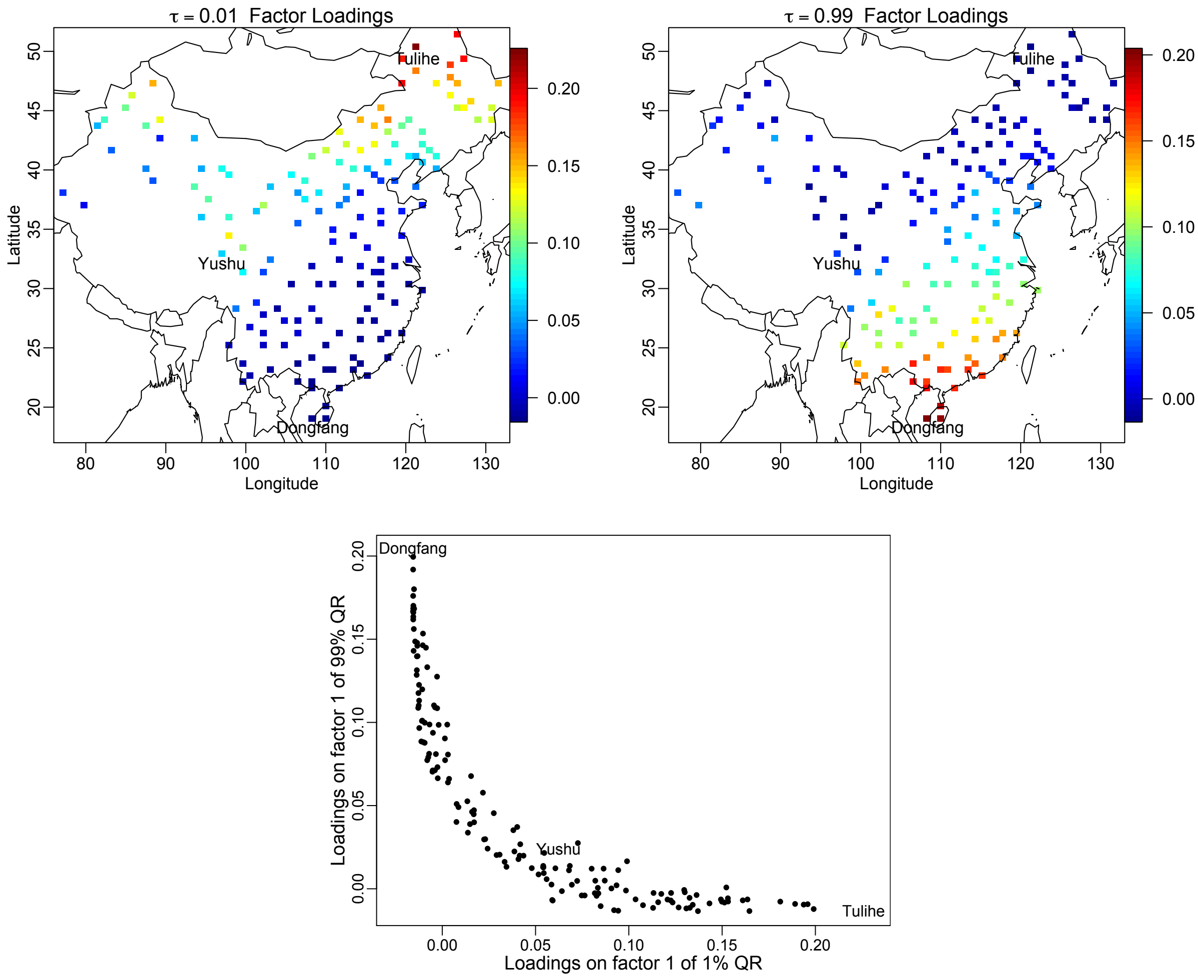

Figure 5.2: Upper panels: plot of the locations of weather stations. The color scale corresponds to the magnitude of their $\tau = 0.01$ (left) and $\tau = 0.99$ factor loading. Lower panel: tail to tail plot for temperature data. Each point is a pair $((\widehat{\Psi}_{0.01})_{1j}, (\widehat{\Psi}_{0.99})_{1j})$ for weather stations $j = 1, ..., 159$.

determining the clusters have to be carefully chosen, which we leave for future study.

The tail to tail plot in Figure 5.2 showing the loadings for the first factor at $\tau = 1\%$ and 99% demonstrates a nearly "L" shape, which shows that the temperature of each station seems to be associated with either the lower tail factor or the upper tail factor, but not both. We highlight three stations in Tulihe, Dongfang, and Yushu which are located in the far right, far top, and center in Figure 5.2.

5.2. Global Financial Risk

Quantifying global financial risk in a high-dimensional setting is a very challenging task. White et al. (2015) estimate the lower quantiles ($\tau = 0.01$) of stock returns from $m = 230$ largest global financial firms with a vector autoregressive (VAR) model, and show that stock returns of the firms with *large market value* and *high leverage* tend to be more vulnerable to systemic shock. However, their method does not scale up to high dimensionality because of excessive computational cost, so they estimate bivariate VAR for the quantiles $(q_{Y_{ij}}(\tau), q_{M_i}(\tau))$ for each stock return Y_j, where M_i is a global market index. In the sequel, we analyze *all* stocks jointly and compare our findings with the results of White et al. (2015).

We analyze the same set of daily stock closing prices as White et al. (2015) with the same time frame from January 1, 2000 to August 6, 2010. The dataset is downloaded from Dr. Manganelli's personal website. See Table 1 of White et al. (2015) for a detailed breakdown of the stocks by sector and country, as well as their averaged market value and leverage (the ratio of short and long term debt over common equity) over the data period. We use daily log-returns of the stock closing prices and this results in $n = 2765$.

We consider a multivariate model which jointly incorporates multiple asset returns. Let $Y_{i,j}$ be the asset return for j firm, where $j = 1, ..., m$ and $i = 1, ..., n$. We consider $q_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top(\boldsymbol{\Gamma}_\tau)_{*j}$, where

$$\boldsymbol{X}_i = (|Y_{i-1,1}|, ..., |Y_{i-1,m}|, Y_{i-1,1}^-, ..., Y_{i-1,m}^-)^\top \in \mathbb{R}^{2m}, \tag{5.1}$$

and $Y^- \stackrel{\text{def}}{=} \max\{-Y, 0\}$. The choice of $\boldsymbol{X}_i$ aims to capture asymmetric contribution of lag return to the quantile of stock price, which is suggested in the Conditional Autoregressive Value-at-Risk (CAViaR) literature, see Engle and Manganelli (2004). We estimate $\boldsymbol{\Gamma}$ via the nuclear norm regularized multivariate quantile regression with $\tau = 0.01$ and 0.99. We estimate the factor and loadings as (2.8) in Section 2.2. To select the tuning parameter λ, applying the procedure described in Section 2.3 gives $\lambda = 0.02468$ for $\tau = 0.01$. By symmetry we also apply the same λ for $\tau = 0.99$.

We present the estimated first factors for the quantile regression at $\tau = 0.01$ and 0.99 in Figure 5.3. The other factors are very small in scale compared to the first factor. Both first factors $f_1^{0.01}(X_i)$ and $f_1^{0.99}(X_i)$ are volatile and moving away from 0 at the end of 2008 and in the first quarter of 2009, and mid 2010, which corresponds to the periods of financial crisis and European debt crisis. In later analysis, we treat $f_1^{0.01}(X_i)$ as an indicator for global financial risk.

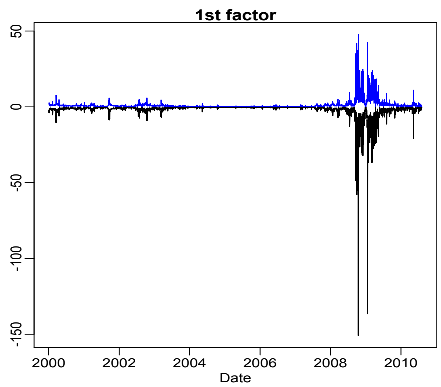

Figure 5.3: The time series plots for the first factor. The black lines correspond to 0.01 quantile factors and the blue lines correspond to 0.99 quantile factors.

The left panel of Figure 5.4 is the "tail to tail" plot with $\tau = 0.01$ and 0.99, in which each point is the pair of loadings $((\widehat{\Psi}_{0.01})_{1j}, (\widehat{\Psi}_{0.99})_{1j})$ defined in (2.8), for $j = 1, ..., 230$. The values $((\widehat{\Psi}_{0.01})_{1j}, (\widehat{\Psi}_{0.99})_{1j})$ are all positive. The fact that they distribute around the reverse diagonal line suggests that the log-returns of these stocks are roughly equally associated to the two extreme quantile factors. However, we observe that the points become more disperse and deviate from the reverse diagonal line when moving northeast.

The right panel of Figure 5.4 plots the firms based on their averaged market value and leverage, and the color scale depends on the magnitude of the $\tau = 0.01$ factor loading of the corresponding stock. Our finding shows that high loadings associated with $f_1^{0.01}(X_i)$ are usually found for those stocks whose underlying firms have large market value and high leverage, which aligns with the results of White et al. (2015). In particular, as shown by the right panel of Figure 5.4, the firms with certain combinations of market value and leverage tend to have high loading associated with $f_1^{0.01}(X_i)$ in 0.01 quantile of their stock returns. This seems to be an interesting direction for future study.

Lastly, we note that the algorithmic convergence results in Section 2.4 apply straightforwardly on financial time series data. However, an extension of the theory in Section 3.1 may be required in order to bound the estimation error for the time series data.

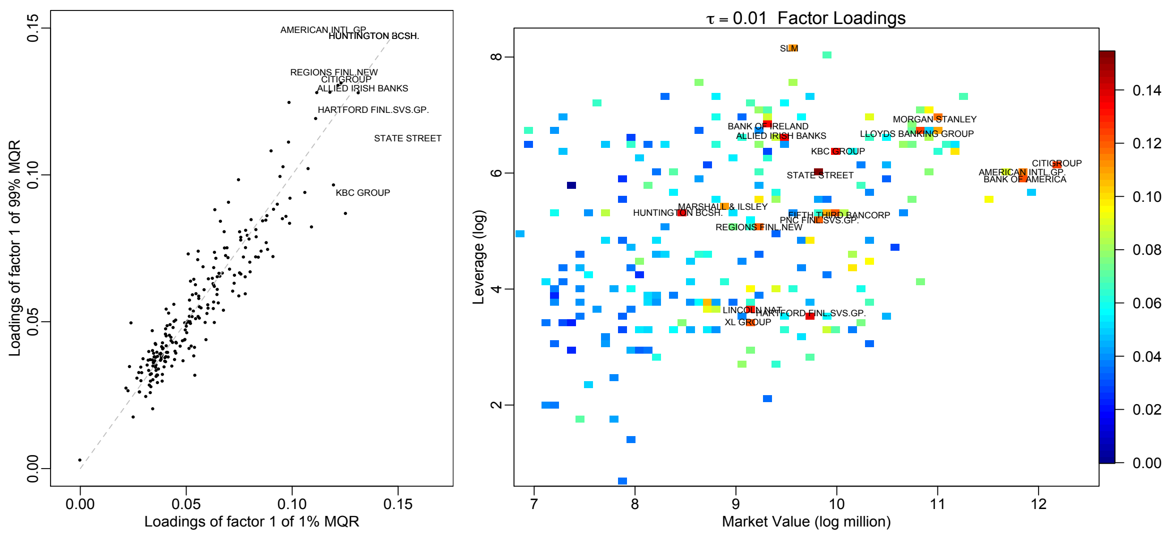

Figure 5.4: Left panel: tail to tail plot. Each point is a pair $((\widehat{\Psi}_{0.01})_{1j}, (\widehat{\Psi}_{0.99})_{1j})$ for stocks $j = 1, ..., 230$; Right panel: the plot of firms based on their averaged market value and leverage over the data period. The color scale corresponds to the magnitude of their $\tau = 0.01$ factor loading.

APPENDIX: Oracle Properties for Exact Optimizer $\widehat{\mathbf{\Gamma}}$

In this section, we present the bounds for the exact minimizer $\widehat{\mathbf{\Gamma}}$ for (1.4). Though $\widehat{\mathbf{\Gamma}}$ is difficult to obtain in practice and is therefore not very useful, it is however very pedagogical to study the bounds of $\widehat{\mathbf{\Gamma}}$, as many ideas applied there will be crucial for proving our main results.

For a this section, define $\nu_\tau \overset{\text{def}}{=} \widetilde{\nu}_\tau(0)$. Note that $\nu_\tau \le \widetilde{\nu}_\tau(g_n) \le \nu_\tau(a_{n,t}(\kappa, \epsilon))$.

