Supplementary Material for "Convoluted Support Matrix Machine in High Dimensions"

Bingzhen Chen and Canyi Chen

Hangzhou Dianzi University and University of Michigan

This supplementary material contains all the proof for the main theoretical results in the main text.

A. Proof of Theorem 1

Proof of Theorem 1. By the definition of elastic-net convoluted support matrix machine (CSMM) in (2.2), we have

$$\begin{aligned} &\frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \widehat{\mathbf{A}}) + \widehat{a}\}] + \lambda\|\widehat{\mathbf{A}}\|_* + \lambda_0\|\widehat{\mathbf{A}}\|_F^2 \\ \le\ &\frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}] + \lambda\|\mathbf{A}^*\|_* + \lambda_0\|\mathbf{A}^*\|_F^2. \end{aligned}$$

Let $\widehat{\delta} \overset{\text{def}}{=} \widehat{a} - a^*$, $\widehat{\mathbf{\Lambda}} \overset{\text{def}}{=} \widehat{\mathbf{A}} - \mathbf{A}^*$, $\widehat{\mathbf{\Delta}}_{r^c} = \mathbf{\Pi}_{\mathcal{N}}\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Delta}}_r \overset{\text{def}}{=} \widehat{\mathbf{\Delta}} - \widehat{\mathbf{\Delta}}_{r^c}$. The above display, together with the fact that $\|\mathbf{A}^* + \widehat{\mathbf{\Delta}}\|_* \ge \|\mathbf{A}^* + \widehat{\mathbf{\Delta}}_{r^c}\|_* - \|\widehat{\mathbf{\Delta}}_r\|_* =$

$\|\mathbf{A}^*\|_* + \|\widehat{\mathbf{\Delta}}_{r^c}\|_* - \|\widehat{\mathbf{\Delta}}_r\|_*$ (Lemma B.1 in Appendix B), implies

$$
\begin{aligned}
& \frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \widehat{\mathbf{A}}) + \widehat{a}\}] \\
& -\frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}] + \lambda_0(\|\widehat{\mathbf{A}}\|_F^2 - \|\mathbf{A}^*\|_F^2) \\
\leq\ & \lambda(\|\mathbf{A}^*\|_* - \|\widehat{\mathbf{A}}\|_*) = \lambda(\|\mathbf{A}^*\|_* - \|\mathbf{A}^* + \widehat{\mathbf{\Delta}}\|_*) \\
\leq\ & \lambda(\|\widehat{\mathbf{\Delta}}_r\|_* - \|\widehat{\mathbf{\Delta}}_{r^c}\|_*).
\end{aligned}
$$

On the other hand, by the convexity of $\mathcal{L}_h(\cdot)$, we have,

$$
\begin{aligned}
& \frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \widehat{\mathbf{A}}) + \widehat{a}\}] \\
& -\frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}] + \lambda_0(\|\widehat{\mathbf{A}}\|_F^2 - \|\mathbf{A}^*\|_F^2) \\
\geq\ & \frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i(\widehat{a} - a^*) \\
& + \left\langle 2\lambda_0 \mathbf{A}^* + \frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i\mathbf{X}_i, \widehat{\mathbf{A}} - \mathbf{A}^* \right\rangle \\
\geq\ & -\left|\frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i\right| |\widehat{\delta}| \\
& -\left\| 2\lambda_0 \mathbf{A}^* + \frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i\mathbf{X}_i \right\| \\
& \big(\|\widehat{\mathbf{\Delta}}_{r^c}\|_* + \|\widehat{\mathbf{\Delta}}_r\|_*\big).
\end{aligned}
$$

Define events

$$\mathcal{E}_1 \stackrel{\text{def}}{=} \left\{ \left| \frac{1}{n} \sum_i \mathcal{L}_h'[Y_i\{\text{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i \right| \le \lambda/2 \right\}$$

and

$$\mathcal{E}_2 \stackrel{\text{def}}{=} \left\{ \left\| 2\lambda_0 \mathbf{A}^* + \frac{1}{n} \sum_i \mathcal{L}_h'[Y_i\{\text{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i \mathbf{X}_i \right\| \le \lambda/2 \right\}.$$

