Supplementary Material to "Factorisable Multitask Quantile Regression"*

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

Section S.1 presents the convergence analysis for the algorithm. Section S.2 provides details on the non-asymptotic risk analysis of $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$. Section S.3 covers miscellaneous technical details. Section S.4 lists some auxiliary results used in our proof.

Additional notation. 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 $\mathbb{P}_n$. For a function $f : \mathbb{R}^p \to \mathbb{R}$, and $\boldsymbol{Z}_i \in \mathbb{R}^p$, define the *empirical process* $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n \{f(\boldsymbol{Z}_i) - \mathsf{E}[f(\boldsymbol{Z}_i)]\}$. The subgradient for $\widehat{Q}_\tau(\mathbf{S})$ is the matrix

$$\nabla\widehat{Q}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} (nm)^{-1}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{W}_{\tau,i*}(\mathbf{S})^\top \stackrel{\text{def}}{=} (nm)^{-1}\mathbf{X}^\top\mathbf{W}_\tau(\mathbf{S}) \in \mathbb{R}^{p\times m}, \tag{0.1}$$

where

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

For the true coefficient matrix $\mathbf{\Gamma}_\tau$, $\boldsymbol{W}_{\tau,i*}(\mathbf{\Gamma}_\tau) \stackrel{\text{def}}{=} \boldsymbol{W}_{\tau,i*}$ and $\mathbf{W}_\tau \stackrel{\text{def}}{=} \mathbf{W}_\tau(\mathbf{\Gamma}_\tau)$.

S.1: Proofs for Algorithmic Convergence Analysis

S.1.1 Proof of (A.2)

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

*Address correspondence to Shih-Kang Chao, Department of Statistics, University of Missouri, Columbia, MO 65211, USA; e-mail: chaosh@missouri.edu.

†University of Missouri

‡Humboldt-Universität zu Berlin

§Columbia University

Remark S.1.1. *It is necessary to choose* $[\tau-1,\tau]$ *rather than* $\{\tau-1,\tau\}$ *for the support of* Θ_{ij} *in (A.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) are fulfilled.*

S.1.2 Proof of Theorem 2.1

Recall the definition of $L_\tau(\mathbf{S})$ and $\widehat{Q}_\tau(\mathbf{S})$ in (A.1), $\widetilde{L}_\tau(\mathbf{S})$ and $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ in (A.5) and (A.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

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

Recall that $\widehat{\boldsymbol{\Gamma}}_\tau$ is a minimizer of $L_\tau(\mathbf{S})$ defined in (A.1). It follows by (S.1.1) that for an arbitrary $\mathbf{S}\in\mathbb{R}^{p\times m}$,

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

where the first inequality is from the first inequality of (S.1.1), the second is the definition of the minimizer $\widehat{\boldsymbol{\Gamma}}_\tau$, 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

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

where the first inequality 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 an arbitrary matrix $\mathbf{S}\in\mathbb{R}^{p\times m}$.

Now from triangle inequality,

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

The third term on the right-hand side of (S.1.4) is bounded by (S.1.3). For any matrix $\mathbf{S}$, we have from (S.1.1) that

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

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

Lemma S.1.3 implies that the gradient of $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ 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

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

where $M = (\kappa m^2 n^2)^{-1}\|\mathbf{X}\|^2$ as given in Lemma S.1.3.

S.1.3 Technical Details for Theorem 2.1

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

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

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

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 $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{S}) = -(mn)^{-1}\mathbf{X}^\top\mathbf{\Theta}^*(\mathbf{S}) \in \mathbb{R}^{p\times m}$, where $\mathbf{\Theta}^*(\mathbf{S})$ is the optimal solution to (A.3), namely*

$$\mathbf{\Theta}^*(\mathbf{S}) = [[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{XS})]]_\tau. \tag{S.1.7}$$

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 (A.3) that

$$\widehat{Q}_{\tau,\kappa}(\mathbf{S}) = \max_{\Theta_{ij}\in[\tau-1,\tau]} \left\{(mn)^{-1}\langle \mathbf{Y}, \mathbf{\Theta}\rangle + (mn)^{-1}\langle -\mathbf{XS}, \mathbf{\Theta}\rangle - \frac{\kappa}{2}\|\mathbf{\Theta}\|_{\mathrm{F}}^2\right\}. \tag{S.1.8}$$

$\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ 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 $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{S}) =$

$-(mn)^{-1}\mathbf{X}^\top\boldsymbol{\Theta}^*(\mathbf{S}) \in \mathbb{R}^{p\times m}$, where $\boldsymbol{\Theta}^*(\mathbf{S})$ is the optimal solution to (A.3):

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

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

Remark S.2.1. *For any $\boldsymbol{\Delta} \in \mathbb{R}^{p\times m}$, from (A2),*

$$\|\boldsymbol{\Delta}\|_{L_2(P_X)}^2 = m^{-1}\mathsf{E}\big[\|\boldsymbol{\Delta}^\top \boldsymbol{X}_i\|_2^2\big] = m^{-1}\sum_{j=1}^m \boldsymbol{\Delta}_{*j}^\top \mathsf{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\top]\boldsymbol{\Delta}_{*j} \geq m^{-1}\sigma_{\min}(\boldsymbol{\Sigma}_X)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2. \quad \text{(S.2.1)}$$

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

$$\|\boldsymbol{\Delta}\|_{L_2(P_X)} \geq \Big(\frac{\sigma_{\min}(\boldsymbol{\Sigma}_X)}{m}\Big)^{1/2}\|\boldsymbol{\Delta}\|_{\mathrm{F}} \geq \Big(\frac{\sigma_{\min}(\boldsymbol{\Sigma}_X)}{m}\Big)^{1/2}\|\mathcal{P}_{\boldsymbol{\Gamma}_\tau}(\boldsymbol{\Delta})\|_{\mathrm{F}}. \quad \text{(S.2.2)}$$

S.2.1 Proof for Lemma 3.1

To prove the first statement, 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

$$\mathrm{P}\bigg(\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}_\tau\| \geq 4s\bigg) = \mathrm{P}\bigg(\sup_{\substack{\mathbf{v}\in S^{p-1}\\ \mathbf{u}\in\mathcal{S}^{m-1}}}\frac{1}{n}\big|\mathbf{v}^\top\mathbf{X}^\top\mathbf{W}_\tau\mathbf{u}\big| \geq 4s\bigg) \leq 8^{p+m}\sup_{\substack{\mathbf{v}\in S^{p-1},\mathbf{u}\in\mathcal{S}^{m-1}\\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}}\mathrm{P}\bigg(\frac{|\langle\mathbf{X}\mathbf{v},\mathbf{W}_\tau\mathbf{u}\rangle|}{n} \geq s\bigg). \quad \text{(S.2.3)}$$

To bound $n^{-1}\langle\mathbf{X}\mathbf{v},\mathbf{W}_\tau\mathbf{u}\rangle = n^{-1}\sum_{i=1}^n\langle\mathbf{v},\boldsymbol{X}_i\rangle\langle\mathbf{u},\boldsymbol{W}_{\tau,i*}\rangle$, first we show the sub-Gaussianity of $\langle\mathbf{u},\boldsymbol{W}_{\tau,i*}\rangle$. Theorem 3.1 of Buldygin and Moskvichova (2013) suggests that the sub-Gaussian norm of the jth component of $\boldsymbol{W}_{\tau,i*}$ is

$$\|W_{\tau,ij}\|_{\psi_2} = \begin{cases} 0, & \tau = 0, 1;\\ \frac{2\tau-1}{2\{\log\tau-\log(1-\tau)\}}, & \tau \in (0,1) - \{1/2\};\\ 1/4, & \tau = 1/2, \end{cases}$$

where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. It follows by Lemma S.4.3 (Hoeffding's inequality) that

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

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

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

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,

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

Therefore, for arbitrary $u > 1$, the event

$$\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}_\tau\| \geq 4\cdot\sqrt{u(\log 8)\frac{K(\tau)c_2\|\boldsymbol{\Sigma}_X\|}{C'}}\sqrt{\frac{p+m}{n}}, \tag{S.2.4}$$

has probability smaller than $3e^{-(u-1)(p+m)\log 8} + \gamma_n$, as $e < 3$.