Lemma A.1. *Under assumptions (A1)-(A3), $\lambda \geq 2\|\nabla\widehat{Q}(\mathbf{\Gamma})\|$ and the growth condition on r:*

$$\underline{f}^{-1}\sqrt{m}\left(32\sqrt{2}c_3\sqrt{r}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{\frac{\log m+\log p}{n}}+\lambda\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\right)<\nu_\tau, \tag{A.1}$$

where c_2 is an absolute constant given by Assumption (A2). Then

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)} \leq 128\sqrt{2}c_3\underline{f}^{-1}\sqrt{r}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{\frac{\log m+\log p}{n}}+\lambda\frac{4\sqrt{2}}{\underline{f}}\sqrt{\frac{mr}{\sigma_{\min}(\mathbf{\Sigma}_X)}}, \tag{A.2}$$

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{\mathrm{F}} \leq 16\sqrt{2}\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)}{\sigma^2_{\min}(\mathbf{\Sigma}_X)}}\sqrt{\frac{r\log(p+m)}{n}}+\lambda\frac{16\sqrt{2}}{\underline{f}}\frac{\sqrt{r}m}{\sigma_{\min}(\mathbf{\Sigma}_X)} \tag{A.3}$$

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{*} \leq 128\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)}{\sigma^2_{\min}(\mathbf{\Sigma}_X)}}\sqrt{\frac{\log(p+m)}{n}}r+\lambda\frac{128}{\underline{f}}\frac{mr}{\sigma_{\min}(\mathbf{\Sigma}_X)} \tag{A.4}$$

$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{\mathrm{F}} \leq \sqrt{m/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)}$ *and* $\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{*} \leq 4\sqrt{rm/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)}$, *with probability greater than $1-16(pm)^{1-c_3^2}-\gamma_n$, where $r=\mathrm{rank}(\mathbf{\Gamma})$.*

Please see Section S.3.1 for a proof of Lemma A.1.

Theorem A.2. *Assume that assumptions (A1)-(A3) hold and select λ as (3.7). Under the growth condition on r:*

$$\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{mn}}<\nu_\tau, \tag{A.5}$$

where $C_\tau(c_3) \stackrel{\text{def}}{=} 16\sqrt{\log 8\{\tau\vee(1-\tau)\}/C'}+32\sqrt{2}c_3$, C' and c_2 are absolute constants given by Lemma S.4.3 in the supplementary material and Assumption (A2). Then

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)} \leq 4\frac{C_\tau(c_3)}{\underline{f}}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X)+B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{r}\sqrt{\frac{(m+p)(\log p+\log m)}{mn}}, \tag{A.6}$$

$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{\mathrm{F}} \leq \sqrt{m/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)}$ *and* $\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_* \leq 4\sqrt{rm/\sigma_{\min}(\mathbf{\Sigma}_X)}\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_{L_2(P_X)}$, *with probability greater than* $1-\gamma_n-16(pm)^{1-c_3^2}-3\exp\{-(p+m)\log 8\}$, *where* $r=\operatorname{rank}(\mathbf{\Gamma})$.

Please see Section S.3.2 for a proof of Theorem A.2.

Remark A.3 (Uniformity in τ). *All the bounds Theorem A.2, 3.5 and 3.6 can be made uniformly in τ by replacing the constant $\tau \vee (1-\tau)$ by 1 and keeping the rest unchanged. This is based on the observation that τ enters those bounds only through the constant $\tau \vee (1-\tau)$.*

References

Ando, T. and Tsay, R. S. (2011). Quantile regression models with factor-augmented predictors and information criterion, *Econometrics Journal* **14**: 1–24.

Bador, M., Naveau, P., Gilleland, E., Castellá, M. and Arivelo, T. (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe, *Weather and Climate Extremes* pp. 17–24.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* **2**(1): 183–202.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics* **39**(1): 82–130.

Belloni, A., Chernozhukov, V. and Fernández-Val, I. (2011). Conditional quantile processes based on series or many regressors. arXiv preprint arXiv:1105.6154.

Bernard, E., Naveau, P. and Vrac, M. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in france, *Journal of Climate* **26**(20): 7929–7937.

Bhatia, R. and Kittaneh, F. (1990). Norm inequalities for partitioned operators and an application, *Mathematische Annalen* **287**: 719–726.

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using NWP model output, *Monthly Weather Review* **132**(1): 338–347.

Briollais, L. and Durrieu, G. (2014). Application of quantile regression to recent genetic and -omic studies, *Human Genetics* **133**: 951–966.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices, *The Annals of Statistics* **39**(2): 1282–1309.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists, *Frontiers in Ecology and the Environment* **1**(8): 412–420.

Cai, J.-F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization* **20**(4): 1956–1982.

Castruccio, S., Huser, R. and Genton, M. G. (2015). High-order composite likelihood inference for max-stable distributions and processes, *Journal of Computational and Graphical Statistics* .

Chen, L., Dolado, J. J. and Gonzalo, J. (2015). Quantile factor models.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing, *Econometrica* **78**(3): 1093–1125.
URL: *http://dx.doi.org/10.3982/ECTA7880*

Davison, A., Padoan, S. A. and Ribatet, M. (2012). Statistical modeling of spatial extremes, *Statistical Science* **27**: 161–186.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves, *Journal of the Royal Statistical Society: Series B* **70**(3): 609–627.

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles, *Journal of Business & Economic Statistics* **22**: 367–381.

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlation, *Communications in Statistics - Simulation and Computation* **28**(3): 785–791.

Fan, J., Xue, L. and Zou, H. (2015). Multi-task quantile regression under the transnormal model, *Journal of the American Statistical Association* .

Friederichs, P. and Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression, *Monthly Weather Review* **135**(6): 2365–2378.
URL: *http://dx.doi.org/10.1175/MWR3403.1*

Jaggi, M. and Sulovský, M. (2010). A simple algorithm for nuclear norm regularized problems, *Proceedings of the 27th International Conference on Machine Learning*.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization, *Proceedings of the 26th International Conference on Machine Learning*.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions, *The Annals of Statistics* **26**(2): 755–770.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, New York.

Koenker, R. and Bassett, G. S. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

Koenker, R. and Hallock, K. F. (2001). Quantile regression, *Journal of Economic Perspectives* **15**(4): 143–156.

Koenker, R. and Portnoy, S. (1990). M estimation of multivariate regressions, *Journal of American Statistical Association* **85**(412): 1060–1068.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *The Annals of Statistics* **39**(5): 2243–2794.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces (Isometry and processes)*, Ergebnis der Mathematik und ihrer Grenzgebiete, Springer-Verlag.

Maurer, A. and Pontil, M. (2013). Excess risk bounds for multitask learning with trace norm regularization, *JMLR: Workshop and Conference Proceedings* **30**: 1–22.

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

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

Nesterov, Y. (2005). Smooth minimization of non-smooth functions, *Mathematical Programming* **103**(1): 127–152.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**(4): 535–553.

Reich, B. J., Fuentes, M. and Dunson, D. B. (2011). Bayesian spatial quantile regression, *Journal of American Statistical Association* **106**(493): 6–20.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression*, Springer, New York.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory*, Academic Press.

Tropp, J. A. (2011). User-friendly tail bounds for sums of random matrices, *Foundations of computational mathematics* **12**(4): 389–434.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*, Springer.

Vershynin, R. (2012a). *Compressed Sensing, Theory and Applications*, Cambridge University Press, chapter 5, pp. 210–268.

Vershynin, R. (2012b). How close is the sample covariance matrix to the actual covariance matrix?, *Journal of Theoretical Probability* **25**(3): 655–686.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), *IEEE Transactions on Information Theory* **55**: 2183–2202.

Wallis, K. F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart, *National Institute Economic Review* **167**: 106–112.

Wallis, K. F. (2014). The two-piece normal, binormal, or double Gaussian distribution: Its origin and rediscoveries, *Statistical Science* **29**(1): 106–112.

White, H., Kim, T.-H. and Manganelli, S. (2015). VAR for VaR: measuring systemic risk using multivariate regression quantiles, *Journal of Econometrics* **187**: 169–188.

Yousefi, N., Lei, Y., Kloft, M., Mollaghasemi, M. and Anagnastapolous, G. (2016). Local rademacher complexity-based learning guarantees for multi-task learning, *ArXiv Preprint Arxiv 1602.05916* .

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians, *Biometrika* **102**(2): 315–323.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society: Series B* **69**(3): 329–346.

SUPPLEMENTARY MATERIAL: FACTORISABLE MUITI-TASK QUANTILE REGRESSION

In this supplementary material, we provide the proofs and technical detail for the materials shown in the main body. Section S.1 presents the convergence analysis for the algorithm. Section S.2 presents the proof for the oracle properties of $\boldsymbol{\Gamma}_{\tau,t}$. Section S.3 contains the proof for the oracle properties of $\widehat{\boldsymbol{\Gamma}}$. Section S.4 discusses technical detail and remarks. Section S.5 lists some auxiliary results.

S.1: Proofs for Algorithmic Convergence Analysis

S.1.1. Proof of (2.2)

To see that this equation holds, note that for each pair of i, j, when $Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{*j} > 0$, $\Theta_{ij} = \tau$, since τ is the largest "positive" value in the interval $[\tau-1, \tau]$. When $Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{*j} \le 0$, $\Theta_{ij} = \tau - 1$ since τ is the smallest "negative" value in the interval $[\tau - 1, \tau]$. This verifies the equation. $\square$

Remark S.1.1. *It is necessary to choose $[\tau - 1, \tau]$ rather than $\{\tau - 1, \tau\}$ for the support of Θ_{ij} in (2.2) (though both choices fulfill the equation). The previous choice is an interval and is therefore a convex set, and the conditions given in Nesterov (2005) is fulfilled.*

S.1.2. Proof of Theorem 2.3

Recall the definition of $L_\tau(\mathbf{S})$ and $\widehat{Q}_\tau(\mathbf{S})$ in (2.1), $\widetilde{L}_\tau(\mathbf{S})$ and $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ in (2.5) and (2.3). We note a comparison property in (2.7) of Nesterov (2005), for an arbitrary $\mathbf{S} \in \mathbb{R}^{p\times m}$,

$$\widehat{Q}_{\tau,\kappa}(\mathbf{S}) \le \widehat{Q}_\tau(\mathbf{S}) \le \widehat{Q}_{\tau,\kappa}(\mathbf{S}) + \kappa \max_{\boldsymbol{\Theta}\in[\tau-1,\tau]^{n\times m}} \frac{\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2}{2} \tag{S.1.1}$$

where

$$\max_{\boldsymbol{\Theta} \in [\tau-1,\tau]^{n\times m}} \|\boldsymbol{\Theta}\|_{\mathrm{F}}^2 = \max_{\boldsymbol{\Theta} \in [\tau-1,\tau]^{n\times m}} \sum_{i\le n, j\le m} \Theta_{ij}^2 \le (\tau \vee \{1-\tau\})^2 nm.$$

Recall that $\widehat{\boldsymbol{\Gamma}}$ is a minimizer of $L_\tau(\mathbf{S})$ defined in (2.1). Thus, for an arbitrary $\mathbf{S} \in \mathbb{R}^{p\times m}$,

$$\widetilde{L}_\tau(\widehat{\boldsymbol{\Gamma}}) \le L_\tau(\widehat{\boldsymbol{\Gamma}}) \le L_\tau(\mathbf{S}) \le \widetilde{L}_\tau(\mathbf{S}) + \kappa(\tau \vee \{1-\tau\})^2 \frac{nm}{2}, \tag{S.1.2}$$

where the first inequality is from the first inequality of (S.1.1), the second is the definition of the minimizer $\widehat{\boldsymbol{\Gamma}}$, and the third inequality is from the second inequality of (S.1.1). Recall that $\boldsymbol{\Gamma}_{\tau,\infty} = \lim_{t\to\infty} \boldsymbol{\Gamma}_{\tau,t}$ is a minimizer of $\widetilde{L}_\tau(\mathbf{S})$, then (S.1.2) gives

$$\widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,\infty}) \le \widetilde{L}_\tau(\widehat{\boldsymbol{\Gamma}}) \le \widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,\infty}) + \kappa(\tau \vee \{1-\tau\})^2 \frac{nm}{2}, \tag{S.1.3}$$

where the first is from the definition of $\boldsymbol{\Gamma}_{\tau,\infty}$ as a minimizer of $\widetilde{L}_\tau(\mathbf{S})$ and the second inequality is from (S.1.2), which holds for any arbitrary matrix $\mathbf{S} \in \mathbb{R}^{p\times m}$.

Now we focus on bounding

$$\begin{aligned} \left|L_\tau(\boldsymbol{\Gamma}_{\tau,t}) - L_\tau(\widehat{\boldsymbol{\Gamma}})\right| \le & \left|L_\tau(\boldsymbol{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,t})\right| + \left|\widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,\infty})\right| + \left|\widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,\infty}) - \widetilde{L}_\tau(\widehat{\boldsymbol{\Gamma}})\right| \\ & + \left|L_\tau(\widehat{\boldsymbol{\Gamma}}) - \widetilde{L}_\tau(\widehat{\boldsymbol{\Gamma}})\right|. \end{aligned} \tag{S.1.4}$$

The third term on the right-hand side of (S.1.4) is bounded by (S.1.3). For any matrix $\mathbf{S}$, by the choice of $\kappa = \epsilon/(2mn)$ in Algorithm 1, we have from (S.1.1) that

$$\left|L_\tau(\mathbf{S}) - \widetilde{L}_\tau(\mathbf{S})\right| \le \kappa \frac{nm(\tau \vee \{1-\tau\})^2}{2} \le \frac{\epsilon(\tau \vee \{1-\tau\})^2}{4}. \tag{S.1.5}$$

Hence, both $\left|L_\tau(\boldsymbol{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,t})\right|$ and $\left|L_\tau(\widehat{\boldsymbol{\Gamma}}) - \widetilde{L}_\tau(\widehat{\boldsymbol{\Gamma}})\right|$ satisfy (S.1.5).