By Lemmas B.3 and B.4, we have

$$\text{pr}(\mathcal{E}_1 \cap \mathcal{E}_2) \quad \ge \quad 1 - 2\exp(-n\lambda^2/8) - 2 \cdot 9^{p+q} \exp\left\{ -1/\eta_0 \min\left(\frac{\lambda^2}{64 m_0^2}, \frac{\lambda}{8 m_0} \right) n \right\},$$

for some absolute constant η_0. Under $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$-\frac{\lambda}{2}(|\widehat{\delta}| + \|\widehat{\mathbf{\Delta}}_{r^c}\|_* + \|\widehat{\mathbf{\Delta}}_r\|_*) \le \lambda(\|\widehat{\mathbf{\Delta}}_r\|_* - \|\widehat{\mathbf{\Delta}}_{r^c}\|_*),$$

which implies

$$\|\widehat{\mathbf{\Delta}}_{r^c}\|_* \le 3\|\widehat{\mathbf{\Delta}}_r\|_* + |\widehat{\delta}|, \tag{A.1}$$

or $(\widehat{\delta}, \widehat{\mathbf{\Delta}}) \in \mathcal{A}$.

Define $F(a, \mathbf{A}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_i \mathcal{L}_h[Y_i\{\text{tr}(\mathbf{X}_i^\top \mathbf{A}) + a\}]$ for any $(a, \mathbf{A}) \in \mathbb{R} \times \mathbb{R}^{p\times q}$ and

$$\mathbb{C}(t) \stackrel{\text{def}}{=} \{(\delta, \mathbf{\Delta}) \in \mathcal{A} \colon |\delta|^2 + \|\mathbf{\Delta}\|_F^2 = t^2(p+q)r/n\},$$

for any $t > 0$. Let $G(a, \mathbf{A}) = F(a, \mathbf{A}) - F(a^*, \mathbf{A}^*)$ and

$$H(t) = \sup_{(a,\mathbf{A})\in(a^*,\mathbf{A}^*)+\mathbb{C}(t)} |G(a, \mathbf{A}) - E\{G(a, \mathbf{A})\}|.$$

We claim that $H(t) = O_p\{t(p+q)r/n\}$. Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. Rademacher random variable. For some constant $c_1 > 0$, we have

$$\begin{aligned}
& E\{H(t)\} \\
\leq\ & 2E\left\{ \sup_{(a,\mathbf{A})\in(a^*,\mathbf{A}^*)+\mathbb{C}(t)} \left| \frac{1}{n}\sum_i \sigma_i \Big(\mathcal{L}_h[Y_i\{\text{tr}(\mathbf{X}_i^\top \mathbf{A}) + a\}] \right.\right. \\
& \left.\left. -\mathcal{L}_h[Y_i\{\text{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}] \Big) \right| \right\} \\
\leq\ & 4E\left\{ \sup_{(a,\mathbf{A})\in(a^*,\mathbf{A}^*)+\mathbb{C}(t)} \left| \frac{1}{n}\sum_i \sigma_i Y_i \Big(\langle \mathbf{X}_i, \mathbf{A} - \mathbf{A}^* \rangle + a - a^* \Big) \right| \right\} \\
\leq\ & \frac{4}{n} E\left(\left\| \sum_i \sigma_i Y_i \mathbf{X}_i \right\| + \left| \sum_i \sigma_i Y_i \right| \right) \cdot 8r^{1/2} t\{(p+q)r/n\}^{1/2} \\
\leq\ & \frac{4}{n} \cdot 4c_1\{n(p+q)\log 9\}^{1/2} \cdot 8r^{1/2} t\{(p+q)r/n\}^{1/2} \\
\leq\ & \frac{300c_1 t(p+q)r}{n},
\end{aligned}$$