To prove the second statement, we note that the event in (S.2.4) has probability less than η by setting $k = 1 - (\eta - \gamma_n)/(3(p+m)\log 8)$.

□

S.2.2 Proof for Theorem 3.2

Before we prove Theorem 3.2, we first define the "support" of matrices by projections.

Definition S.2.2. *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 $\mathbf{U}_{[1:r]} = [\mathbf{u}_1 ... \mathbf{u}_r] \in \mathbb{R}^{p\times r}$; $\mathbf{P}_2 \overset{\text{def}}{=} \mathbf{V}_{[1:r]}\mathbf{V}_{[1:r]}^\top$, where $\mathbf{V}_{[1:r]} = [\mathbf{v}_1 ... \mathbf{v}_r] \in \mathbb{R}^{m\times r}$. 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*

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

Define for any $a \geq 0$,

$$\mathcal{K}(\mathbf{\Gamma}_\tau; a) \stackrel{\text{def}}{=} \left\{\mathbf{S} \in \mathbb{R}^{p\times m} : \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\mathbf{S})\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{S})\|_* + a\right\}. \tag{S.2.5}$$

We note that the nuclear norm is *decomposable* under the projection: for any $\mathbf{S}, \mathbf{A} \in \mathbb{R}^{p\times m}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathbf{A}}(\mathbf{S})\|_* + \|\mathcal{P}^\perp_{\mathbf{A}}(\mathbf{S})\|_*$. This is analogous to the ℓ_1 norm for vectors: for any vector $\boldsymbol{v}$ and support S, $\|\boldsymbol{v}\|_1 = \|\boldsymbol{v}_S\|_1 + \|\boldsymbol{v}_{S^c}\|_1$; see Definition 1 on page 541 of Negahban et al. (2012). Moreover, the rank of $\mathcal{P}_{\mathbf{A}}(\mathbf{S})$ is at most $\text{rank}(\mathbf{A})$.

The shape of $\mathcal{K}(\mathbf{\Gamma}_\tau; a)$ 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). See also their Figure 1 on page 544.

To simplify the notations in this proof, let

$$\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau, \tag{S.2.6}$$

$$\alpha_r = 4\sqrt{r/\sigma_{\min}(\mathbf{\Sigma}_X)}, \tag{S.2.7}$$

$$\alpha_{r,m} = m^{1/2}\alpha_r, \tag{S.2.8}$$

$$c_n = 16\sqrt{2}m^{-1/2}\delta\lambda^{-1}\sqrt{c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p}\sqrt{\log m + \log p}, \tag{S.2.9}$$

$$d_n = 8\sqrt{2}\alpha_r\sqrt{c_2\sigma_{\max}(\mathbf{\Sigma}_X) + B_p}\sqrt{\log m + \log p}. \tag{S.2.10}$$

Let the events

Ω_1 : Assumption (A2) holds;

$\Omega_2 : \mathcal{A}(t) \leq u(td_n + c_n)$ for $u > 1$, where

$$\mathcal{A}(t) \stackrel{\text{def}}{=} \sup_{\|\mathbf{\Delta}\|_{L_2(P_X)}\leq t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)} \left|\mathbb{G}_n\left[m^{-1}\sum_{j=1}^m \left(\rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top(\mathbf{\Gamma}_{\tau,*j} + \mathbf{\Delta}_{*j})\} - \rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}\}\right)\right]\right|. \tag{S.2.11}$$

$$\Omega_3 : \frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq C^*\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)}\sqrt{\frac{p+m}{n}},$$

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

The probability of event $\text{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \geq 1 - \gamma_n - 16(pm)^{1-u^2} - 3e^{-(p+m)\log 8}$ by Assumption (A2), Lemma 3.1 and Lemma S.2.5.

Recall that $\alpha_{r,m}$, c_n and d_n are defined in (S.2.8), (S.2.9) and (S.2.10). Set

$$t = \sqrt{n^{-1/2} u c_n \frac{4}{\underline{f}^{\tau}} + \frac{8}{\underline{f}^{\tau}}\delta} + \frac{4}{\underline{f}^{\tau}}(n^{-1/2} u d_n + \lambda \alpha_{r,m}). \tag{S.2.12}$$

We will prove by contradiction. Suppose to the contrary that $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} \geq t$. Since $\widehat{\mathbf{\Gamma}}_\tau$ minimizes $L_\tau(\mathbf{S}) = \widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*$ (defined in (1.3)) and $L_\tau(\widehat{\mathbf{\Gamma}}_\tau) - L_\tau(\mathbf{\Gamma}_\tau) < 0$, we have

$$\begin{aligned}
&\widehat{Q}_\tau(\mathbf{\Gamma}_\tau + \widehat{\mathbf{\Delta}}) - \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) + \lambda(\|\mathbf{\Gamma}_\tau + \widehat{\mathbf{\Delta}}\|_* - \|\mathbf{\Gamma}_\tau\|_*) \\
&= L_\tau(\widehat{\mathbf{\Gamma}}_\tau) - L_\tau(\mathbf{\Gamma}_\tau) + L_\tau(\mathbf{\Gamma}_\tau + \widehat{\mathbf{\Delta}}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau) \\
&\leq \delta,
\end{aligned} \tag{S.2.13}$$

where we recall (2.1).

Observe that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau \in \mathcal{K}(\mathbf{\Gamma}_\tau; 0) \subset \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)$ with probability $1-\eta$ by applying (3.6) and Lemma S.2.3. Hence, from (2.1),

$$\delta > \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)} \geq t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)} \widehat{Q}_\tau(\mathbf{\Gamma}_\tau + \mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) + \lambda(\|\mathbf{\Gamma}_\tau + \mathbf{\Delta}\|_* - \|\mathbf{\Gamma}_\tau\|_*). \tag{S.2.14}$$

Note the facts that

1. $\widehat{Q}_\tau(\cdot) + \lambda\|\cdot\|_*$ is convex (unique optimum);
2. $\mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)$ is star-shaped (see Figure 1 of Negahban et al. (2012)).