Lemma S.1.3 implies that the gradient of $\widehat{Q}_{\tau,\kappa}(\boldsymbol{\Gamma})$ is Lipschitz continuous with Lipschitz constant M. By Theorem 4.1 of Ji and Ye (2009) or Theorem 4.4 of Beck and Teboulle (2009) (applied in general real Hilbert space, see their Remark 2.1), we have

$$\big|\widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\boldsymbol{\Gamma}_{\tau,\infty})\big| \leq \frac{2M\|\boldsymbol{\Gamma}_{\tau,0} - \boldsymbol{\Gamma}_{\tau,\infty}\|_{\mathrm{F}}^2}{(t+1)^2}, \tag{S.1.6}$$

where M is given in Lemma S.1.3. Since $\kappa = \epsilon/(2mn)$, $M = \frac{2}{mn\epsilon}\|\mathbf{X}\|^2$ by Lemma S.1.3.

Putting the bounds (S.1.3), (S.1.5) and (S.1.6) into (S.1.4), we have

$$\big|L_\tau(\boldsymbol{\Gamma}_{\tau,t}) - L_\tau(\widehat{\boldsymbol{\Gamma}})\big| \leq \frac{3\epsilon(\tau \vee \{1-\tau\})^2}{4} + \frac{4\|\boldsymbol{\Gamma}_{\tau,0} - \boldsymbol{\Gamma}_{\tau,\infty}\|_{\mathrm{F}}^2}{(t+1)^2}\frac{\|\mathbf{X}\|^2}{mn\epsilon}. \tag{S.1.7}$$

Hence, the proof of (2.11) is completed. Setting the right-hand side of (S.1.7) to be ϵ and solve it for T yields the bound (2.12). □

S.1.3. Technical Details for Theorem 2.3

Lemma S.1.2. *For any* $\mathbf{S}, \boldsymbol{\Theta} \in \mathbb{R}^{p\times m}$, $\widetilde{Q}_\tau(\mathbf{S}, \boldsymbol{\Theta})$ *can be expressed as* $\widetilde{Q}_\tau(\mathbf{S}, \boldsymbol{\Theta}) = \langle -\mathbf{X}\mathbf{S}, \boldsymbol{\Theta}\rangle + \langle \mathbf{Y}, \boldsymbol{\Theta}\rangle$.

Proof of Lemma S.1.2. One can show by elementary matrix algebra that

$$\begin{aligned}\widetilde{Q}_\tau(\mathbf{S}, \boldsymbol{\Theta}) &= \sum_{i=1}^n \sum_{j=1}^m \Theta_{ij}\left(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\right) = \sum_{i=1}^n \sum_{j=1}^m \Theta_{ij} Y_{ij} - \sum_{i=1}^n \sum_{j=1}^m \Theta_{ij}\boldsymbol{X}_i^\top \mathbf{S}_{*j} \\ &= \langle \mathbf{Y}, \boldsymbol{\Theta}\rangle + \langle -\mathbf{X}\mathbf{S}, \boldsymbol{\Theta}\rangle.\end{aligned}$$

The proof is therefore completed. □

Lemma S.1.3. *For any* $\kappa > 0$, $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ *is a well-defined, convex and continuously differentiable function in* $\mathbf{S}$ *with the gradient* $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{S}) = -(mn)^{-1}\mathbf{X}^\top\boldsymbol{\Theta}^*(\mathbf{S}) \in \mathbb{R}^{p\times m}$, *where* $\boldsymbol{\Theta}^*(\mathbf{S})$

is the optimal solution to (2.3), *namely*

$$\mathbf{\Theta}^*(\mathbf{S}) = [[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{S})]]_\tau. \tag{S.1.8}$$

The gradient $\nabla \widehat{Q}_{\tau,\kappa}(\mathbf{S})$ *is Lipschitz continuous with the Lipschitz constant* $M = (\kappa m^2 n^2)^{-1}\|\mathbf{X}\|^2$.

Proof of Lemma S.1.3. In view of Lemma S.1.2, we have from (2.3) that

$$\widehat{Q}_{\tau,\kappa}(\mathbf{S}) = \max_{\Theta_{ij}\in[\tau-1,\tau]} \left\{(mn)^{-1}\langle \mathbf{Y}, \mathbf{\Theta}\rangle + (mn)^{-1}\langle -\mathbf{X}\mathbf{S}, \mathbf{\Theta}\rangle - \frac{\kappa}{2}\|\mathbf{\Theta}\|_{\mathrm{F}}^2\right\}. \tag{S.1.9}$$

$\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ matches the form in (2.5) on page 131 of Nesterov (2005), with their $\widehat{\phi}(\mathbf{\Theta}) = (mn)^{-1}\langle \mathbf{Y}, \mathbf{\Theta}\rangle$ which is a continuous convex function, and their $A = -(mn)^{-1}\mathbf{X}$ which maps from the vector space $\mathbb{R}^{p\times m}$ to the space $\mathbb{R}^{n\times m}$ (the model setting described below (2.2) on page 129 of Nesterov (2005)), and their $d_2(\mathbf{\Theta}) = \frac{\kappa}{2}\|\mathbf{\Theta}\|_{\mathrm{F}}^2$. Therefore, applying Theorem 1 of Nesterov (2005), with $\sigma_2 = 1$, $d(\mathbf{\Theta}) = \|\mathbf{\Theta}\|_{\mathrm{F}}^2/2$, the gradient $\nabla \widehat{Q}_{\tau,\kappa}(\mathbf{S}) = -(mn)^{-1}\mathbf{X}^\top \mathbf{\Theta}^*(\mathbf{S}) \in \mathbb{R}^{p\times m}$, where $\mathbf{\Theta}^*(\mathbf{S})$ is the optimal solution to (2.3):

$$\mathbf{\Theta}^*(\mathbf{S}) = [[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{S})]]_\tau,$$

and the Lipschitz constant of $\nabla \widehat{Q}_{\tau,\kappa}(\mathbf{S})$ is $\|\mathbf{X}\|/(\kappa n^2 m^2)$, where $\|\mathbf{X}\|$ is the spectral norm of $\mathbf{X}$ (see line 8 on page 129 of Nesterov (2005)). Hence, the proof is completed. □

S.2: Proofs for Non-Asymptotic Bounds

S.2.1. Proof for Lemma 3.4

Applying the same $\mathcal{E}$-net argument on the unit Euclidean sphere $\mathcal{S}^{m-1} = \{\mathbf{u} \in \mathbb{R}^m : \|\mathbf{u}\|_2 = 1\}$ as in the first part of the proof of Lemma 3 in Negahban and Wainwright (2011) (page 6 to the beginning of page 7 in their supplemental materials), we obtain

$$\mathrm{P}\bigg(\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \geq 4s\bigg) = \mathrm{P}\bigg(\sup_{\substack{\mathbf{v}\in\mathcal{S}^{p-1}\\ \mathbf{u}\in\mathcal{S}^{m-1}}} \frac{1}{n}\big|\mathbf{v}^\top\mathbf{X}^\top\mathbf{W}\mathbf{u}\big| \geq 4s\bigg) \leq 8^{p+m} \sup_{\substack{\mathbf{v}\in\mathcal{S}^{p-1},\mathbf{u}\in\mathcal{S}^{m-1}\\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}} \mathrm{P}\bigg(\frac{|\langle\mathbf{X}\mathbf{v},\mathbf{W}\mathbf{u}\rangle|}{n} \geq s\bigg). \tag{S.2.1}$$

To bound $n^{-1}\langle\mathbf{X}\mathbf{v},\mathbf{W}\mathbf{u}\rangle = n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle$, first we show the sub-Gaussianity of $\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle$. Since $|W_{ij}| \leq \tau\vee(1-\tau)$. It follows by Lemma S.4.3 (Hoeffding's inequality) that

$$\mathrm{P}\big(\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle \geq s\big) \leq \exp\left(1 - \frac{C's^2}{\{\tau\vee(1-\tau)\}\|\mathbf{u}\|_2^2}\right) = \exp\left(1 - \frac{C's^2}{\tau\vee(1-\tau)}\right).$$

It can also be concluded that (see Definition 5.7 and discussion of Vershynin (2012a)) $\|\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle\|_{\psi_2} = \sqrt{\tau\vee(1-\tau)}$.

We apply Lemma S.4.3 again to bound $n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle$. Conditioning on $\boldsymbol{X}_i$, we have

$$\begin{aligned}\mathrm{P}\left(\left|n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle\langle\mathbf{u},\boldsymbol{W}_{\tau,i}\rangle\right| \geq s\right) &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau\vee(1-\tau)\}n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle^2}\right)\\ &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau\vee(1-\tau)\}c_2\|\boldsymbol{\Sigma}_X\|}\right).\end{aligned}$$

where the second inequality follows from the fact that $\|\mathbf{v}\|_2 = 1$ and $n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle^2 \leq \|\mathbf{X}^\top\mathbf{X}/n\| \leq c_2\|\boldsymbol{\Sigma}_X\|$ on the event that (A2) holds.

To summarize, on the event that (A2) holds,

$$\begin{aligned}\mathrm{P}\bigg(\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \geq 4s\bigg) &\leq 8^{p+m}\exp\left(1 - \frac{C'ns^2}{\{\tau\vee(1-\tau)\}c_2\|\boldsymbol{\Sigma}_X\|}\right)\\ &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau\vee(1-\tau)\}c_2\|\boldsymbol{\Sigma}_X\|} + (p+m)\log 8\right).\end{aligned}$$

Therefore,

$$\frac{1}{n}\|\mathbf{X}^\top \mathbf{W}\| \leq 4 \cdot \sqrt{2\log 8 \frac{\{\tau \vee (1-\tau)\}c_2\|\mathbf{\Sigma}_X\|}{C'}}\sqrt{\frac{p+m}{n}},$$

with probability greater than $1 - 3e^{-(p+m)\log 8} - \gamma_n$, as $e < 3$. $\square$

S.2.2. Proof for Lemma 3.5

We proceed as the proof for Lemma A.1. To simplify the notations in this proof, let $\widehat{\mathbf{\Delta}}_\infty = \mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}$, $\alpha_r \stackrel{\text{def}}{=} 4\sqrt{r/\sigma_{\min}(\mathbf{\Sigma}_X)}$, $\alpha_{r,m} \stackrel{\text{def}}{=} m^{1/2}\alpha_r$. Define

$$c_n \stackrel{\text{def}}{=} 16\sqrt{2}m^{-1/2}g_n\lambda^{-1}\sqrt{c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p}\sqrt{\log m + \log p},$$

where λ is chosen as in (3.7); recall from (S.3.1),

$$d_n = 8\sqrt{2}\alpha_r\sqrt{c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p}\sqrt{\log m + \log p}.$$

Let

Ω_1 : event that Assumption (A2) holds;

Ω_2 : event $\widetilde{\mathcal{A}}(u) \leq c_3(ud_n + c_n)$ for $c_3 > 1$;

Ω_3 : event $\dfrac{1}{n}\|\mathbf{X}^\top \mathbf{W}\| \leq C^*\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)\{\tau \vee (1-\tau)\}}\sqrt{\dfrac{p+m}{n}}$,

where $C^* = 4\sqrt{2\frac{c_2}{C'}\log 8}$,

$$\widetilde{\mathcal{A}}(u) \stackrel{\text{def}}{=} \sup_{\|\mathbf{\Delta}\|_{L_2(P_X)} \leq u, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, g_n)} \left| \mathbb{G}_n \left[m^{-1} \sum_{j=1}^m \left(\rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top(\mathbf{\Gamma}_{*j} + \mathbf{\Delta}_{*j})\} - \rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}\} \right) \right] \right|. \tag{S.2.2}$$

Note that the probability of event $\mathrm{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \geq 1 - \gamma_n - 16(pm)^{1-c_3^2} - 3e^{-(p+m)\log 8}$ from Assumption (A2), Lemma 3.4 and Lemma S.2.3. Set

$$u = \sqrt{n^{-1/2} c_3 c_n \frac{4}{\underline{f}} + \frac{4}{\underline{f}} g_n} + \frac{4}{\underline{f}}(n^{-1/2} c_3 d_n + \lambda \alpha_{r,m}). \tag{S.2.3}$$

It can be shown via the relation (S.2.6) and similar steps as in the proof for Theorem A.1 in Section S.3.1 (here, using Lemma S.2.3 and Lemma S.2.2 instead), that on $\Omega_1 \cap \Omega_2$ we have an expression similar to (S.3.5),

$$0 > \inf_{\|\boldsymbol{\Delta}\|_{L_2(P_X)} = u, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, g_n)} Q_\tau(\boldsymbol{\Gamma} + \boldsymbol{\Delta}) - Q_\tau(\boldsymbol{\Gamma}) - n^{-1/2} c_3 (d_n u + c_n) - \lambda(\alpha_{r,m} u + 2g_n/\lambda) - g_n,$$

Finally, since $\widetilde{\nu}_\tau(2g_n/\lambda) > u/4$ by (3.10), we obtain from Lemma S.2.2 (i) that

$$0 > \inf_{\|\boldsymbol{\Delta}\|_{L_2(P_X)} = u, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, g_n)} \frac{1}{4} \underline{f} u^2 - n^{-1/2} c_3 (d_n u + c_n) - \lambda(\alpha_{r,m} u + 2g_n/\lambda) - g_n. \tag{S.2.4}$$

With our choice of u in (S.2.3), the right-hand side of (S.2.4) is 0, and we get a contradiction. To complete the proof, by the choice for λ in (3.7), we can bound the expression in (S.2.3) by

$$\begin{aligned}
&n^{-1/2} c_n \\
&= \frac{16}{\sqrt{2}C^*} m^{-1/2} g_n(\kappa) m \sqrt{\frac{n}{p+m}} (\sigma_{\max}(\boldsymbol{\Sigma}_X)\{\tau \vee (1-\tau)\})^{-1/2} \sqrt{c_2 \sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p \frac{\log m + \log p}{n}} \\
&\leq C \frac{16}{\sqrt{2}C^*} g_n(\kappa) \sqrt{c_2 + \frac{B_p}{\sigma_{\max}(\boldsymbol{\Sigma}_X)} \frac{m(\log m + \log p)}{p+m}} \\
&\leq C_1 \sqrt{\frac{B_p}{\sigma_{\max}(\boldsymbol{\Sigma}_X)}} g_n(\kappa),
\end{aligned} \tag{S.2.5}$$

where C_1 is a constant depending on $\boldsymbol{X}$. Combining (S.2.5) with other terms in (S.2.3) we complete the proof of (3.11).