where the first inequality is by the standard symmetrization technique (see, e.g., Lemma 2.3.1 in van der Vaart and Wellner (2000)), the second one is due to the contraction inequality (see, Theorem 4.12 in Ledoux and Talagrand (2011)) the third inequality is owing to Cauchy–Schwarz inequality and the fact that $|\delta| + \|\mathbf{\Delta}\|_* = |\delta| + \|\mathbf{\Delta}_{r^c}\|_* + \|\mathbf{\Delta}_r\|_* \leq 4\|\mathbf{\Delta}_r\|_* + 2|\delta| \leq 4r^{1/2}\|\mathbf{\Delta}\|_F + 2|\delta| \leq 4r^{1/2}t\{(p+q)r/n\}^{1/2} + 2t\{(p+q)r/n\}^{1/2} \leq 8r^{1/2}t\{(p+q)r/n\}^{1/2}$, and the penultimate is by Lemma B.5. This implies that $H(t) = O_p\{t(p+q)r/n\}$. For any $T > 0$, define event

$$\mathcal{G}_T \stackrel{\text{def}}{=} \left\{ H(t) \leq \frac{Tt(p+q)r}{n} \right\},$$

and then we have $\lim_{T\to\infty} \limsup_{n\to\infty} \text{pr}(\mathcal{G}_T^c) = 0$.

We next derive a loweer bound for $E\{G(a, \mathbf{A})\}$ for any $(a, \mathbf{A}) \in (a^*, \mathbf{A}^*) + \mathbb{C}(t)$. For sufficiently large n and any $(a, \mathbf{A}) \in (a^*, \mathbf{A}^*) + \mathbb{C}(t)$, by Taylor's expansion and Assumption (A2), there exists $a \in [0, 1]$ such that

$$\begin{aligned}
& E\{G(a, \mathbf{A})\} \\
= \; & E\{\mathcal{L}_h[Y\{\text{tr}(\mathbf{X}^\top \mathbf{A}) + a\}] - \mathcal{L}_h[Y\{\text{tr}(\mathbf{X}^\top \mathbf{A}^*) + a^*\}]\} \\
\geq \; & 1/2\kappa\{|a - a^*|^2 + \|\mathbf{A} - \mathbf{A}^*\|_F^2\} \\
\geq \; & 1/2\kappa t^2 \frac{(p+q)r}{n}. \qquad \text{(A.2)}
\end{aligned}$$

On the other hand, with our choice for tuning parameters, for any $(a, \mathbf{A}) \in (a^*, \mathbf{A}^*) + \mathbb{C}(t)$ with $\mathbf{\Delta} = \mathbf{A} - \mathbf{A}^*$ and $\delta = a - a^*$, we have

$$\begin{aligned}
& \lambda(\|\mathbf{A}^*\|_* - \|\mathbf{A}\|_*) \\
\leq\ & \lambda\|\mathbf{\Delta}\|_* = \lambda\|\mathbf{\Delta}_{r^c}\|_* + \lambda\|\mathbf{\Delta}_r\|_* \\
\leq\ & 4\lambda\|\mathbf{\Delta}_r\|_* + \lambda|\delta| \\
\leq\ & 4\lambda r^{1/2}\|\mathbf{\Delta}\|_F + \lambda|\delta| \\
\leq\ & 4\lambda r^{1/2} t\{(p+q)r/n\}^{1/2} + \lambda t\{(p+q)r/n\}^{1/2} \\
\leq\ & 5c_0 rt(p+q)/n, \qquad \text{(A.3)}
\end{aligned}$$