Hence, $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} \geq t$ can be replaced by $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} = t$ and the strict inequality in (S.2.14) is maintained

$$\delta \geq \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)} \widehat{Q}_\tau(\mathbf{\Gamma}_\tau + \mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) + \lambda(\|\mathbf{\Gamma}_\tau + \mathbf{\Delta}\|_* - \|\mathbf{\Gamma}_\tau\|_*).$$

It can be deducted from the last display that

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

By triangle inequality, $\big|\|\mathbf{\Gamma}_\tau + \mathbf{\Delta}\|_* - \|\mathbf{\Gamma}_\tau\|_*\big| \leq \|\mathbf{\Delta}\|_* \leq \alpha_{r,m} t + 2\delta/\lambda$ on the set $\{\|\mathbf{\Delta}\|_{L_2(P_X)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)\}$ by Lemma S.2.4(ii). Applying the bound in Ω_2 obtains

$$\delta \geq \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)} Q_\tau(\mathbf{\Gamma}_\tau + \mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}_\tau) - n^{-1/2} u(d_n t + c_n) - \lambda(\alpha_{r,m} t + 2\delta/\lambda).$$

Since $\delta \le C\lambda\sqrt{m/n}$, by Remark 3.3,

$$\nu_\tau(2\delta/\lambda) \ge \nu_\tau(2C\sqrt{m/n}) > u\epsilon_{n,\tau,r} \ge t/4$$

(where the second inequality is from (3.7); the last inequality will be shown in (S.2.18) below), invoking Lemma S.2.4 (i) to get the minorization

$$\delta \ge \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma}_\tau;2\delta/\lambda)} \frac{1}{4}\underline{f}^\tau t^2 - n^{-1/2}u(d_n t + c_n) - \lambda(\alpha_{r,m}t + 2\delta/\lambda). \tag{S.2.15}$$

Rearranging terms to get

$$0 \ge \inf_{\|\mathbf{\Delta}\|_{L_2(P_X)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma}_\tau;2\delta/\lambda)} \frac{1}{4}\underline{f}^\tau t^2 - n^{-1/2}u(d_n t + c_n) - \lambda\alpha_{r,m}t - 3\delta. \tag{S.2.16}$$

However, the right-hand side of (S.2.16) is strictly greater than 0 whenever

$$t > \frac{2}{\underline{f}^\tau}(n^{-1/2}ud_n + \lambda\alpha_{r,m}) + \frac{2}{\underline{f}^\tau}\sqrt{(n^{-1/2}ud_n + \lambda\alpha_{r,m})^2 + \underline{f}^\tau(n^{-1/2}uc_n + 3\delta)}. \tag{S.2.17}$$

The right hand side of the last display is upper bounded by (by $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ for all $a, b > 0$)

$$t = \frac{2}{\underline{f}^\tau}(n^{-1/2}ud_n + \lambda\alpha_{r,m}) + \frac{2}{\underline{f}^\tau}(n^{-1/2}ud_n + \lambda\alpha_{r,m}) + \sqrt{\frac{4}{\underline{f}^\tau}n^{-1/2}uc_n + \frac{12}{\underline{f}^\tau}\delta},$$

which leads to the t in (S.2.12). We get a contradiction, so $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} \ge t$ does not hold. Namely, $\|\widehat{\mathbf{\Delta}}\|_{L_2(P_X)} < t$.

To show (3.8), we will prove

$$t \le u\epsilon_{n,\tau,r}, \tag{S.2.18}$$

where $\epsilon_{n,\tau,r}$ is defined in (3.7). To see this, first note that,

$$\begin{aligned}\lambda \overset{(3.6)}{\le} 2\bar{\lambda} \overset{(3.4)}{\le} 2\frac{C^*}{m}\sqrt{\Big(1 - \frac{\eta - \gamma_n}{3(p+m)\log 8}\Big)\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)}\sqrt{\frac{p+m}{n}} \\ \le \frac{2C^*}{m}\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)}\sqrt{\frac{p+m}{n}}\end{aligned} \tag{S.2.19}$$

since $0 < \eta < 1$ and $\gamma_n \to 0$.

Elementary calculation shows that for $u \geq 1$,

$$\begin{aligned}
&\max\left\{2\lambda\alpha_{r,m}/\underline{f}^{\tau}, 2n^{-1/2}ud_n/\underline{f}^{\tau}\right\} \\
&\quad \leq \frac{2(32\sqrt{2}+8C^*)u}{\underline{f}^{\tau}\wedge 1}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)\vee 1}{\sigma_{\min}(\mathbf{\Sigma}_X)\wedge 1}}\sqrt{\frac{r(m+p\vee B_p)(\log p+\log m)}{mn}}. \qquad \text{(S.2.20)}
\end{aligned}$$

Under the condition that $\delta < \lambda m^{1/2}n^{-1/2}$, $r \geq 1$,

$$\begin{aligned}
\sqrt{\frac{1}{\underline{f}^{\tau}}n^{-1/2}uc_n} &\leq \sqrt{\frac{u}{\underline{f}^{\tau}}\frac{d_n}{\alpha_r n}} \leq \alpha_r^{-1/2}\frac{ud_n}{(\underline{f}^{\tau}\wedge 1)\sqrt{n}} \leq \frac{(\sigma_{\min}(\mathbf{\Sigma}_X)^{1/2}\vee 1)ud_n}{(\underline{f}^{\tau}\wedge 1)\sqrt{n}} \\
&\leq \frac{(32\sqrt{2}+8C^*)u}{\underline{f}^{\tau}\wedge 1}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)\vee 1}{\sigma_{\min}(\mathbf{\Sigma}_X)\wedge 1}}\sqrt{\frac{r(m+p\vee B_p)(\log p+\log m)}{mn}} \qquad \text{(S.2.21)}
\end{aligned}$$

since $u \geq 1$, $d_n \geq 1$ (as $m, p \to \infty$).

Lastly, again from $\delta < \lambda m^{1/2}n^{-1/2}$ and (S.2.19),

$$\begin{aligned}
\delta &\leq \lambda m^{1/2}n^{-1/2} \leq 2C^*\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)\frac{p+m}{n^2m}} \leq C^*n^{-1}\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)\frac{p+m}{m}} \\
&\leq C^*\frac{p+m}{nm}\sqrt{\sigma_{\max}(\mathbf{\Sigma}_X)}, \qquad \text{(S.2.22)}
\end{aligned}$$

where in the third inequality the fact $\sup_{\tau}|K(\tau)| \leq 1/4$ (noted below Lemma 3.1, or in (K4) of Lemma 2.1 on p.35 of Buldygin and Moskvichova (2013)) is applied, where $K(\tau)$ is defined in (3.3); in the last inequality, the fact $\sqrt{1+p/m} \leq 1+p/m$ is applied. Hence,

$$\begin{aligned}
\sqrt{\frac{1}{\underline{f}^{\tau}}\delta} &\leq \frac{1}{\underline{f}^{\tau}\wedge 1}\sqrt{C^*\frac{p+m}{nm}}\sigma_{\max}(\mathbf{\Sigma}_X)^{1/4} \\
&\leq \frac{(32\sqrt{2}+8C^*)u}{\underline{f}^{\tau}\wedge 1}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)\vee 1}{\sigma_{\min}(\mathbf{\Sigma}_X)\wedge 1}}\sqrt{\frac{r(m+p\vee B_p)(\log p+\log m)}{mn}} \qquad \text{(S.2.23)}
\end{aligned}$$

where the inequality follows by the facts:

- $\dfrac{\sqrt{C^*}}{\underline{f}^\tau \wedge 1} \le \dfrac{C^*}{\underline{f}^\tau \wedge 1} \le \dfrac{(32\sqrt{2}+8C^*)u}{\underline{f}^\tau \wedge 1}$, $\quad$ ($u > 1$ from the hypothesis of the Theorem, and $C^* \ge 1$ from Lemma 3.1)
- $\sigma_{\max}(\mathbf{\Sigma}_X)^{1/4} \le (\sigma_{\max}(\mathbf{\Sigma}_X) \vee 1)^{1/4} \le (\sigma_{\max}(\mathbf{\Sigma}_X) \vee 1)^{1/2} \le \sqrt{\dfrac{\sigma_{\max}(\mathbf{\Sigma}_X) \vee 1}{\sigma_{\min}(\mathbf{\Sigma}_X) \wedge 1}}$
- $\sqrt{\dfrac{p+m}{nm}} \le \sqrt{\dfrac{r(m+p \vee B_p)(\log p + \log m)}{mn}}$, $\quad$ $B_p \ge 1$ by (A2), $r \ge 1$, $p, m \ge 3$ in (A1).

Note that if $r = \operatorname{rank}(\mathbf{\Gamma}_\tau) = 0$, then the matrix $\mathbf{\Gamma}_\tau = 0$ and this case is excluded.

Combining (S.2.20), (S.2.21) and (S.2.23) gives (S.2.18).

□

S.2.3 Technical Details for Theorem 3.2

The following lemma asserts that $\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau$ lies in the cone $\mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)$.

Lemma S.2.3. *Suppose $\lambda \ge 2\|\nabla \widehat{Q}(\mathbf{\Gamma}_\tau)\|$ and $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau$, where $\nabla \widehat{Q}(\mathbf{\Gamma}_\tau)$ is the subgradient of $\widehat{Q}(\mathbf{\Gamma}_\tau)$ defined in (A.10). Then $\|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* \le 3\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* + 2\delta'/\lambda$ for all $\delta' \ge \delta$. That is, $\widehat{\mathbf{\Delta}} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta'/\lambda)$ for all $\delta' \ge \delta$.*

Proof for Lemma S.2.3.

$$
\begin{aligned}
0 &\le \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) - \widehat{Q}_\tau(\widehat{\mathbf{\Gamma}}_\tau) + \lambda(\|\mathbf{\Gamma}_\tau\|_* - \|\widehat{\mathbf{\Gamma}}_\tau\|_*) \quad (\widehat{\mathbf{\Gamma}}_\tau \text{ is the minimizer of } \widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*) \\
&\le \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) - \widehat{Q}_\tau(\widehat{\mathbf{\Gamma}}_{\tau,\delta}) + \lambda(\|\mathbf{\Gamma}_\tau\|_* - \|\widehat{\mathbf{\Gamma}}_{\tau,\delta}\|_*) + \delta \quad (\text{by (2.1)}) \\
&\le \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)\|\|\widehat{\mathbf{\Delta}}\|_* + \lambda(\|\mathbf{\Gamma}_\tau\|_* - \|\widehat{\mathbf{\Gamma}}_{\tau,\delta}\|_*) + \delta \\
&\le \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)\|\big(\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* + \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{\Gamma}_\tau)\|_* - \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Gamma}}_{\tau,\delta})\|_* - \|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Gamma}}_{\tau,\delta})\|_*) + \delta \\
&\le \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)\|\big(\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* + \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* - \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_*) + \delta, \qquad \text{(S.2.24)}
\end{aligned}
$$

where the second inequality follows from the definition of subgradient:

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

and Hölder's inequality; the third inequality is from the fact that $\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\mathbf{\Gamma}_\tau) = 0$ and for any $\mathbf{S}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{S})\|_* + \|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\mathbf{S})\|_*$ (the discussion after Definition S.2.2) ; the fourth inequality is from the triangle inequality.

Rearrange expression (S.2.24) to get,

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

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

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

Hence, $\|\mathcal{P}^{\perp}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* + 2\delta/\lambda \leq 3\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\widehat{\mathbf{\Delta}})\|_* + 2\delta'/\lambda$ for all $\delta' \geq \delta$. □

Lemma S.2.4. *Under assumptions (A2), (A3), we have for all $\delta > 0$,*

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

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

Proof for Lemma S.2.4. 1. Let $Q_{\tau,j}(\mathbf{\Gamma}_{\tau,*j}) = \mathsf{E}[\rho_\tau(Y_{ij}-\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j})]$. From Knight's identity (Knight; 1998), for any $v, u \in \mathbb{R}$,

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v \big(\mathbf{1}\{u \leq z\} - \mathbf{1}\{u \leq 0\}\big)dz. \tag{S.2.25}$$

where $\psi_\tau(u) \stackrel{\text{def}}{=} \tau - \mathbf{1}(u \leq 0)$. Putting $u = Y_{ij} - \boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}$ in (S.2.25), 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 $\mathbf{\Gamma}_\tau = \arg\min_{\mathbf{S}} \mathsf{E}[\widehat{Q}_\tau(\mathbf{S})]$. Therefore, using the law of iterative expectation and mean value theorem, we have by (A3) that

$$\begin{aligned}
&Q_{\tau,j}(\mathbf{\Gamma}_{\tau,*j} + \mathbf{\Delta}_{*j}) - Q_{\tau,j}(\mathbf{\Gamma}_{\tau,*j}) \\
&\quad= \mathsf{E}\bigg[\int_0^{\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}} F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j} + z|\boldsymbol{X}_i) - F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}|\boldsymbol{X}_i)dz\bigg] \\
&\quad= \mathsf{E}\bigg[\int_0^{\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}} z f_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}|\boldsymbol{X}_i) + \frac{z^2}{2}f'_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j} + z^\dagger|\boldsymbol{X}_i)dz\bigg] \\
&\quad\geq \underline{f}^\tau\frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2\big]}{4} + \underline{f}^\tau\frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j})^2\big]}{4} - \frac{1}{6}\bar{f}'\mathsf{E}[|\boldsymbol{X}_i^\top\mathbf{\Delta}_{*j}|^3]
\end{aligned} \tag{S.2.26}$$

for $z^\dagger \in [0, z]$. Now, for $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}_\tau; 2\delta/\lambda)$, the condition

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

implies

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

Therefore,

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

2. By the decomposability of the nuclear norm, $\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma}_\tau;2\delta/\lambda)$ and (S.2.2) in Remark S.2.1, we can estimate

$$\begin{aligned}\|\mathbf{\Delta}\|_*&=\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{\Delta})\|_*+\|\mathcal{P}^\perp_{\mathbf{\Gamma}_\tau}(\mathbf{\Delta})\|_*\leq 4\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{\Delta})\|_*+2\delta/\lambda\leq 4\sqrt{r}\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{\Delta})\|_{\mathrm{F}}+2\delta/\lambda\\&\leq 4\sqrt{\frac{rm}{\sigma_{\min}(\mathbf{\Sigma}_X)}}\|\mathbf{\Delta}\|_{L_2(P_X)}+2\delta/\lambda.\end{aligned}$$

□

Lemma S.2.5. *Under Assumptions (A1)-(A3), recall that $\mathcal{A}(t)$ is defined in* (S.2.11)*, then for an arbitrary $u>1$,*

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

where $\alpha_r=4\sqrt{r/\sigma_{\min}(\mathbf{\Sigma}_X)}$ and $r=\mathrm{rank}(\mathbf{\Gamma}_\tau)$.