The bounds in Frobenius norm is from $\|\mathbf{\Delta}\|_{L_2(P_X)}^2 \geq (\sigma_{\min}(\mathbf{\Sigma}_X)/m)\|\mathbf{\Delta}\|_{\mathrm{F}}^2$ implied by (3.4) in Remark 3.3. Thus, the proof is completed. $\square$

S.2.3. Technical Details for Lemma 3.5

Lemma S.2.1. *Suppose $\lambda \geq 2\|\nabla \widehat{Q}(\mathbf{\Gamma})\|$ and $\mathbf{\Delta}_\infty = \mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}$. Then $\mathbf{\Delta}_\infty \in \mathcal{K}(\mathbf{\Gamma}, 2g_n/\lambda)$.*

Proof for Lemma S.2.1. We recall that $\mathbf{\Gamma}_{\tau,\infty}$ minimizes $\widetilde{L}_\tau(\mathbf{S})$, where $\widetilde{L}_\tau(\mathbf{S})$ is defined in (2.5). Also recall that $L_\tau(\mathbf{S})$ is defined in (2.1). For $g_n(\kappa)$ defined in (3.8), we have

$$L_\tau(\mathbf{\Gamma}_{\tau,\infty}) \leq \widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty}) + g_n \leq \widetilde{L}_\tau(\widehat{\mathbf{\Gamma}}) + g_n \leq L_\tau(\widehat{\mathbf{\Gamma}}) + g_n \leq L_\tau(\mathbf{\Gamma}) + g_n, \tag{S.2.6}$$

where the first inequality is by the second inequality in (S.1.1), the second follows by the definition of $\mathbf{\Gamma}_{\tau,\infty}$, the third inequality is from the first inequality in (S.1.1), and the last inequality is from the definition of $\widehat{\mathbf{\Gamma}}$.

Now, by exactly the same argument for obtaining (S.3.7), we have

$$(\lambda - \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Delta}_\infty)\|_* \leq (\lambda + \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta}_\infty)\|_* + g_n.$$

By $\lambda \geq 2\|\nabla \widehat{Q}(\mathbf{\Gamma})\|$, we get

$$\frac{1}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Delta}_\infty)\|_* \leq \frac{3}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta}_\infty)\|_* + g_n.$$

Hence, $\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Delta}_\infty)\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta}_\infty)\|_* + 2g_n/\lambda$. $\square$

Lemma S.2.2. *Under assumptions (A2) and (A3), we have*

(i) If $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 2g_n/\lambda)$, $\|\mathbf{\Delta}\|_{L_2(P_X)} \leq \widetilde{\nu}_\tau(2g_n/\lambda)$, then $Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) \geq \frac{1}{4}\underline{f}\|\mathbf{\Delta}\|_{L_2(P_X)}^2$, where $\widetilde{\nu}$ is defined in (3.9);

(ii) If $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 2g_n/\lambda)$, $\|\mathbf{\Delta}\|_* \leq 4\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\|\mathbf{\Delta}\|_{L_2(P_X)} + 2g_n/\lambda$, where $r = \mathrm{rank}(\mathbf{\Gamma})$.

Proof for Lemma S.2.2. The proof follows by similar argument for obtaining Lemma S.3.2 and is omitted for brevity. $\square$

Lemma S.2.3. *Under Assumptions (A1)-(A3),*

$$\mathrm{P}\Big\{\widetilde{\mathcal{A}}(u) \le 8\sqrt{2}c_3(\alpha_r u + 2m^{-1/2}g_n/\lambda)\sqrt{(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)}\sqrt{\log m + \log p}\Big\} \ge 1 - 16(pm)^{1-c_3^2} - \gamma_n,$$

where $c_3 > 1$, $\alpha_r = 4\sqrt{r/\sigma_{\min}(\boldsymbol{\Sigma}_X)}$ *and* $r = \mathrm{rank}(\boldsymbol{\Gamma})$.

Proof of Lemma S.2.3. Proceed analogously as the proof of Lemma S.3.3, we arrive with the same equation as (S.3.16):

$$\begin{aligned}
\sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le u \\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma},g_n)}} \left|\sum_{i=1}^n\sum_{j=1}^m \varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right| &\le \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le u \\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma},g_n)}} m^{1/2}\|\boldsymbol{\Delta}\|_* \max_{j\le m}\left\|\sum_{i=1}^n \varepsilon_{ij}\boldsymbol{X}_i\right\| \\
&\le m^{1/2}(\alpha_{m,r}\|\boldsymbol{\Delta}\|_{L_2(P_X)} + 2g_n/\lambda)\max_{j\le m}\left\|\sum_{i=1}^n \varepsilon_{ij}\boldsymbol{X}_i\right\|,
\end{aligned}$$

Continue as in the proof of Lemma S.3.3, we get an expression similar to (S.3.20),

$$\mathrm{P}\{\widetilde{\mathcal{A}}(u) > s|\Omega\} \le 8m(p+1)\exp\left(\frac{-\mu s}{4}\right)\exp\left\{2\mu^2(\alpha_r u + 2m^{-1/2}g_n/\lambda)^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)\right\}. \tag{S.2.7}$$

Minimize the expression (S.2.7) with respect to μ gives

$$\mathrm{P}\{\widetilde{\mathcal{A}}(u) > s|\Omega\} \le 8m(p+1)\exp\left\{-\frac{s^2}{128(\alpha_r u + 2m^{-1/2}g_n/\lambda)^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)}\right\}.$$

Take

$$s = 8\sqrt{2}c_3(\alpha_r u + 2m^{-1/2}g_n/\lambda)\sqrt{(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)}\sqrt{\log m + \log p}$$

to finish the proof. $\square$

S.2.4. Proof of Theorem 3.6

We proceed as the proof for Theorem A.2. To simplify the notations, let $\boldsymbol{\Delta}_{\tau,t} = \boldsymbol{\Gamma}_{\tau,t} - \boldsymbol{\Gamma}$, $\alpha_r \overset{\text{def}}{=} 4\sqrt{r/\sigma_{\min}(\boldsymbol{\Sigma}_X)}$, $\alpha_{r,m} \overset{\text{def}}{=} m^{1/2}\alpha_r$. Define

$$\widetilde{c}_n \overset{\text{def}}{=} 16\sqrt{2}m^{-1/2}a_{n,t}(\kappa,\epsilon)\lambda^{-1}\sqrt{c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p}\sqrt{\log m + \log p},$$

where λ is chosen as in (3.7); recall from (S.3.1),

$$d_n = 8\sqrt{2}\alpha_r\sqrt{c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p}\sqrt{\log m + \log p}.$$

Let

Ω_1 : event that Assumption (A2) holds;

Ω_2 : event $\boldsymbol{\Delta}_{\tau,t} \in \mathcal{K}(\boldsymbol{\Gamma}, a_{n,t}(\kappa,\epsilon))$;

Ω_3 : event $\dfrac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq C^*\sqrt{\sigma_{\max}(\boldsymbol{\Sigma}_X)\{\tau \vee (1-\tau)\}}\sqrt{\dfrac{p+m}{n}}$;

Ω_4 : event $\widetilde{\mathcal{B}}(u) \leq c_3(ud_n + \widetilde{c}_n)$ for $c_3 > 1$,

where

$$\widetilde{\mathcal{B}}(u) \overset{\text{def}}{=} \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\leq u,\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma},a_{n,t}(\kappa,\epsilon))}} \left|\mathbb{G}_n\left[m^{-1}\sum_{j=1}^m \big(\rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top(\boldsymbol{\Gamma}_{*j} + \boldsymbol{\Delta}_{*j})\} - \rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top\boldsymbol{\Gamma}_{*j}\}\big)\right]\right|. \tag{S.2.8}$$

Note that the probability of event $\mathrm{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4) \geq 1 - 2\gamma_n - 32(pm)^{1-c_3^2} - 6\exp\{-(p+m)\log 8\}$ from Assumption (A2), Lemma 3.4, Lemma S.2.4 and Lemma S.2.6.

Set

$$u = \sqrt{n^{-1/2} c_3 \widetilde{c}_n \frac{4}{\underline{f}} + \frac{4}{\underline{f}} a_{n,t}(\kappa, \epsilon)} + \frac{4}{\underline{f}} (n^{-1/2} c_3 d_n + \lambda \alpha_{r,m}). \tag{S.2.9}$$

It can be shown via the relation (S.2.11) and similar steps as in the proof for Lemma A.1 in Section S.3.1 (here, using Lemma S.2.6 and Lemma S.2.5 instead), that on $\Omega_1 \cap \Omega_2 \cap \Omega_3$ we have an expression similar to (S.3.5),

$$0 > \inf_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)} = u, \\ \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, a_{n,t}(\kappa,\epsilon))}} Q_\tau(\boldsymbol{\Gamma} + \boldsymbol{\Delta}) - Q_\tau(\boldsymbol{\Gamma}) - n^{-1/2} c_3 (d_n u + \widetilde{c}_n) - \lambda(\alpha_{r,m} u + 2a_{n,t}(\kappa, \epsilon)/\lambda) - a_{n,t}(\kappa, \epsilon),$$

Finally, since $\widetilde{\nu}_\tau(2a_{n,t}(\kappa, \epsilon)/\lambda) > u/4$ by (3.14), we obtain from Lemma S.2.5 (i) that

$$0 > \inf_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)} = u, \\ \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, a_{n,t}(\kappa,\epsilon))}} \frac{1}{4} \underline{f} u^2 - n^{-1/2} c_3 (d_n u + \widetilde{c}_n) - \lambda(\alpha_{r,m} u + 2a_{n,t}(\kappa, \epsilon)/\lambda) - a_{n,t}(\kappa, \epsilon). \tag{S.2.10}$$

With our choice of u in (S.2.9), the right-hand side of (S.2.10) is 0, and we get a contradiction.

The bounds in Frobenius norm is from $\|\boldsymbol{\Delta}\|^2_{L_2(P_X)} \geq (\sigma_{\min}(\boldsymbol{\Sigma}_X)/m) \|\boldsymbol{\Delta}\|^2_{\mathrm{F}}$ implied by (3.4) in Remark 3.3. Thus, the proof is completed. □

S.2.5. Technical Details for the Proof of Theorem 3.6

Lemma S.2.4. *Let* $\boldsymbol{\Delta}_{\tau,t} = \boldsymbol{\Gamma}_{\tau,t} - \boldsymbol{\Gamma}$ *and* $\lambda \geq 2\|\nabla \widehat{Q}_\tau(\boldsymbol{\Gamma})\|$. *Suppose (A1)-(A3) hold. Then* $\boldsymbol{\Delta}_{\tau,t} \in \mathcal{K}(\boldsymbol{\Gamma}; a_{n,t}(\kappa, \epsilon))$ *with probability* $1 - \gamma_n - 16(pm)^{1-c_3^2} - 3\exp\{-(p+m)\log 8\}$, *where* $\mathcal{K}(\boldsymbol{\Gamma}; a_{n,t}(\kappa, \epsilon))$, $a_{n,t}(\kappa, \epsilon)$ *are defined in* (3.3) *and* (3.13).

Proof of Lemma S.2.4. Recall the function $\widehat{Q}_{\kappa,\tau}(\cdot)$ defined in (2.3). $\boldsymbol{\Gamma}_{\tau,\infty}$ is the minimizer of the loss function $\widehat{Q}_{\kappa,\tau}(\mathbf{S}) + \lambda\|\mathbf{S}\|_*$. Therefore,

$$\begin{aligned}
0 &\le \widehat{Q}_{\kappa,\tau}(\mathbf{\Gamma}) - \widehat{Q}_{\kappa,\tau}(\mathbf{\Gamma}_{\tau,\infty}) + \lambda\|\mathbf{\Gamma}\|_* - \lambda\|\mathbf{\Gamma}_{\tau,\infty}\|_* \\
&\le \widehat{Q}_{\kappa,\tau}(\mathbf{\Gamma}) - \widehat{Q}_{\kappa,\tau}(\mathbf{\Gamma}_{\tau,t}) + \lambda\|\mathbf{\Gamma}\|_* - \lambda\|\mathbf{\Gamma}_{\tau,t}\|_* + \left|\widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty})\right| \\
&\le \widehat{Q}_{\tau}(\mathbf{\Gamma}) - \widehat{Q}_{\tau}(\mathbf{\Gamma}_{\tau,t}) + \lambda\|\mathbf{\Gamma}\|_* - \lambda\|\mathbf{\Gamma}_{\tau,t}\|_* + R_{n,t}(\kappa,\epsilon) \\
&\le \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|\big(\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\tau,t})\|_* + \|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Delta}_{\tau,t})\|_*\big) + \lambda(\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\tau,t})\|_* - \|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Delta}_{\tau,t})\|_*) + R_{n,t}(\kappa,\epsilon),
\end{aligned} \tag{S.2.11}$$

where the first inequality is from the definition of $\mathbf{\Gamma}^\infty$, the second inequality is by the definition of $\widetilde{L}$ in (2.5), and $R_{n,t}(\kappa,\epsilon)$ in the third inequality is defined by

$$R_{n,t}(\kappa,\epsilon) \stackrel{\text{def}}{=} 2\sup_{\mathbf{S}\in\mathbb{R}^{p\times m}} \left|\widehat{Q}_\tau(\mathbf{S}) - \widehat{Q}_{\kappa,\tau}(\mathbf{S})\right| + \left|\widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,t}) - \widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty})\right|; \tag{S.2.12}$$

the last inequality follows by exactly the same argument for obtaining S.3.7 in Lemma S.3.1.