and we also have, by the convexity of Frobenius norm,

$$\begin{aligned}
& \lambda_0(\|\mathbf{A}\|_F^2 - \|\mathbf{A}^*\|_F^2) \\
\geq\ & 2\lambda_0\langle \mathbf{A}^*, \mathbf{A} - \mathbf{A}^*\rangle \\
\geq\ & -2\lambda_0\|\mathbf{A}^*\|\|\mathbf{A} - \mathbf{A}^*\|_* \\
\geq\ & -\frac{\lambda}{4}\|\mathbf{A} - \mathbf{A}^*\|_* \\
\geq\ & -\frac{\lambda}{4}(4\|\mathbf{\Delta}_r\|_* + |\delta|) \\
\geq\ & -\lambda r^{1/2} t\{r(p+q)/n\}^{1/2} - \frac{\lambda}{4} t\{r(p+q)/n\}^{1/2} \\
\geq\ & -2c_0 rt(p+q)/n, \qquad \text{(A.4)}
\end{aligned}$$

where the second inequality is by von Neumann's trace inequality. Thus, combining (A.2), (A.3) and (A.4), under $\mathcal{G}_T$, we have for any $(a, \mathbf{A}) \in (a^*, \mathbf{A}^*) + \mathbb{C}(t)$,

$$\begin{aligned} & F(a, \mathbf{A}) + \lambda_0 \|\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_* - \{F(a^*, \mathbf{A}^*) + \lambda_0 \|\mathbf{A}^*\|_F^2 + \lambda \|\mathbf{A}^*\|_*\} \\ \geq \; & G(a, \mathbf{A}) - 7c_0 rt(p+q)/n \\ \geq \; & E\{G(a, \mathbf{A})\} - H(t) - 7c_0 rt(p+q)/n \\ \geq \; & E\{G(a, \mathbf{A})\} - Tt(p+q)r/n - 7c_0 rt(p+q)/n \\ \geq \; & (1/2\kappa t - T - 7c_0) t(p+q)r/n. \end{aligned}$$

Now, taking $t = (4T + 28c_0)/\kappa$, we have under $\mathcal{G}_T$,

$$\begin{aligned} \inf_{(a,\mathbf{A}) \in (a^*, \mathbf{A}^*) + \mathbb{C}(t)} F(a, \mathbf{A}) + \lambda_0 \|\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_* > F(a^*, \mathbf{A}^*) \\ + \lambda_0 \|\mathbf{A}^*\|_F^2 + \lambda \|\mathbf{A}^*\|_*. \end{aligned} \tag{A.5}$$

Recall that under $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$(\widehat{a}, \widehat{\mathbf{A}}) \in (a^*, \mathbf{A}^*) + \mathcal{A}.$$

We next claim that under $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T$,

$$|\widehat{a} - a^*|^2 + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_F^2 \le t^2(p+q)r/n.$$

In fact, if

$$|\widehat{a} - a^*|^2 + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_F^2 > t^2(p+q)r/n,$$

let

$$\omega_0^2 \stackrel{\text{def}}{=} \frac{t^2(p+q)r/n}{|\widehat{a} - a^*|^2 + \|\widehat{\mathbf{A}} - \mathbf{A}^*\|_F^2},$$

then $\omega_0 \in (0,1)$. In addition, define

$$(a', \mathbf{A}') = \omega_0(\widehat{a}, \widehat{\mathbf{A}}) + (1-\omega_0)(a^*, \mathbf{A}^*),$$

and then we have

$$|a' - a^*|^2 + \|\mathbf{A}' - \mathbf{A}^*\|_F^2 = t^2(p+q)r/n.$$

Moreover, because $(\widehat{a}, \widehat{\mathbf{A}}) - (a^*, \mathbf{A}^*) \in \mathcal{A}$ under $\mathcal{E}_1 \cap \mathcal{E}_2$ and $\mathcal{A}$ is a cone, we

get

$$(a', \mathbf{A}') - (a^*, \mathbf{A}^*) = \omega_0\{(\widehat{a}, \widehat{\mathbf{A}}) - (a^*, \mathbf{A}^*)\} \in \mathcal{A}.$$

This means that under $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$(a', \mathbf{A}') \in (a^*, \mathbf{A}^*) + \mathbb{C}(t).$$