Proof of Lemma S.2.5. To simplify notations, let

$$\alpha_r\overset{\text{def}}{=}4\sqrt{r/\sigma_{\min}(\mathbf{\Sigma}_X)}\tag{S.2.27}$$

Let $\{\varepsilon_{ij}\}_{i\leq n,j\leq m}$ 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 $\{\varepsilon_{ij}\}_{i\leq n,j\leq m}$, given Y_{ij} and $\boldsymbol{X}_i$. Denote

$$\chi_{ij}^\tau(\cdot)\overset{\text{def}}{=}\rho_\tau\{Y_{ij}-\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}-\cdot\}-\rho_\tau\{Y_{ij}-\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\tau,*j}\}.\tag{S.2.28}$$

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

$$\big|\chi_{ij}^\tau(a)-\chi_{ij}^\tau(b)\big|\leq|a-b|.\quad\forall i=1,...,n,\ j=1,...,m.\tag{S.2.29}$$

First, we note that for any $\boldsymbol{\Delta}$ satisfying $\boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}_\tau; 2\delta/\lambda)$ and $\|\boldsymbol{\Delta}\|_{L_2(P_X)} \le t$,

$$\begin{aligned}
&\mathrm{Var}\left(\mathbb{G}_n\left(m^{-1}\sum_{j=1}^{m}\chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right)\right)\\
&= \mathrm{Var}\left(m^{-1}\sum_{j=1}^{m}\chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right) \le m^{-1}\sum_{j=1}^{m}\mathsf{E}\big[(\chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}))^2\big]\\
&\le m^{-1}\sum_{j=1}^{m}\mathsf{E}\big[(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})^2\big] \le t^2, \qquad \text{(S.2.30)}
\end{aligned}$$

where the first equality and the second inequality follow from elementary computations and i.i.d. assumption (A1), the third inequality is a result of (S.2.29), and the last inequality applies (S.2.1) in Remark S.2.1.

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

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

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

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

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

$$\mathrm{P}\{\mathcal{A}(t) > s\} \le 4\mathrm{P}\left(\sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le t\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma}_\tau;2\delta/\lambda)}}\left|n^{-1/2}m^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\chi_{ij}^{\tau}(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right| > \frac{s}{4}\right). \qquad \text{(S.2.31)}$$

Now we restrict the $\mathcal{A}(t)$ on the event Ω on which (3.1) 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.2.31)

can be bounded by

$$\begin{aligned}&\mathrm{P}\{\mathcal{A}(t) > s|\Omega\}\\&\leq 4\exp\left(\frac{-\mu s}{4}\right)\mathsf{E}\left[\mathsf{E}_\varepsilon\left[\exp\left\{\mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\leq t\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma}_\tau;2\delta/\lambda)}}\left|n^{-1/2}m^{-1}\sum_{i=1}^n\sum_{j=1}^m \varepsilon_{ij}\chi_{ij}^\tau(\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j})\right|\right\}\right]\Big|\Omega\right].\end{aligned} \tag{S.2.32}$$

Now recall (S.2.29), the comparison theorem for Rademacher processes (Lemma 4.12 in Ledoux and Talagrand (1991)) implies the right-hand side of (S.2.32) is bounded by

$$\begin{aligned}&\mathrm{P}\{\mathcal{A}(t) > s|\Omega\}\\&\leq 4\exp\left(\frac{-\mu s}{4}\right)\mathsf{E}\left[\mathsf{E}_\varepsilon\left[\exp\left\{2\mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\leq t\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma}_\tau;2\delta/\lambda)}}\left|n^{-1/2}m^{-1}\sum_{i=1}^n\sum_{j=1}^m \varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right|\right\}\right]\Big|\Omega\right].\end{aligned} \tag{S.2.33}$$

To obtain a bound for the right-hand side of (S.2.33), we note that

$$\begin{aligned}\left|\sum_{i=1}^n\sum_{j=1}^m \varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\right| &= \left|\mathrm{tr}\Big(\Big[\sum_{i=1}^n\varepsilon_{i1}\boldsymbol{X}_i\ \sum_{i=1}^n\varepsilon_{i2}\boldsymbol{X}_i\ \ldots\ \sum_{i=1}^n\varepsilon_{im}\boldsymbol{X}_i\Big]^\top\boldsymbol{\Delta}\Big)\right|\\&\leq \|\boldsymbol{\Delta}\|_* \sup_{\boldsymbol{a}\in\mathcal{S}^{p-1}}\left|\sum_{j=1}^m\Big(\sum_{i=1}^n\varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{a}\Big)^2\right|^{1/2}\\&\leq m^{1/2}\|\boldsymbol{\Delta}\|_*\max_{j\leq m}\left\|\sum_{i=1}^n\varepsilon_{ij}\boldsymbol{X}_i\right\|,\end{aligned} \tag{S.2.34}$$

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.2.33). Using matrix dilations (see, for example Section 2.6 of Tropp (2011)), we have

$$\left\|\sum_{i=1}^n\varepsilon_{ij}\boldsymbol{X}_i\right\| = \left\|\sum_{i=1}^n\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\right\|. \tag{S.2.35}$$

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 symmetrically distributed

conditional on $\boldsymbol{X}_i$. We now obtain

$$
\begin{aligned}
&\mathsf{E}_\varepsilon\bigg[\exp\bigg\{2\mu \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(P_X)}\le t\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma}_\tau;2\delta/\lambda)}}\bigg|n^{-1/2}m^{-1}\sum_{i=1}^n\sum_{j=1}^m \varepsilon_{ij}\boldsymbol{X}_i^\top\boldsymbol{\Delta}_{*j}\bigg|\bigg\}\bigg]\\
&\le \mathsf{E}_\varepsilon\bigg[\exp\bigg\{2\mu(\alpha_r t+m^{-1/2}2\delta/\lambda)\max_{j\le m}\bigg\|n^{-1/2}\sum_{i=1}^n\varepsilon_{ij}\boldsymbol{X}_i^\top\bigg\|\bigg\}\bigg]\\
&\le m\max_{j\le m}\mathsf{E}_\varepsilon\bigg[\exp\bigg\{2\mu(\alpha_r t+m^{-1/2}2\delta/\lambda)\bigg\|n^{-1/2}\sum_{i=1}^n\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\bigg\|\bigg\}\bigg]\\
&\le m2(p+1)\max_{j\le m}\exp\bigg\{\sigma_{\max}\bigg(\sum_{i=1}^n\log\mathsf{E}_\varepsilon\bigg[\exp\bigg\{2\mu(\alpha_r t+m^{-1/2}2\delta/\lambda)n^{-1/2}\varepsilon_{ij}\begin{pmatrix}\mathbf{0}_p & \boldsymbol{X}_i\\ \boldsymbol{X}_i^\top & 0\end{pmatrix}\bigg\}\bigg]\bigg)\bigg\}
\end{aligned}
\tag{S.2.36}
$$

where the first inequality is from Lemma S.2.4(ii) and (S.2.34) and recall α_r in (S.2.27), the second inequality follows from (S.2.35), Lemma S.2.4 (ii) ($\boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma}_\tau;2\delta/\lambda)$), 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 3(ii) of Maurer and Pontil (2013) by the symmetric distribution of ε_{ij}, where for a self adjoint matrix $\mathbf{A}$,