We note that with probability $1 - \gamma_n - 16(pm)^{1-c_3^2} - 3\exp\{-(p+m)\log 8\}$,

$$\begin{aligned}
R_{n,t}(\kappa,\epsilon) &\le \kappa(\tau \vee \{1-\tau\})^2 nm + \frac{4\|\mathbf{\Gamma}_{\tau,0} - \mathbf{\Gamma}_{\tau,\infty}\|_{\mathrm{F}}^2}{(t+1)^2}\frac{\|\mathbf{X}\|^2}{mn\epsilon} \\
&\le \kappa(\tau \vee \{1-\tau\})^2 nm + \frac{8c_2^2(\|\mathbf{\Gamma}\|_{\mathrm{F}}^2 + h_n^2)\sigma_{\max}^2(\mathbf{\Sigma}_X)}{(t+1)^2\epsilon m} \\
&= a_{n,t}(\kappa,\epsilon),
\end{aligned}$$

where the first inequality is from (S.1.1) and (S.1.6), and the second follows by Lemma 3.5, and $\|\mathbf{X}\|^2/n = \sigma_{\max}^2(\widehat{\mathbf{\Sigma}}_X) \le c_2\sigma_{\max}(\mathbf{\Sigma}_X)$ with probability greater than $1-\gamma_n$ from Assumption (A2). The last equality is the definition of $a_{n,t}(\kappa,\epsilon)$ in (3.13).

Rearrange expression (S.2.11) to get,

$$(\lambda - \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \le (\lambda + \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* + a_{n,t}(\kappa,\epsilon).$$

By $\lambda \geq 2\|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|$,

$$\frac{1}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \leq \frac{3}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* + a_{n,t}(\kappa,\epsilon).$$

Hence, $\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* + 2a_{n,t}(\kappa,\epsilon)/\lambda$.

□

Lemma S.2.5. *Under assumptions (A2) and (A3), we have*

(i) If $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 2a_{n,t}(\kappa,\epsilon)/\lambda)$, $\|\mathbf{\Delta}\|_{L_2(P_X)} \leq \widetilde{\nu}_\tau(2a_{n,t}(\kappa,\epsilon)/\lambda)$, *where* $\widetilde{\nu}$ *is defined in* (3.9), *then* $Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) \geq \frac{1}{4}\underline{f}\|\mathbf{\Delta}\|_{L_2(P_X)}^2$;

(ii) If $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 2g_n/\lambda)$, $\|\mathbf{\Delta}\|_* \leq 4\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\|\mathbf{\Delta}\|_{L_2(P_X)} + 2g_n/\lambda$, *where* $r = \operatorname{rank}(\mathbf{\Gamma})$.

Proof for Lemma S.2.2. The proof follows by similar argument for obtaining Lemma S.3.2 and is omitted for brevity. □

Lemma S.2.6. *Under assumptions (A1)-(A3),*

$$\mathrm{P}\Big\{\mathcal{B}(t) \leq 8\sqrt{2}c_3(\alpha_r u + 2m^{-1/2}a_{n,t}(\kappa,\epsilon)/\lambda)\sqrt{(c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p)}\sqrt{\log m + \log p}\Big\} \geq 1 - 16(pm)^{1-c_3^2} - \gamma_n,$$

where $c_3 > 1$, $\alpha_r = 4\sqrt{r/\sigma_{\min}(\mathbf{\Sigma}_X)}$ *and* $r = \operatorname{rank}(\mathbf{\Gamma})$.

Proof for Lemma S.2.6. The proof follows by similar arguments in the proof of Lemma S.2.3, and replace g_n by $a_{n,t}(\kappa,\epsilon)$ there. We omit the details for brevity. □

S.2.6. Proof of Theorem 3.10

In this proof, we abbreviate $\sigma_k^2(\mathbf{\Gamma})$, $\sigma_k^2(\mathbf{\Gamma}_{\tau,t})$, $(\widehat{\mathbf{V}}_\tau)_{*k}$ and $(\mathbf{V}_\tau)_{*k}$, $(\widehat{\mathbf{U}}_\tau)_{*k}$ and $(\mathbf{U}_\tau)_{*k}$ by σ_k, $\widehat{\sigma}_k$, $\widehat{\mathbf{V}}_{*k}$ and $\mathbf{V}_{*k}$, $\widehat{\mathbf{U}}_{*k}$ and $\mathbf{U}_{*k}$.

To prove (3.18), since $\boldsymbol{\Psi}_\tau = \mathbf{V}_\tau$ and $\widehat{\boldsymbol{\Psi}}_\tau = \mathbf{V}_{\tau,t}$, by Theorem 3 of Yu et al. (2015),

$$\sin\cos^{-1}(|\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}|) \le \frac{2(2\|\boldsymbol{\Gamma}\| + \|\boldsymbol{\Gamma}_{\tau,t} - \boldsymbol{\Gamma}\|_{\mathrm{F}})\|\boldsymbol{\Gamma}_{\tau,t} - \boldsymbol{\Gamma}\|_{\mathrm{F}}}{\min\{\sigma_{j-1}^2(\boldsymbol{\Gamma}) - \sigma_j^2(\boldsymbol{\Gamma}), \sigma_j^2(\boldsymbol{\Gamma}) - \sigma_{j+1}^2(\boldsymbol{\Gamma})\}} \tag{S.2.13}$$

where by the fact that $|\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}| \le 1$,

$$\begin{aligned}\sin\cos^{-1}(|\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}|) &= \sqrt{1 - (\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j})^2} = \sqrt{(1 - \widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j})(1 + \widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j})} \\ &\ge \sqrt{(1 - |\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}|)^2} = 1 - \left|\widehat{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}\right|.\end{aligned}$$

Similar bound like (3.18) also holds for $\widehat{\mathbf{U}}_{*j}$, by the discussion below Theorem 3 of Yu et al. (2015).

For a proof for inequality (3.19), by direct calculation,

$$\begin{aligned}\left|\widehat{f}_k^\tau(\boldsymbol{X}_i) - f_k^\tau(\boldsymbol{X}_i)\right| &= \left|\widehat{\sigma}_k \widehat{\mathbf{U}}_{*k}^\top \boldsymbol{X}_i - \sigma_k \mathbf{U}_{*k}^\top \boldsymbol{X}_i\right| \\ &\le \left\|\widehat{\sigma}_k \widehat{\mathbf{U}}_{*k}^\top - \sigma_k \mathbf{U}_{*k}^\top\right\| \|\boldsymbol{X}_i\| \\ &\le \left(\left|\widehat{\sigma}_k - \sigma_k\right| \left\|\widehat{\mathbf{U}}_{*k}\right\| + \sigma_k \left\|\widehat{\mathbf{U}}_{*k} - \mathbf{U}_{*k}\right\|\right) \|\boldsymbol{X}_i\| \\ &\le \left(\left|\widehat{\sigma}_k - \sigma_k\right| + \sigma_k \sqrt{(\widehat{\mathbf{U}}_{*k} - \mathbf{U}_{*k})^\top (\widehat{\mathbf{U}}_{*k} - \mathbf{U}_{*k})}\right) \|\boldsymbol{X}_i\| \\ &\le \left(\left|\widehat{\sigma}_k - \sigma_k\right| + \sigma_k \sqrt{2(1 - \widehat{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k})}\right) \|\boldsymbol{X}_i\| \end{aligned} \tag{S.2.14}$$

where we apply the fact that $\|\widehat{\mathbf{U}}_{*k}\| = 1$. By assumption $\widehat{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k} \ge 0$, $\widehat{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k} = |\widehat{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k}|$. Apply Lemma 3.9 and the bound (S.2.13) with $\mathbf{V}$ being replaced by $\mathbf{U}$ to (S.2.14), then (3.19) is proved. Thus, the proof for this theorem is completed. □

S.3: Proof for Oracle Properties for Exact Optimizer $\widehat{\boldsymbol{\Gamma}}$

S.3.1. Proof for Lemma A.1

To simplify the notations in this proof, let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}$, $\alpha_{r,m} \stackrel{\text{def}}{=} 4\sqrt{2m/\sigma_{\min}(\mathbf{\Sigma}_X)}$ and

$$d_n \stackrel{\text{def}}{=} 32\sqrt{2}\sqrt{r}\sqrt{\frac{c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\sqrt{\log m + \log p}. \tag{S.3.1}$$

Ω_1 : event that Assumption (A2) holds;

Ω_2 : event $\mathcal{A}(t) \leq tc_3 d_n$ for $c_3 > 1$,

where

$$\mathcal{A}(t) \stackrel{\text{def}}{=} \sup_{\|\mathbf{\Delta}\|_{L_2(P_X)} \leq t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})} \left| \mathbb{G}_n \left[m^{-1} \sum_{j=1}^{m} \left(\rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top (\mathbf{\Gamma}_{*j} + \mathbf{\Delta}_{*j})\} - \rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}\} \right) \right] \right|. \tag{S.3.2}$$

Note that the probability of event $\mathrm{P}(\Omega_1 \cap \Omega_2) \geq 1 - \gamma_n - 16(pm)^{1-c_3^2}$ from Assumption (A2) and Lemma S.3.3. Set

$$t = 4\underline{f}^{-1} c_3 n^{-1/2} d_n + 4\lambda \frac{\alpha_{r,m}}{\underline{f}} > 0. \tag{S.3.3}$$

Suppose to the contrary that $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} > t$ is true, together with $\widehat{\mathbf{\Delta}} \in \mathcal{K}(\mathbf{\Gamma})$ from Lemma S.3.1, so from the fact that $\widehat{\mathbf{\Gamma}}$ minimizes $\widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*$,

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)} \geq t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})} \widehat{Q}_\tau(\mathbf{\Gamma} + \mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}) + \lambda(\|\mathbf{\Gamma} + \mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*), \tag{S.3.4}$$

where the strict negativity is from the uniqueness of minimizer $\widehat{\mathbf{\Gamma}}$ as argued in Remark 2.1 in Koenker (2005). As argued in the proof of Theorem 2 of Belloni and Chernozhukov (2011), from the facts that

1. $\widehat{Q}_\tau(\cdot) + \lambda\|\cdot\|_*$ is convex;

2. $\mathcal{K}(\mathbf{\Gamma})$ is a cone,

(S.3.4) forces the value of $\widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) + \lambda\|\mathbf{\Gamma}+\mathbf{\Delta}\|_*$ on $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{L_2(P_X)} \geq t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})\}$ to be less than that evaluated at $\mathbf{\Delta} = 0$. Convexity implies that $\widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) + \lambda\|\mathbf{\Gamma}+\mathbf{\Delta}\|_*$ evaluated at $\{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{L_2(P_X)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})\}$ must be smaller than that evaluated at $\mathbf{\Delta} = 0$. Thus, we have the inequality

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})} \widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}) + \lambda(\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*).$$

With regard of the definition $\mathcal{A}(t)$ in (S.3.2), it can be further deducted that

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})} Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) - n^{-1/2}\mathcal{A}(t) + \lambda(\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*),$$

By triangle inequality, $\big|\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*\big| \leq \|\mathbf{\Delta}\|_* \leq \alpha_{r,m}\|\mathbf{\Delta}\|_{L_2(P_X)} = \alpha_{r,m}t$ on the set $\{\|\mathbf{\Delta}\|_{L_2(P_X)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})\}$. Furthermore, on event $\Omega_1 \cap \Omega_2$, it holds from Lemma S.3.2 (ii) that

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})} Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) - n^{-1/2}c_3 d_n t - \lambda\alpha_{r,m}t, \tag{S.3.5}$$

Finally, since $\nu_\tau > t/4$ by (A.1) and $t = \|\mathbf{\Delta}\|_{L_2(P_X)}$, we obtain from Lemma S.3.2 (i) that

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})} \frac{1}{4}\underline{f}t^2 - n^{-1/2}c_3 d_n t - \lambda\alpha_{r,m}t. \tag{S.3.6}$$

With our choice of t in (S.3.3), the right-hand side of (S.3.6) is 0, and we get a contradiction. Thus, we established the inequality (A.2).