By the convexity of $F(\cdot)$ and norm, and (A.5), we have

$$\begin{aligned}
& \omega_0\{F(\widehat{a}, \widehat{\mathbf{A}}) + \lambda_0\|\widehat{\mathbf{A}}\|_F^2 + \lambda\|\widehat{\mathbf{A}}\|_*\} \\
& +(1-\omega_0)\{F(a^*, \mathbf{A}^*) + \lambda_0\|\mathbf{A}^*\|_F^2 + \lambda\|\mathbf{A}^*\|_*\} \\
\geq\ & F(a', \mathbf{A}') + \lambda_0\|\mathbf{A}'\|_F^2 + \lambda\|\mathbf{A}'\|_* \\
\geq\ & \inf_{(a,\mathbf{A})\in(a^*,\mathbf{A}^*)+\mathbb{C}(t)} F(a, \mathbf{A}) + \lambda_0\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_* \\
>\ & F(a^*, \mathbf{A}^*) + \lambda_0\|\mathbf{A}^*\|_F^2 + \lambda\|\mathbf{A}^*\|_*,
\end{aligned}$$

under $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T$. The above display means

$$F(\widehat{a}, \widehat{\mathbf{A}}) + \lambda_0\|\widehat{\mathbf{A}}\|_F^2 + \lambda\|\widehat{\mathbf{A}}\|_* > F(a^*, \mathbf{A}^*) + \lambda_0\|\mathbf{A}^*\|_F^2 + \lambda\|\mathbf{A}^*\|_*,$$

which is a contradiction with the definition of $(\widehat{a}, \widehat{\mathbf{A}})$. So we proved the claim.

By union bound, previous results, and the choice of tuning parameters, we have

$$
\begin{aligned}
& \mathrm{pr}[\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T\}^c] \\
\leq\ & \mathrm{pr}(\mathcal{E}_1^c) + \mathrm{pr}(\mathcal{E}_2^c) + \mathrm{pr}(\mathcal{G}_T^c) \\
\leq\ & 2\exp(-n\lambda^2/8) + 2\cdot 9^{p+q}\exp\left\{-\frac{1}{\eta_0}\min\left(\frac{\lambda^2}{64m_0^2}, \frac{\lambda}{8m_0}\right)n\right\} + \mathrm{pr}(\mathcal{G}_T^c) \\
\leq\ & 2\exp\left\{-c_0^2\frac{(p+q)}{8}\right\} + 2\cdot 9^{p+q}\exp\left\{-\frac{1}{\eta_0}\frac{\lambda^2 n}{64m_0^2}\right\} \\
& +2\cdot 9^{p+q}\exp\left\{-\frac{1}{\eta_0}\frac{\lambda n}{8m_0}\right\} + \mathrm{pr}(\mathcal{G}_T^c)] \\
=\ & 2\exp\left\{-c_0^2\frac{(p+q)}{8}\right\} + 2\cdot 9^{p+q}\exp\left\{-\frac{1}{\eta_0}\frac{c_0^2(p+q)}{64m_0^2}\right\} \\
& +2\cdot 9^{p+q}\exp\left\{-\frac{1}{\eta_0}\frac{c_0\{n(p+q)\}^{1/2}}{8m_0}\right\} + \mathrm{pr}(\mathcal{G}_T^c)]
\end{aligned}
$$

Since $(p+q)/n = o(1)$, as long as c_0 is sufficiently large, we have

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathrm{pr}[\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{G}_T\}^c] = 0.$$

Combining this result and the previous claim, we have proved Theorem 1. □

B. Some Useful Lemmas

Lemma B.1. For $\mathbf{A}_1 \in \mathcal{M}$ and $\mathbf{A}_2 \in \mathcal{N}$, we have $\|\mathbf{A}_1+\mathbf{A}_2\|_* = \|\mathbf{A}_1\|_*+\|\mathbf{A}_2\|_*$.

In addition, for any $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\mathrm{rank}(\mathbf{A} - \mathbf{\Pi}_{\mathcal{N}}\mathbf{A}) \le 2r$.