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

From equation (2.4) on page 399 of Tropp (2011), for any j and $c>0$,

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

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, (S.2.36) is bounded by

$$2m(p+1)\exp\left\{2\mu^2(\alpha_r t + m^{-1/2}2\delta/\lambda)^2\sigma_{\max}\left(n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{X}_i\boldsymbol{X}_i^\top & \mathbf{0}_p \\ 0 & \boldsymbol{X}_i^\top\boldsymbol{X}_i\end{pmatrix}\right)\right\}$$
$$\leq 2m(p+1)\exp\left\{2\mu^2(\alpha_r t + m^{-1/2}2\delta/\lambda)^2(\sigma_{\max}(\widehat{\boldsymbol{\Sigma}}_X) + B_p)\right\}, \tag{S.2.37}$$

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) (with Shatten-∞ norm), and Assumption (A2).

Putting (S.2.37) into (S.2.32), we obtain

$$\begin{aligned}\mathrm{P}\{\mathcal{A}(t) > s|\Omega\} &\leq 8m(p+1)\exp\left(\frac{-\mu s}{4}\right)\mathsf{E}\left[\exp\left\{2\mu^2(\alpha_r t + m^{-1/2}2\delta/\lambda)^2(\sigma_{\max}(\widehat{\boldsymbol{\Sigma}}_X) + B_p)\right\}\middle|\Omega\right] \\ &\leq 8m(p+1)\exp\left(\frac{-\mu s}{4}\right)\exp\left\{2\mu^2(\alpha_r t + m^{-1/2}2\delta/\lambda)^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p)\right\}.\end{aligned} \tag{S.2.38}$$

Minimizing the expression (S.2.38) with respect to μ gives

$$\mathrm{P}\{\mathcal{A}(t) > s|\Omega\} \leq 8m(p+1)\exp\left\{-\frac{s^2}{128(\alpha_r t + m^{-1/2}2\delta/\lambda)^2(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p)}\right\}. \tag{S.2.39}$$

Taking

$$s = 8\sqrt{2}u(\alpha_r t + m^{-1/2}2\delta/\lambda)\sqrt{(c_2\sigma_{\max}(\boldsymbol{\Sigma}_X) + B_p)}\sqrt{\log m + \log p}. \tag{S.2.40}$$

Notice that $s \geq \sqrt{8}t$ for large enough p, m, so the symmetrization (S.2.31) is valid. Recall that $\mathrm{P}(\Omega) \geq 1 - \gamma_n$. The proof is then completed. □

Remark S.2.6. *The 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.2.5 require only independence in the random variables $(Y_{ij}, \boldsymbol{X}_i)$, without needing identical distribution. 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.2.7. *It can be observed that Lemma S.2.5 is valid uniformly for any $0 < \tau < 1$.*

S.2.4 Proof of Theorem 3.7

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

To prove (3.13), since $\Psi_\tau = \mathbf{V}_\tau$ and $\widehat{\Psi}_\tau = \widetilde{\mathbf{V}}_\tau$, by Theorem 3 of Yu et al. (2015),

$$\sin\cos^{-1}(|\widetilde{\mathbf{V}}_{*j}^\top \mathbf{V}_{*j}|) \le \frac{2(2\|\mathbf{\Gamma}_\tau\| + \|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}})\|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}}{\min\{\sigma_{j-1}^2(\mathbf{\Gamma}_\tau) - \sigma_j^2(\mathbf{\Gamma}_\tau), \sigma_j^2(\mathbf{\Gamma}_\tau) - \sigma_{j+1}^2(\mathbf{\Gamma}_\tau)\}} \tag{S.2.41}$$

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

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

A similar bound like (3.13) also holds for $\widetilde{\mathbf{U}}_{*j}$, by the discussion below Theorem 3 of Yu et al. (2015).

For a proof for inequality (3.14), by direct calculation,

$$\begin{aligned}\big|\widehat{f}_k^\tau(\boldsymbol{X}_i) - f_k^\tau(\boldsymbol{X}_i)\big| &= \big|\widetilde{\sigma}_k \widetilde{\mathbf{U}}_{*k}^\top \boldsymbol{X}_i - \sigma_k \mathbf{U}_{*k}^\top \boldsymbol{X}_i\big| \\ &\le \big\|\widetilde{\sigma}_k \widetilde{\mathbf{U}}_{*k}^\top - \sigma_k \mathbf{U}_{*k}^\top\big\| \|\boldsymbol{X}_i\| \\ &\le \big(\big|\widetilde{\sigma}_k - \sigma_k\big| \big\|\widetilde{\mathbf{U}}_{*k}\big\| + \sigma_k \big\|\widetilde{\mathbf{U}}_{*k} - \mathbf{U}_{*k}\big\|\big)\|\boldsymbol{X}_i\| \\ &\le \Big(\big|\widetilde{\sigma}_k - \sigma_k\big| + \sigma_k \sqrt{(\widetilde{\mathbf{U}}_{*k} - \mathbf{U}_{*k})^\top (\widetilde{\mathbf{U}}_{*k} - \mathbf{U}_{*k})}\Big)\|\boldsymbol{X}_i\| \\ &\le \Big(\big|\widetilde{\sigma}_k - \sigma_k\big| + \sigma_k \sqrt{2(1 - \widetilde{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k})}\Big)\|\boldsymbol{X}_i\|\end{aligned} \tag{S.2.42}$$

where we apply the fact that $\big\|\widetilde{\mathbf{U}}_{*k}\big\| = 1$. By assumption $\widetilde{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k} \ge 0$, $\widetilde{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k} = |\widetilde{\mathbf{U}}_{*k}^\top \mathbf{U}_{*k}|$. Apply Lemma 3.6 and the bound (S.2.41) with $\mathbf{V}$ being replaced by $\mathbf{U}$ to (S.2.42), then (3.14) is proved. Thus, the proof for this theorem is completed. □

S.3: Miscellaneous Technical Details

S.3.1 Detail on Remark 3.3

For (3.7) to hold, it is enough to have $\mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^3] \le C\mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^2]^{3/2}$ for all $j = 1, 2, ..., m$, where $C > 0$ is a constant independent of j, because

$$\Big(\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^2]^{3/2}\Big)^{2/3} \le \sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^2] \tag{S.3.1}$$

by the inequality $\|\boldsymbol{a}\|_{3/2} \le \|\boldsymbol{a}\|_1$ for an arbitrary $\boldsymbol{a} = (a_1, a_2, ..., a_m)$ with $a_j \ge 0$, $\forall j$. If $\boldsymbol{X}_i$ is i.i.d. sampled from a *log-concave* density, then Theorem 5.22 of Lovász and Vempala (2007) implies $\mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^3] \le 3^{3/2}\mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^2]^{3/2}$ for any $\mathbf{\Delta}$. See also Design 1 on p.2 of the

supplemental materials of Belloni and Chernozhukov (2011). This implies (3.7) as $\epsilon_{n,\tau,r}$ is small as $n \gtrsim B_p r(p+m)(\log p + \log m)$.