The bounds in Frobenius and nuclear norm are from $\|\mathbf{\Delta}\|_{L_2(P_X)}^2 \geq (\sigma_{\min}(\mathbf{\Sigma}_X)/m)\|\mathbf{\Delta}\|_{\mathrm{F}}^2$

implied by (3.4) in Remark 3.3 and $\|\widehat{\boldsymbol{\Delta}}\|_* \le \alpha_{r,m}\|\widehat{\boldsymbol{\Delta}}\|_{L_2(P_X)}$ from the fact that $\widehat{\boldsymbol{\Delta}} \in \mathcal{K}(\boldsymbol{\Gamma})$ (Lemma S.3.1) and Lemma S.3.2 (ii). Thus, the proof is completed. □

S.3.2. Proof for Theorem A.2

Let events Ω_1 and Ω_2 be defined as in the proof of Theorem A.2, and

$$\Omega_3 = \text{ the event that } \frac{1}{n}\|\mathbf{X}^\top \mathbf{W}_\tau\| \le C^* \sqrt{\|\boldsymbol{\Sigma}_X\|\{\tau \vee (1-\tau)\}}\sqrt{\frac{p+m}{n}}.$$

Note that the probability $\mathrm{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \ge 1-\gamma_n - 16(pm)^{1-c_3^2} - 3e^{-(p+m)\log 8}$. On $\Omega_1 \cap \Omega_2 \cap \Omega_3$, the bounds (A.2) and (3.6) hold. Substituting λ with (3.7) in (A.2) yields bounds (A.6).

The bounds in Frobenius and nuclear norm can be deducted by the same argument as in the proof of Theorem A.2. Hence, the proof is completed. □

S.3.3. Technical Details for Theorem A.2

The following lemma asserts that the empirical error $\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$ lies in the cone $\mathcal{K}(\boldsymbol{\Gamma})$.

Lemma S.3.1. *Suppose $\lambda \ge 2\|\nabla \widehat{Q}(\boldsymbol{\Gamma})\|$ and $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$. Then $\|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Delta}})\|_* \le 3\|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}})\|_*$. That is, $\widehat{\boldsymbol{\Delta}} \in \mathcal{K}(\boldsymbol{\Gamma})$.*

Proof for Lemma S.3.1. Recall that $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$,

$$\begin{aligned}
0 &\le \widehat{Q}_\tau(\boldsymbol{\Gamma}) - \widehat{Q}_\tau(\widehat{\boldsymbol{\Gamma}}) + \lambda(\|\boldsymbol{\Gamma}\|_* - \|\widehat{\boldsymbol{\Gamma}}\|_*) \quad (\widehat{\boldsymbol{\Gamma}} \text{ is the minimizer of } \widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*) \\
&\le \|\nabla \widehat{Q}_\tau(\boldsymbol{\Gamma})\|\|\widehat{\boldsymbol{\Delta}}\|_* + \lambda(\|\boldsymbol{\Gamma}\|_* - \|\widehat{\boldsymbol{\Gamma}}\|_*) \\
&\le \|\nabla \widehat{Q}_\tau(\boldsymbol{\Gamma})\|\big(\|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}})\|_* + \|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma})\|_* - \|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Gamma}})\|_* - \|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Gamma}})\|_*) \\
&\le \|\nabla \widehat{Q}_\tau(\boldsymbol{\Gamma})\|\big(\|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}})\|_* + \|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}})\|_* - \|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Delta}})\|_*), \qquad \text{(S.3.7)}
\end{aligned}$$

where the second inequality follows from the definition of subgradient:

$$\widehat{Q}_\tau(\widehat{\mathbf{\Gamma}}) - \widehat{Q}_\tau(\mathbf{\Gamma}) \geq \langle \nabla \widehat{Q}_\tau(\mathbf{\Gamma}), \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma} \rangle,$$

and Hölder's inequality; the third inequality is from the fact that $\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{\Gamma}) = 0$ and for any $\mathbf{S}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{S})\|_* + \|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\mathbf{S})\|_*$ (the discussion after Definition 3.1) ; the fourth inequality is from the triangle inequality.

Rearrange expression (S.3.7) to get,

$$(\lambda - \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \leq (\lambda + \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*.$$

Choose $\lambda \geq 2\|\nabla \widehat{Q}_\tau(\mathbf{\Gamma})\|$,

$$\frac{1}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \leq \frac{3}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*.$$

Hence, $\|\mathcal{P}_{\mathbf{\Gamma}}^{\perp}(\widehat{\mathbf{\Delta}})\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*$. □

Lemma S.3.2. *Under assumptions (A2), (A3), we have*

(i) If $\|\mathbf{\Delta}\|_{L_2(P_X)} \leq 4\nu_\tau$, where $\nu_\tau = \widetilde{\nu}_\tau(0)$, and $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})$, then $Q_\tau(\mathbf{\Gamma} + \mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) \geq \frac{1}{4}\underline{f}\|\mathbf{\Delta}\|^2_{L_2(P_X)}$;

(ii) If $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma})$, $\|\mathbf{\Delta}\|_* \leq 4\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\|\mathbf{\Delta}\|_{L_2(P_X)}$, where $r = \mathrm{rank}(\mathbf{\Gamma})$.

Proof for Lemma S.3.2.

1. Let $Q_{\tau,j}(\mathbf{\Gamma}_{*j}) = \mathsf{E}[\rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j})]$. From Knight's identity (Knight; 1998), for any $v, u \in \mathbb{R}$,

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v \big(\mathbf{1}\{u \leq z\} - \mathbf{1}\{u \leq 0\}\big)dz. \tag{S.3.8}$$

Putting $u = Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}$ in (S.3.8), and $v = \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}$, $\mathsf{E}[-v\psi_\tau(u)] = 0$ for all j and i, by the definition of $\mathbf{\Gamma} = \arg\min_{\mathbf{S}} \mathsf{E}[\widehat{Q}_\tau(\mathbf{S})]$. Therefore, using law of iterative expectation and mean value theorem, we have by (A3) that

$$
\begin{aligned}
&Q_{\tau,j}(\mathbf{\Gamma}_{*j}+\mathbf{\Delta}_{*j})-Q_{\tau,j}(\mathbf{\Gamma}_{*j})\\
&=\mathsf{E}\Big[\int_0^{\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}} F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{*j}+z|\boldsymbol{X}_i)-F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{*j}|\boldsymbol{X}_i)dz\Big]\\
&=\mathsf{E}\Big[\int_0^{\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}} zf_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{*j}|\boldsymbol{X}_i)+\frac{z^2}{2}f'_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{*j}+z^\dagger|\boldsymbol{X}_i)dz\Big]\\
&\geq \underline{f}\frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2\big]}{4}+\underline{f}\frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2\big]}{4}-\frac{1}{6}\bar{f}'\mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^3]
\end{aligned}
\tag{S.3.9}
$$

for $z^\dagger\in[0,z]$. Now, for $\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})$, the condition

$$
\|\mathbf{\Delta}\|_{L_2(P_X)}\leq 4\nu_\tau=\frac{3}{2}\frac{\underline{f}}{\bar{f}'}\inf_{\substack{\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})\\ \mathbf{\Delta}\neq 0}}\frac{\big(\sum_{j=1}^m\mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^2]\big)^{3/2}}{\sum_{j=1}^m\mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^3]}
$$

implies

$$
\underline{f}m^{-1}\sum_{j=1}^m\frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2\big]}{4}\geq\frac{1}{6}\bar{f}'m^{-1}\sum_{j=1}^m\mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^3]
$$

Therefore,

$$
Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta})-Q_\tau(\mathbf{\Gamma})\geq\underline{f}m^{-1}\sum_{j=1}^m\frac{\mathsf{E}(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2}{4}=\frac{1}{4}\underline{f}\|\mathbf{\Delta}\|_{L_2(P_X)}^2.
$$

2. By the decomposability of nuclear norm, $\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma})$ and (3.5) in Remark 3.3, we can estimate

$$
\begin{aligned}
\|\mathbf{\Delta}\|_*&=\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_*+\|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\mathbf{\Delta})\|_*\leq 4\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_*\leq 4\sqrt{r}\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_{\mathrm{F}}\\
&\leq 4\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\|\mathbf{\Delta}\|_{L_2(P_X)}.
\end{aligned}
$$

□

Lemma S.3.3. *Under Assumptions (A1)-(A3),*

$$\mathrm{P}\left\{\mathcal{A}(t) \le t32\sqrt{2}c_3\sqrt{r}\sqrt{\frac{c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p}{\sigma_{\min}(\boldsymbol{\Sigma}_X)}}\sqrt{\log m+\log p}\right\} \ge 1-16(pm)^{1-c_3^2}-\gamma_n,$$

where $c_3>1$ *and* $r=\operatorname{rank}(\boldsymbol{\Gamma})$.

Proof for Lemma S.3.3. To simplify notations, let $\alpha_r \overset{\text{def}}{=} 4\sqrt{r/\sigma_{\min}(\boldsymbol{\Sigma}_X)}$. Let $\{\varepsilon_{ij}\}_{i\le n, j\le m}$ be independent Rademacher random variables independent from Y_{ij} and $\boldsymbol{X}_i$ for all i,j. Denote P_ε and E_ε as the conditional probability and the conditional expectation with respect to $\{\varepsilon_{ij}\}_{i\le n, j\le m}$, given Y_{ij} and $\boldsymbol{X}_i$. Denote

$$\chi_{ij}^{\tau}(\cdot) \overset{\text{def}}{=} \rho_\tau\{Y_{ij}-\boldsymbol{X}_i^\top\boldsymbol{\Gamma}_{*j}-\cdot\}-\rho_\tau\{Y_{ij}-\boldsymbol{X}_i^\top\boldsymbol{\Gamma}_{*j}\}. \tag{S.3.10}$$

$\chi_{ij}^{\tau}(\cdot)$ is a contraction in the sense that $\chi_{ij}^{\tau}(0)=0$, and for all $a,b\in\mathbb{R}$,

$$\left|\chi_{ij}^{\tau}(a)-\chi_{ij}^{\tau}(b)\right| \le |a-b|. \quad \forall i=1,...,n,\ j=1,...,m. \tag{S.3.11}$$

First, we note that for any $\boldsymbol{\Delta}$ satisfying $\boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma})$ and $\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le t$,

$$\begin{aligned}
&\operatorname{Var}\left(\mathbb{G}_n\left(m^{-1}\sum_{j=1}^m \chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right)\right)\\
&=\operatorname{Var}\left(m^{-1}\sum_{j=1}^m \chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right) \le m^{-1}\sum_{j=1}^m \mathsf{E}\big[(\chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}))^2\big]\\
&\le m^{-1}\sum_{j=1}^m \mathsf{E}\big[(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})^2\big] \le t^2,
\end{aligned} \tag{S.3.12}$$

where the first equality and the second inequality follows from elementary computations and i.i.d. assumption (A1), the third inequality is a result of (S.3.11), and the last inequality applies (3.4) in Remark 3.3.

To apply Lemma 2.3.7 of van der Vaart and Wellner (1996), we observe from Chebyshev's inequality that for any $s > 0$,

$$\inf_{\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma})} \mathrm{P}\left(\left| \mathbb{G}_n \left(m^{-1} \sum_{j=1}^{m} \chi_{ij}^{\tau}(\boldsymbol{X}_i^{\top} \boldsymbol{\Delta}_{*j}) \right) \right| < \frac{s}{2} \right)$$
$$= 1 - \sup_{\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma})} \mathrm{P}\left(\left| \mathbb{G}_n \left(m^{-1} \sum_{j=1}^{m} \chi_{ij}^{\tau}(\boldsymbol{X}_i^{\top} \boldsymbol{\Delta}_{*j}) \right) \right| \ge \frac{s}{2} \right) \ge 1 - 4\frac{t^2}{s^2}.$$

Taking $s \ge \sqrt{8}t$, we have

$$\frac{1}{2} \le \inf_{\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma})} \mathrm{P}\left(\left| \mathbb{G}_n \left(m^{-1} \sum_{j=1}^{m} \chi_{ij}^{\tau}(\boldsymbol{X}_i^{\top} \boldsymbol{\Delta}_{*j}) \right) \right| < \frac{s}{2} \right).$$

Thus, applying Lemma 2.3.7 of van der Vaart and Wellner (1996), we have

$$\mathrm{P}\{\mathcal{A}(t) > s\} \le 4\mathrm{P}\left(\sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t \\ \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma})}} \left| n^{-1/2} m^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \varepsilon_{ij} \chi_{ij}^{\tau}(\boldsymbol{X}_i^{\top} \boldsymbol{\Delta}_{*j}) \right| > \frac{s}{4} \right). \tag{S.3.13}$$

Now we restrict the $\mathcal{A}(t)$ on the event Ω on which (3.2) in (A2) holds, with $\mathrm{P}(\Omega) \ge 1 - \gamma_n$. Applying Markov's inequality, for an arbitrary constant $\mu > 0$, the right-hand side of (S.3.13) can be bounded by

$$\mathrm{P}\{\mathcal{A}(t) > s | \Omega\}$$
$$\le 4 \exp\left(\frac{-\mu s}{4} \right) \mathsf{E}\left[\mathsf{E}_{\varepsilon}\left[\exp\left\{ \mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t \\ \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma})}} \left| n^{-1/2} m^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \varepsilon_{ij} \chi_{ij}^{\tau}(\boldsymbol{X}_i^{\top} \boldsymbol{\Delta}_{*j}) \right| \right\} \right] \middle| \Omega \right]. \tag{S.3.14}$$

Now recall (S.3.11), the comparison theorem for Rademacher processes (Lemma 4.12 in

Ledoux and Talagrand (1991)) implies the right-hand side of (S.3.14) is bounded by

$$
\begin{aligned}
&\mathrm{P}\{\mathcal{A}(t) > s|\Omega\} \\
&\leq 4\exp\left(\frac{-\mu s}{4}\right)\mathsf{E}\left[\mathsf{E}_\varepsilon\left[\exp\left\{2\mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\leq t \\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma})}}\left|n^{-1/2}m^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right|\right\}\right]\middle|\Omega\right]. \qquad \text{(S.3.15)}
\end{aligned}
$$

To obtain a bound for the right-hand side of (S.3.15), we note that

$$
\begin{aligned}
\left|\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right| &= \left|\mathrm{tr}\Big(\Big[\sum_{i=1}^{n}\varepsilon_{i1}\boldsymbol{X}_i \;\; \sum_{i=1}^{n}\varepsilon_{i2}\boldsymbol{X}_i \;\ldots\; \sum_{i=1}^{n}\varepsilon_{im}\boldsymbol{X}_i\Big]^\top\boldsymbol{\Delta}\Big)\right| \\
&\leq \|\boldsymbol{\Delta}\|_* \sup_{\boldsymbol{a}\in\mathcal{S}^{p-1}}\left|\sum_{j=1}^{m}\Big(\sum_{i=1}^{n}\varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{a}\Big)^2\right|^{1/2} \\
&\leq m^{1/2}\|\boldsymbol{\Delta}\|_*\max_{j\leq m}\left\|\sum_{i=1}^{n}\varepsilon_{ij}\boldsymbol{X}_i\right\|, \qquad \text{(S.3.16)}
\end{aligned}
$$

where the first inequality is from Hölder's inequality, and the second inequality is elementary.