S.3.2 Detail on Remark 3.5

We need some extra notations. 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 $\mathbf{P}_{\mathcal{V}} = \mathbf{V}\mathbf{V}^{\top}$, $\mathbf{P}^{\perp}_{\mathcal{V}} = \mathbf{I}_{m\times r} - \mathbf{P}_{\mathcal{V}}$, $\mathbf{V} = [\mathbf{v}_1 ... \mathbf{v}_r]$, and $\{\mathbf{v}_j\}_{j=1}^r$ is a set of orthonormal basis for $\mathcal{V}$; analogously, $\mathbf{P}_{\mathcal{U}} = \mathbf{U}\mathbf{U}^{\top}$, $\mathbf{P}^{\perp}_{\mathcal{U}} = \mathbf{I}_{p\times r} - \mathbf{P}_{\mathcal{U}}$, $\mathbf{U} = [\mathbf{u}_1 ... \mathbf{u}_r]$, and $\{\mathbf{u}_j\}_{j=1}^r$ is a set of orthonormal basis for $\mathcal{U}$. Moreover, for any $\mathbf{S} \in \mathbb{R}^{p\times m}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathcal{M}}(\mathbf{S})\|_* + \|\mathcal{P}^{\top}_{\overline{\mathcal{M}}}(\mathbf{S})\|_*$.

It can be shown that when $\lambda \geq 2\|\nabla \widehat{Q}(\mathbf{\Gamma}_\tau)\|$, the difference $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau$ lies in the set

$$\begin{aligned}
&\mathcal{K}(\overline{\mathcal{M}}, 4\|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Gamma}_\tau)\| + 2\delta'/\lambda) \\
&\quad \stackrel{\text{def}}{=} \left\{\mathbf{\Delta} \in \mathbb{R}^{p\times m} : \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\mathbf{\Delta})\| \leq 3\|\mathcal{P}_{\overline{\mathcal{M}}}(\mathbf{\Delta})\| + 4\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma}_\tau)\| + \frac{2\delta'}{\lambda}\right\}, \qquad \text{(S.3.2)}
\end{aligned}$$

where $\delta' \geq \delta$. Under this situation, the recovery property of $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ can be shown via similar argument as for Theorem 3.2 (possibly under more restrictive conditions), and we leave out the details.

To show (S.3.2), we first note an inequality

$$\|\widehat{\mathbf{\Gamma}}_{\tau,\delta}\|_* - \|\mathbf{\Gamma}_\tau\|_* \leq 2\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma}_\tau)\|_* + \|\mathcal{P}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* - \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_*, \qquad \text{(S.3.3)}$$

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 seen that from similar argument as (S.2.24),

$$\begin{aligned}
0 &\leq \widehat{Q}_\tau(\mathbf{\Gamma}_\tau) - \widehat{Q}_\tau(\mathbf{\Gamma}_{\tau,T}) + \lambda\|\mathbf{\Gamma}_\tau\|_* - \lambda\|\mathbf{\Gamma}_{\tau,T}\|_* + \delta \\
&\leq \|\nabla \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)\|\big(\|\mathcal{P}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* + \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_*\big) \\
&\qquad + \lambda(2\|\mathcal{P}^{\perp}_{\mathcal{M}}(\mathbf{\Gamma}_\tau)\|_* + \|\mathcal{P}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_* - \|\mathcal{P}^{\perp}_{\overline{\mathcal{M}}}(\widehat{\mathbf{\Delta}})\|_*) + \delta, \qquad \text{(S.3.4)}
\end{aligned}$$

where the second inequality is from (S.3.3). Rearrange expression (S.3.4) to get,

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

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

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

S.3.3 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 remaining singular value is 0. For $\mathbf{S}_2^{Sym}$, the first two nonzero singular values are $(\sigma_1(\mathbf{S}_2^{Sym}), \sigma_2(\mathbf{S}_2^{Sym})) = (175.48, 25.74)$ and the rest is 0. For $\mathbf{S}_2^{Sym}$, the first six nonzero singular values are $(\sigma_1(\mathbf{S}_2^{Asym}), ..., \sigma_6(\mathbf{S}_2^{Asym})) = (473.40, 29.87, 25.66, 23.89, 23.58, 22.16)$ and the rest is 0.

S.4: 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, $S_f : \mathcal{X} \to \mathcal{X}$:*

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

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*

$$S_\lambda(\mathbf{Y}) \overset{\text{def}}{=} \mathbf{U}(\mathbf{D} - \lambda\mathbf{I}_{pm})_+\mathbf{V}^\top, \tag{S.4.1}$$

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 (2012))**.** *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*

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

S.5: Selecting the Matrix B in Section 4.3

The $\mathbf{B}$ in (4.3) is the coefficient estimator obtained by fitting a VAR(1) model (Lütkepohl; 2005) to the $\boldsymbol{X}_i$ in (5.1) and $\Sigma_{\boldsymbol{\varepsilon}}$ is the sample covariance matrix from the residuals. Due to the high dimensionality (460), the VAR model may be over-parameterized especially when the order is high, and straightforwardly estimating the VAR may yield unreliable estimates. Therefore, as suggested by multiple authors (e.g. Davis et al. (2016); Nicholson et al. (2017) and the references therein), we estimate the VAR model with the ℓ_1 norm penalty, or Lasso (Tibshirani; 1996), to alleviate the problem of over-parameterization. Henceforth, the VAR model estimated with the Lasso penalty will be called Lasso-VAR. The computation can be carried out with the `R` package `BigVAR` (Nicholson et al.; 2017). The Lasso tuning parameter is selected optimally by the cross-validation procedure provided in the package.

To evaluate the adequacy of the VAR(1) model for the real data in (5.1) Table S.5.1 provides the 1-step-ahead mean square forecasting error (MSFE) of Lasso-VAR (see Eq. (12) of Nicholson et al. (2017)) with different lags. As it requires excessive computational time and resource for model estimation and cross-validation, the maximal order under consideration here is three. Lasso-VAR(3) has the smallest MSFE, but the difference between the models seems small, so we take Lasso-VAR(1). The MSFE of VAR with order selected by AIC or BIC (Lütkepohl; 2005; Nicholson et al.; 2017) is 2805 with optimal order of both being 0, which is higher than that of Lasso-VAR as shown in Table S.5.1.

For a simple diagnosis of Lasso-VAR(1), we check the autocorrelation and partial autocorrelation function of each individual residual series. Autocorrelation and partial autocorrelation functions of some series are significant. However, increasing the order of the VAR model does not improve the situation. To our knowledge, we are not aware of any literature on vector ARIMA models for high dimensional time series, which might provide a better fit of our data. Fitting a very high dimensional VAR like ours is very subtle. As the Lasso-VAR(1) has demonstrated competent forecasting performance as shown in Table S.5.1, we

adopt Lasso-VAR(1). A full exploration of the time series structure of the data is left for future research.

Order	1	2	3
Lasso-VAR MSFE	2364.82	2353.075	2341.046
% of active coef.	3.9	2.836	2.381
MSFE of VAR-AIC/BIC (optimal order = 0): 2805			

Table S.5.1: The mean square forecasting error (MSFE) and the percentage of active coefficients (total number of coefficients = $460 \times (1 + 460\times$ order)) with different orders, where "1" is from the intercept. For the matrix $\boldsymbol{B}$, we do not include the intercept.

S.6: Additional Numerical Results: AR(1) Model

In this section, we consider the same data generating model as (4.1) in Section 4.1, but now the regressor $\boldsymbol{X}_i$ follows an AR(1) model

$$\boldsymbol{X}_i = 0.5\boldsymbol{X}_{i-1} + \boldsymbol{u}_i, \tag{S.6.1}$$

where $\boldsymbol{u}_i$ follows the multivariate $U([0,1])$ distribution with covariance matrix $\boldsymbol{\Sigma}$ in which $\boldsymbol{\Sigma}_{ij} = 0.1 * 0.8^{|i-j|}$. Because $\boldsymbol{Y}_i$ is generated as (4.1), the true number of factors is 2 for $\tau = 0.2$ and 6 for $\tau = 0.8$ as in the i.i.d. case. The computational setting is the same as the i.i.d. case.

Figure S.6.1 shows the relative frequency of the estimated number of factors and the estimated penalized validation loss when the regressors follow (S.6.1). It appears that the presence of time dependency slightly decreases the recovery accuracy, but the pattern of the penalized validation loss and the estimation performance of the number of factors remain similar to the i.i.d. case in Section 4.2. However, for $\tau = 0.8$, smaller κ and greater T than than those for $\tau = 0.2$ are selected to ensure estimation accuracy, which is due to the fact that the true number of factors for $\tau = 0.8$ is greater than that of $\tau = 0.2$.

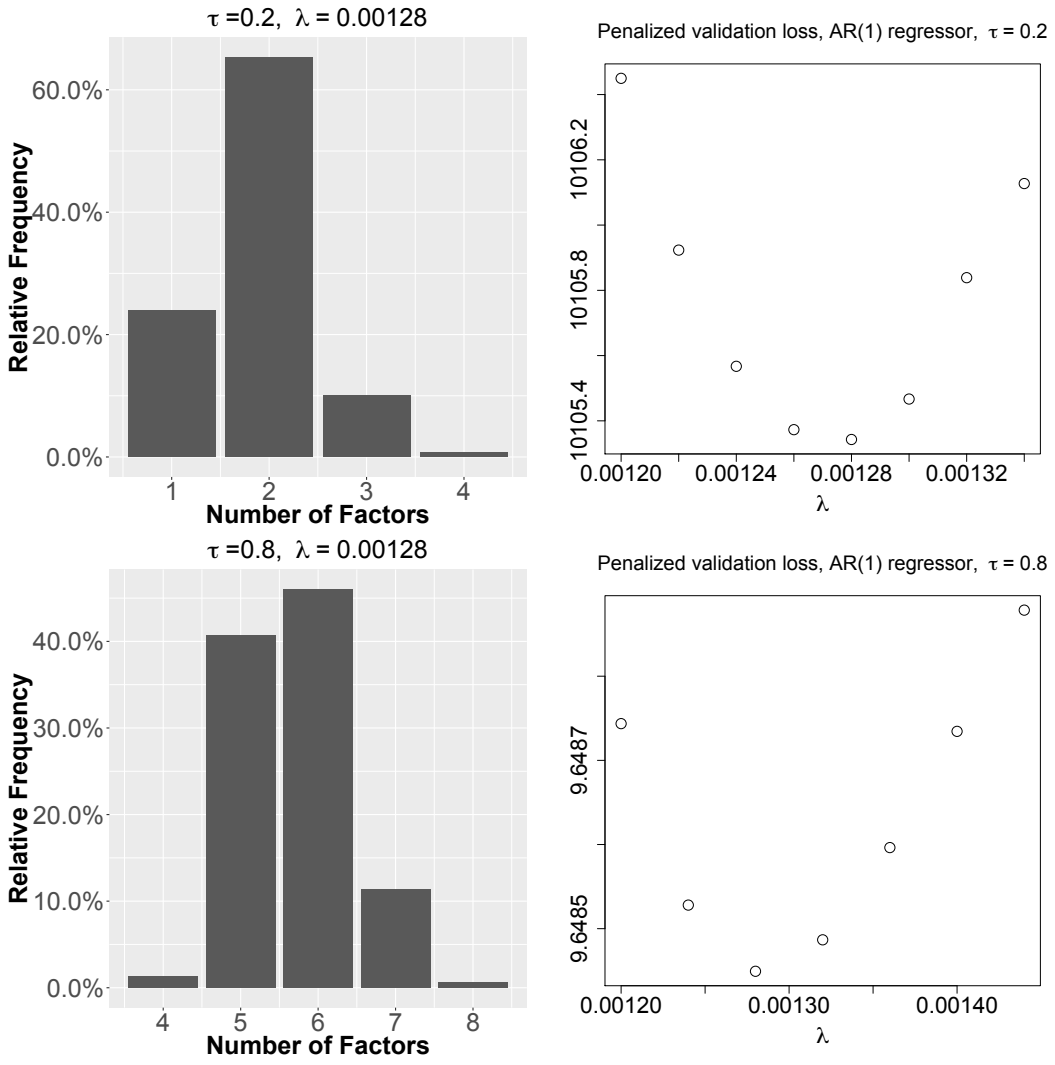

Figure S.6.1: The histogram of the estimated number of factors and the plot for the penalized validation loss computed by the average of 150 Monte Carlo repetitions, $\tau = 0.2$ and 0.8. Data are generated as (4.1), with AR(1) regressor $\boldsymbol{X}_i$ generated as in (S.6.1). The true number of factors is 2 for $\tau = 0.2$ and 6 for $\tau = 0.8$. $(\kappa, T) = (6.66 * 10^{-6}, 3500)$ for $\tau = 0.2$ and $(\kappa, T) = (8 * 10^{-7}, 4000)$ for $\tau = 0.8$.

References

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.

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

Buldygin, V. V. and Moskvichova, K. K. (2013). The sub-Gaussian norm of a binary random variable, *Theory of Probability and Mathematical Statistics* **86**: 33–49.

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.

Davis, R. A., Zang, P. and Zheng, T. (2016). Sparse vector autoregressive modeling, *Journal of Computational and Graphical Statistics* **25**(4): 1077–1096.
URL: *https://doi.org/10.1080/10618600.2015.1092978*

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.

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

Lovász, L. and Vempala, S. (2007). The geometry of logconcave functions and sampling algorithms, *Random Structures & Algorithms* **30**(3): 307–358.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20135*

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, 1st edn, Springer-Verlag Berlin Heidelberg.

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 noise 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.

Nicholson, W. B., Matteson, D. S. and Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables, *International Journal of Forecasting* **33**(3): 627 – 651.
URL: *http://www.sciencedirect.com/science/article/pii/S0169207017300080*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1): 267–288.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x*

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

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