Now we apply random matrix theory to bound the right-hand side of (S.3.16). Using matrix dilations (see, for example Section 2.6 of Tropp (2011)), we have

$$
\left\|\sum_{i=1}^{n}\varepsilon_{ij}\boldsymbol{X}_i\right\| = \left\|\sum_{i=1}^{n}\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i \\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\|. \qquad \text{(S.3.17)}
$$

Notice that the random matrix $\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i \\ \boldsymbol{X}_i^\top & 0\end{pmatrix}$ is self adjoint and symmetric conditional on $\boldsymbol{X}_i$. We now obtain

$$
\begin{aligned}
&\mathsf{E}_\varepsilon\left[\exp\left\{2\mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le t\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma})}}\left|n^{-1/2}m^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right|\right\}\right]\\
&\le \mathsf{E}_\varepsilon\left[\exp\left\{2\mu\alpha_r t\max_{j\le m} n^{-1/2}\left\|n^{-1/2}\sum_{i=1}^{n}\varepsilon_{ij}\boldsymbol{X}_i^\top\right\|\right\}\right]\\
&\le m\max_{j\le m}\mathsf{E}_\varepsilon\left[\exp\left\{2\mu\alpha_r t n^{-1/2}\left\|n^{-1/2}\sum_{i=1}^{n}\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\|\right\}\right]\\
&\le m2(p+1)\max_{j\le m}\exp\left\{4\mu^2\alpha_r^2t^2\sigma_{\max}\left(n^{-1}\sum_{i=1}^{n}\log\mathsf{E}_\varepsilon\left[\exp\left\{\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\}\right]\right)\right\}
\end{aligned}
\tag{S.3.18}
$$

where the first inequality is from Lemma S.3.2(ii) and (S.3.16), the second inequality follows from (S.3.17), Lemma S.3.2 (ii) ($\boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma})$), and the fact that

$$
\mathsf{E}[\max_{j\le m}\exp(|Z_j|)]\le m\max_{j\le m}\mathsf{E}[\exp(|Z_j|)],\quad \text{for any random variable } Z_j\in\mathbb{R}.
$$

The third inequality is by Theorem (ii) of Maurer and Pontil (2013) by the symmetric distribution of ε_{ij}, where for a self adjoint matrix $\mathbf{A}$,

$$
\begin{aligned}
\exp(\mathbf{A}) &\overset{\text{def}}{=} \mathbf{I}+\sum_{j=1}^{\infty}\frac{\mathbf{A}^j}{j!}\\
\log(\exp(\mathbf{A})) &\overset{\text{def}}{=} \mathbf{A}.
\end{aligned}
$$

From equation (2.4) on page 399 of Tropp (2011), for any j,

$$\mathsf{E}_\varepsilon\left[\exp\left\{\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\}\right] = \frac{1}{2}\Bigg(\exp\left\{\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\} + \exp\left\{-\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\}\Bigg\}$$
$$\preccurlyeq \exp\left\{\frac{1}{2}\begin{pmatrix}\boldsymbol{X}_i\boldsymbol{X}_i^\top & \mathbf{0}_p\\ 0 & \boldsymbol{X}_i^\top\boldsymbol{X}_i\end{pmatrix}\right\},$$

where "$\mathbf{A} \preccurlyeq \mathbf{B}$" means the $\mathbf{B} - \mathbf{A}$ is positive semidefinite for two matrices $\mathbf{A}, \mathbf{B}$. From equation (2.8) on page 399 of Tropp (2011), the logarithm defined above preserves the order $\preccurlyeq$. Hence, the last inequality in (S.3.18) is bounded by

$$\begin{aligned}
&2m(p+1)\exp\left\{4\mu^2\alpha_r^2t^2\sigma_{\max}\Bigg(n^{-1}\sum_{i=1}^n\frac{1}{2}\begin{pmatrix}\boldsymbol{X}_i\boldsymbol{X}_i^\top & \mathbf{0}_p\\ 0 & \boldsymbol{X}_i^\top\boldsymbol{X}_i\end{pmatrix}\Bigg)\right\}\\
&\le 2m(p+1)\exp\big\{2\mu^2\alpha_r^2t^2\sigma_{\max}(\widehat{\boldsymbol{\Sigma}}_X + B_p)\big\},
\end{aligned} \tag{S.3.19}$$

where the last inequality follows from a bound for the spectral norm for block matrices in equation (2) of Theorem 1 in Bhatia and Kittaneh (1990), and Assumption (A2).

Putting (S.3.19) into (S.3.14), we obtain

$$\begin{aligned}
\mathrm{P}\{\mathcal{A}(t) > s|\Omega\} &\le 8m(p+1)\exp\left(\frac{-\mu s}{4}\right)\mathsf{E}\big[\exp\big\{2\mu^2\alpha_r^2t^2\sigma_{\max}(\widehat{\boldsymbol{\Sigma}}_X + B_p)\big\}\big|\Omega\big]\\
&\le 8m(p+1)\exp\left(\frac{-\mu s}{4}\right)\exp\big\{2\mu^2\alpha_r^2t^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p)\big\}.
\end{aligned} \tag{S.3.20}$$

Minimizing the expression (S.3.20) with respect to μ gives

$$\mathrm{P}\{\mathcal{A}(t) > s|\Omega\} \le 8m(p+1)\exp\left\{-\frac{s^2}{128\alpha_r^2t^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p)}\right\}. \tag{S.3.21}$$

Taking

$$
\begin{aligned}
s &= t8\sqrt{2}c_3\alpha_r\sqrt{(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)}\sqrt{\log m+\log p} \\
&= t32\sqrt{2}c_3\sqrt{r}\sqrt{(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X)+B_p)/\sigma_{\min}(\boldsymbol{\Sigma}_X)}\sqrt{\log m+\log p}.
\end{aligned}
$$

Notice that by the above choice, $s \geq \sqrt{8}t$ for large enough p, m, so that the symmetrization (S.3.13) is valid. Recall that $\mathrm{P}(\Omega) \geq 1-\gamma_n$. The proof is then completed.

□

Remark S.3.4. *Note that both Lemma 2.3.7 of van der Vaart and Wellner (1996) and Lemma 4.12 of Ledoux and Talagrand (1991) applied in the proof of Lemma S.3.3 can be applied on* arbitrary $(Y_{ij}, \boldsymbol{X}_i)$, *regardless whether they are i.i.d. or not. The random matrix theory applied in the proof may also be generalized to matrix martingales; see Section 7 of Tropp (2011) for more details.*

Remark S.3.5. *It can be observed that Lemma S.3.3 is valid uniformly for any* $0 < \tau < 1$.

S.4: Miscellaneous Technical Detail

S.4.1. Detail on Remark 3.7

Suppose $\|\boldsymbol{X}\| \leq B$ for some constant $B > 0$ almost surely, if not, under (A2) this holds with high probability. For any $\boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, a)$, where $a = 0, 2g_n(\kappa)/\lambda$ or $2a_{n,t}(\kappa,\epsilon)/\lambda$,

$$
\begin{aligned}
\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^3] &\leq \sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^2] B\|\boldsymbol{\Delta}\|_* \\
&\leq \Big(\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^2]\Big)^{3/2} B\left(4\sqrt{\frac{r}{\sigma_{\min}(\boldsymbol{\Sigma}_X)}} + \frac{a}{m^{1/2}\|\boldsymbol{\Delta}\|_{L_2(P_X)}}\right)
\end{aligned}
$$

where the first inequality is from Hölder's inequality, the second is from Lemma S.3.2 (ii), Lemma S.2.2 (ii), and Lemma S.2.5 (ii). Hence,

$$\frac{\mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^2]^{3/2}}{\mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^3]} \geq B^{-1}\left(\sqrt{\frac{r}{\sigma_{\min}(\boldsymbol{\Sigma}_X)}} + \frac{a}{m^{1/2}\|\boldsymbol{\Delta}\|_{L_2(P_X)}}\right)^{-1} \tag{S.4.1}$$

Below we discuss three cases corresponding to the conditions required for the theoretical results in Section 3.

Case I: $a = 0$**.** (A.1) holds when r is small and n is large enough. In particular, the right-hand side of (S.4.1) is large when r is small enough. On the other hand, the left-hand side of (A.1) is small whenever n is large enough, because that is a constant multiplied by the rate of $\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{L_2(P_X)}$.

Case II: $a = 2g_n(\kappa)/\lambda$**.** (3.10) holds when r(resp. n) is sufficiently small(resp. large), and the smoothing error $g_n(\kappa)$ is sufficiently small. If $\kappa = \epsilon/(2mn)$, we need to select ϵ small enough.

Case III: $a = 2a_{n,t}(\kappa, \epsilon)/\lambda$**.** (3.14) holds when r(resp. n) is sufficiently small(resp. large), and the rate $a_{n,t}(\kappa, \epsilon)$ is sufficiently small. $a_{n,t}(\kappa, \epsilon)$ is made small when we increase t and choose a small ϵ, if $\kappa = \epsilon/(2mn)$.

S.4.2. Detail on Remark 3.8

We first note an inequality

$$\|\boldsymbol{\Gamma}_{\tau,t}\|_* - \|\boldsymbol{\Gamma}\|_* \leq 2\|\mathcal{P}_{\mathcal{M}}^{\perp}(\boldsymbol{\Gamma})\|_* + \|\mathcal{P}_{\overline{\mathcal{M}}}(\boldsymbol{\Delta}_{\tau,t})\|_* - \|\mathcal{P}_{\overline{\mathcal{M}}}^{\perp}(\boldsymbol{\Delta}_{\tau,t})\|_*, \tag{S.4.2}$$

which can be shown by exactly the same argument for showing inequality (52) in Lemma 3 on page 27 in the supplementary material of Negahban et al. (2012), because the nuclear norm is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$.

It can be shown by similar argument for showing (S.2.11) that

$$
\begin{aligned}
0 &\leq \widehat{Q}_\tau(\mathbf{\Gamma}) - \widehat{Q}_\tau(\mathbf{\Gamma}_{\tau,t}) + \lambda\|\mathbf{\Gamma}\|_* - \lambda\|\mathbf{\Gamma}_{\tau,t}\|_* + R_{n,t}(\kappa,\epsilon) \\
&\leq \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|\big(\|\mathcal{P}_{\overline{\mathcal{M}}}(\mathbf{\Delta}_{\tau,t})\|_* + \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Delta}_{\tau,t})\|_*\big) \\
&\qquad + \lambda(2\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma})\|_* + \|\mathcal{P}_{\overline{\mathcal{M}}}(\mathbf{\Delta}_{\tau,t})\|_* - \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Delta}_{\tau,t})\|_*) + R_{n,t}(\kappa,\epsilon),
\end{aligned} \tag{S.4.3}
$$

where the first inequality follows by the first three lines in (S.2.11), and the second inequality is from (S.4.2).

Rearrange expression (S.4.3) to get,

$$
(\lambda - \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* \leq (\lambda + \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* + 2\lambda\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma})\|_* + R_{n,t}(\kappa,\epsilon).
$$

By $\lambda \geq 2\|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|$,

$$
\frac{1}{2}\lambda\|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* \leq \frac{3}{2}\lambda\|\mathcal{P}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* + 2\lambda\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma})\|_* + R_{n,t}(\kappa,\epsilon).
$$

As argued in the proof for Lemma S.2.4, we have $\mathrm{P}(R_{n,t}(\kappa,\epsilon) \leq a_{n,t}(\kappa,\epsilon)) \geq 1 - \gamma_n - 16(pm)^{1-c_3^2} - 3\exp\{-(p+m)\log 8\}$. Thus, the proof for (3.16) is completed.

S.4.3. Details for Generating matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ in Section 4

Given (r_1, r_2), $\mathbf{S}_1$ and $\mathbf{S}_2$ are selected with the following procedure:

1. Generate vectors $\{\boldsymbol{a}_1, ..., \boldsymbol{a}_{r_1}\}$ and $\{\boldsymbol{b}_1, ..., \boldsymbol{b}_{r_2}\}$, where $\boldsymbol{a}_{j_1}, \boldsymbol{b}_{j_2} \in \mathbb{R}^p$, and $a_{j_1k_1}, b_{j_2k_2} \sim U(0,1)$ i.i.d. for $j_1 = 1, ..., r_1$, $j_2 = 1, ..., r_2$, $k_1, k_2 = 1, ..., p$;

2. Set the columns of $\mathbf{S}_1$ and $\mathbf{S}_2$ by $(\mathbf{S}_1)_{*j} = \sum_{k=1}^{r_1} \alpha_{k,j}\boldsymbol{a}_k$ and $(\mathbf{S}_2)_{*j} = \sum_{k=1}^{r_2} \beta_{k,j}\boldsymbol{b}_k$ for $j = 1, ..., m$, where $\alpha_{k,j}$, $\beta_{k,j}$ are independent random variables in $U[0,1]$ for $k = 1, ..., p$ and $j = 1, ..., m$.

In our simulation, the first two nonzero singular values for $\mathbf{S}_1$ are $(\sigma_1(\mathbf{S}_1), \sigma_2(\mathbf{S}_1)) = (179.91, 26.51)$ and the rest singular value is 0. For $\mathbf{S}_2^{ES}$, the first two nonzero singular values are $(\sigma_1(\mathbf{S}_2^{ES}), \sigma_2(\mathbf{S}_2^{ES})) = (175.48, 25.74)$ and the rest is 0. For $\mathbf{S}_2^{ES}$, the first six nonzero singular values are $(\sigma_1(\mathbf{S}_2^{AS}), ..., \sigma_6(\mathbf{S}_2^{AS})) = (473.40, 29.87, 25.66, 23.89, 23.58, 22.16)$ and the rest is 0.

S.4.4. Detail on Mean Removing

Estimation of mean function and smoothing are done jointly by minimizing

$$\widehat{\mu}(s) \stackrel{\text{def}}{=} \arg\min_{\mu \in \mathbb{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[Y_{ij} - \mu(i/365)\right]^2 + \eta \int [D^2 \mu(s)]^2 ds \tag{S.4.4}$$

where $\eta > 0$ is a smoothing parameter selected by generalized cross-validation, and $\mathbb{S}$ is a space of cubic B-splines. The computation is performed with the command `smooth.spline` in `R`.

S.5: Auxiliary Lemmas

Definition S.4.1. *Let $\mathcal{X} = \mathbb{R}^{p \times n}$ with inner product $\langle \mathbf{A}, \mathbf{B} \rangle = tr(\mathbf{A}^\top \mathbf{B})$ and $\|\cdot\|$ be the induced norm. $f : \mathcal{X} \to \mathbb{R}$ a lower semicontinuous convex function. The* proximity operator of f*, $S_f : \mathcal{X} \to \mathcal{X}$:*

$$S_f(\mathbf{Y}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{X} \in \mathcal{X}} \left\{ f(\mathbf{X}) + \frac{1}{2}\|\mathbf{X} - \mathbf{Y}\|^2 \right\}, \forall \mathbf{Y} \in \mathcal{X}.$$

Theorem S.4.2 (Theorem 2.1 of Cai et al. (2010))**.** *Suppose the singular decomposition of $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{p \times m}$, where $\mathbf{D}$ is a $p \times m$ rectangular diagonal matrix and $\mathbf{U}$ and $\mathbf{V}$ are* *unitary matrices. The proximity operator $S_\lambda(\cdot)$ associated with $\lambda\|\cdot\|_*$ is*

$$S_\lambda(\mathbf{Y}) \stackrel{\text{def}}{=} \mathbf{U}(\mathbf{D}-\lambda\mathbf{I}_{pm})_+\mathbf{V}^\top, \tag{S.5.1}$$

where $\mathbf{I}_{pm}$ is the $p\times m$ rectangular identity matrix with diagonal elements equal to 1.

Lemma S.4.3 (Hoeffding's Inequality, Proposition 5.10 of Vershynin (2012a))**.** *Let $X_1, ..., X_n$ be independent centered sub-gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $\mathbf{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$\mathrm{P}\bigg(\bigg|\sum_{i=1}^n a_i X_i\bigg| \geq t\bigg) \leq e\cdot\exp\bigg(-\frac{C' t^2}{K^2\|\mathbf{a}\|_2^2}\bigg),$$

where $C' > 0$ is a universal constant.

Lemma S.4.4 (Hoeffding's Inequality: classical form)**.** *Let $X_1, ..., X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely, then*

$$\mathrm{P}\bigg(\bigg|\sum_{i=1}^n X_i\bigg| \geq t\bigg) \leq 2\exp\bigg(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\bigg).$$

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Downside risk and stock returns: An empirical analysis of the long-run and short-run dynamics from the G-7 Countries" by Cathy Yi-Hsuan Chen, Thomas C. Chiang and Wolfgang Karl Härdle, January 2016.
002 "Uncertainty and Employment Dynamics in the Euro Area and the US" by Aleksei Netsunajev and Katharina Glass, January 2016.
003 "College Admissions with Entrance Exams: Centralized versus Decentralized" by Isa E. Hafalir, Rustamdjan Hakimov, Dorothea Kübler and Morimitsu Kurino, January 2016.
004 "Leveraged ETF options implied volatility paradox: a statistical study" by Wolfgang Karl Härdle, Sergey Nasekin and Zhiwu Hong, February 2016.
005 "The German Labor Market Miracle, 2003 -2015: An Assessment" by Michael C. Burda, February 2016.
006 "What Derives the Bond Portfolio Value-at-Risk: Information Roles of Macroeconomic and Financial Stress Factors" by Anthony H. Tu and Cathy Yi-Hsuan Chen, February 2016.
007 "Budget-neutral fiscal rules targeting inflation differentials" by Maren Brede, February 2016.
008 "Measuring the benefit from reducing income inequality in terms of GDP" by Simon Voigts, February 2016.
009 "Solving DSGE Portfolio Choice Models with Asymmetric Countries" by Grzegorz R. Dlugoszek, February 2016.
010 "No Role for the Hartz Reforms? Demand and Supply Factors in the German Labor Market, 1993-2014" by Michael C. Burda and Stefanie Seele, February 2016.
011 "Cognitive Load Increases Risk Aversion" by Holger Gerhardt, Guido P. Biele, Hauke R. Heekeren, and Harald Uhlig, March 2016.
012 "Neighborhood Effects in Wind Farm Performance: An Econometric Approach" by Matthias Ritter, Simone Pieralli and Martin Odening, March 2016.
013 "The importance of time-varying parameters in new Keynesian models with zero lower bound" by Julien Albertini and Hong Lan, March 2016.
014 "Aggregate Employment, Job Polarization and Inequalities: A Transatlantic Perspective" by Julien Albertini and Jean Olivier Hairault, March 2016.
015 "The Anchoring of Inflation Expectations in the Short and in the Long Run" by Dieter Nautz, Aleksei Netsunajev and Till Strohsal, March 2016.
016 "Irrational Exuberance and Herding in Financial Markets" by Christopher Boortz, March 2016.
017 "Calculating Joint Confidence Bands for Impulse Response Functions using Highest Density Regions" by Helmut Lütkepohl, Anna Staszewska-Bystrova and Peter Winker, March 2016.
018 "Factorisable Sparse Tail Event Curves with Expectiles" by Wolfgang K. Härdle, Chen Huang and Shih-Kang Chao, March 2016.
019 "International dynamics of inflation expectations" by Aleksei Netšunajev and Lars Winkelmann, May 2016.
020 "Academic Ranking Scales in Economics: Prediction and Imdputation" by Alona Zharova, Andrija Mihoci and Wolfgang Karl Härdle, May 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

021 "CRIX or evaluating blockchain based currencies" by Simon Trimborn and Wolfgang Karl Härdle, May 2016.
022 "Towards a national indicator for urban green space provision and environmental inequalities in Germany: Method and findings" by Henry Wüstemann, Dennis Kalisch, June 2016.
023 "A Mortality Model for Multi-populations: A Semi-Parametric Approach" by Lei Fang, Wolfgang K. Härdle and Juhyun Park, June 2016.
024 "Simultaneous Inference for the Partially Linear Model with a Multivariate Unknown Function when the Covariates are Measured with Errors" by Kun Ho Kim, Shih-Kang Chao and Wolfgang K. Härdle, August 2016.
025 "Forecasting Limit Order Book Liquidity Supply-Demand Curves with Functional AutoRegressive Dynamics" by Ying Chen, Wee Song Chua and Wolfgang K. Härdle, August 2016.
026 "VAT multipliers and pass-through dynamics" by Simon Voigts, August 2016.
027 "Can a Bonus Overcome Moral Hazard? An Experiment on Voluntary Payments, Competition, and Reputation in Markets for Expert Services" by Vera Angelova and Tobias Regner, August 2016.
028 "Relative Performance of Liability Rules: Experimental Evidence" by Vera Angelova, Giuseppe Attanasi, Yolande Hiriart, August 2016.
029 "What renders financial advisors less treacherous? On commissions and reciprocity" by Vera Angelova, August 2016.
030 "Do voluntary payments to advisors improve the quality of financial advice? An experimental sender-receiver game" by Vera Angelova and Tobias Regner, August 2016.
031 "A first econometric analysis of the CRIX family" by Shi Chen, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, TM Lee and Bobby Ong, August 2016.
032 "Specification Testing in Nonparametric Instrumental Quantile Regression" by Christoph Breunig, August 2016.
033 "Functional Principal Component Analysis for Derivatives of Multivariate Curves" by Maria Grith, Wolfgang K. Härdle, Alois Kneip and Heiko Wagner, August 2016.
034 "Blooming Landscapes in the West? - German reunification and the price of land." by Raphael Schoettler and Nikolaus Wolf, September 2016.
035 "Time-Adaptive Probabilistic Forecasts of Electricity Spot Prices with Application to Risk Management." by Brenda López Cabrera , Franziska Schulz, September 2016.
036 "Protecting Unsophisticated Applicants in School Choice through Information Disclosure" by Christian Basteck and Marco Mantovani, September 2016.
037 "Cognitive Ability and Games of School Choice" by Christian Basteck and Marco Mantovani, Oktober 2016.
038 "The Cross-Section of Crypto-Currencies as Financial Assets: An Overview" by Hermann Elendner, Simon Trimborn, Bobby Ong and Teik Ming Lee, Oktober 2016.
039 "Disinflation and the Phillips Curve: Israel 1986-2015" by Rafi Melnick and Till Strohsal, Oktober 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2016

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

040 "Principal Component Analysis in an Asymmetric Norm" by Ngoc M. Tran, Petra Burdejová, Maria Osipenko and Wolfgang K. Härdle, October 2016.
041 "Forward Guidance under Disagreement - Evidence from the Fed's Dot Projections" by Gunda-Alexandra Detmers, October 2016.
042 "The Impact of a Negative Labor Demand Shock on Fertility - Evidence from the Fall of the Berlin Wall" by Hannah Liepmann, October 2016.
043 "Implications of Shadow Bank Regulation for Monetary Policy at the Zero Lower Bound" by Falk Mazelis, October 2016.
044 "Dynamic Contracting with Long-Term Consequences: Optimal CEO Compensation and Turnover" by Suvi Vasama, October 2016.
045 "Information Acquisition and Liquidity Dry-Ups" by Philipp Koenig and David Pothier, October 2016.
046 "Credit Rating Score Analysis" by Wolfgang Karl Härdle, Phoon Kok Fai and David Lee Kuo Chuen, November 2016.
047 "Time Varying Quantile Lasso" by Lenka Zbonakova, Wolfgang Karl Härdle, Phoon Kok Fai and Weining Wang, November 2016.
048 "Unraveling of Cooperation in Dynamic Collaboration" by Suvi Vasama, November 2016.
049 "Q3-D3-LSA" by Lukas Borke and Wolfgang K. Härdle, November 2016.
050 "Network Quantile Autoregression" by Xuening Zhu, Weining Wang, Hangsheng Wang and Wolfgang Karl Härdle, November 2016.
051 "Dynamic Topic Modelling for Cryptocurrency Community Forums" by Marco Linton, Ernie Gin Swee Teo, Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, November 2016.
052 "Beta-boosted ensemble for big credit scoring data" by Maciej Zieba and Wolfgang Karl Härdle, November 2016.
053 "Central Bank Reputation, Cheap Talk and Transparency as Substitutes for Commitment: Experimental Evidence" by John Duffy and Frank Heinemann, December 2016.
054 "Labor Market Frictions and Monetary Policy Design" by Anna Almosova, December 2016.
055 "Effect of Particulate Air Pollution on Coronary Heart Disease in China: Evidence from Threshold GAM and Bayesian Hierarchical Model" by Xiaoyu Chen, December 2016.
056 "The Effect of House Price on Stock Market Participation in China: Evidence from the CHFS Micro-Datal" by Xiaoyu Chen and Xiaohao Ji, December 2016.
057 "Factorisable Multi-Task Quantile Regression" by Shih-Kang Chao, Wolfgang K. Härdle and Ming Yuan, December 2016.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".