Proof of Lemma B.1. For any $\mathbf{A}_1 \in \mathcal{M}$ and $\mathbf{A}_2 \in \mathcal{N}$, by the definition of $\mathcal{M}$ and $\mathcal{N}$, we have $\mathbf{A}_1 = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ and $\mathbf{A}_2 = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^\top$ with $\mathbf{U}^\top\widetilde{\mathbf{U}} = \mathbf{0}$ and $\mathbf{V}^\top\widetilde{\mathbf{V}} = \mathbf{0}$. Simple calculation leads to $\mathbf{A}_1 + \mathbf{A}_2 = (\mathbf{U}, \widetilde{\mathbf{U}})\mathrm{diag}(\mathbf{\Lambda}, \widetilde{\mathbf{\Lambda}})(\mathbf{V}, \widetilde{\mathbf{V}})^\top$, which implies $\|\mathbf{A}_1 + \mathbf{A}_2\|_* = \|\mathbf{A}_1\|_* + \|\mathbf{A}_2\|_*$.

Suppose that

$$\mathbf{A} = (\mathbf{U}, \widetilde{\mathbf{U}}) \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix} (\mathbf{V}, \widetilde{\mathbf{V}})^\top.$$

It is straightforward to check that $\mathbf{\Pi}_{\mathcal{N}}\mathbf{A} = \widetilde{\mathbf{U}}\mathbf{H}\widetilde{\mathbf{V}}^\top$. It follows that the rank of

$$\mathbf{A} - \mathbf{\Pi}_{\mathcal{N}}\mathbf{A} = (\mathbf{U}, \widetilde{\mathbf{U}}) \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{0} \end{pmatrix} (\mathbf{V}, \widetilde{\mathbf{V}})^\top$$

is at most $2r$. □

Lemma B.2. For any matrix $\mathbf{E} \in \mathbb{R}^{p\times q}$,

$$\|\mathbf{E}\| \le 2 \max_{\mathbf{u}_j \in \mathcal{U}, \mathbf{v}_k \in \mathcal{V}} \mathbf{u}_j^\top \mathbf{E} \mathbf{v}_k,$$

where $\mathcal{U} \subseteq \mathbb{R}^p$ is a finite set containing at most 9^p unit vectors, and $\mathcal{V} \subseteq \mathbb{R}^q$

is a finite set containing at most 9^q unit vectors.

Proof of Lemma B.2. This is an immediate result of Corollary 4.2.13 and Exercise 4.4.3 in Vershynin (2018). □

Lemma B.3. Under the condition of Theorem 1, we have

$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i\right| > \lambda/2\right) \leq 2\exp(-n\lambda^2/8).$$

Proof of Lemma B.3. By the definition of $(a^*, \mathbf{A}^*)$ and the first order optimality condition, we have $E(\mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i) = 0$. Note that $|\mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]| \leq 1$. Combining these two facts with Hoeffding's inequality, we have

$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top \mathbf{A}^*) + a^*\}]Y_i\right| > \lambda/2\right) \leq 2\exp(-n\lambda^2/8).$$

This completes the proof of Lemma B.3. □

Lemma B.4. Under the condition of Theorem 1, we have

$$\begin{aligned}\mathrm{pr}\left(\left\|2\lambda_0\mathbf{A}^* + \frac{1}{n}\sum_i \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top\mathbf{A}^*) + a^*\}]Y_i\mathbf{X}_i\right\| > \lambda/2\right) \\ \leq 2\cdot 9^{p+q}\exp\left\{-1/\eta_0 \min\left(\frac{\lambda^2}{64m_0^2}, \frac{\lambda}{8m_0}\right)n\right\}.\end{aligned}$$

Proof of Lemma B.4. By the definition of $(a^*, \mathbf{A}^*)$ and the first order optimality condition, we have $E(\mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top\mathbf{A}^*) + a^*\}]Y_i\mathbf{X}_i) = \mathbf{0}$. Let $\epsilon_i \overset{\mathrm{def}}{=} \mathcal{L}_h'[Y_i\{\mathrm{tr}(\mathbf{X}_i^\top\mathbf{A}^*) + a^*\}]Y_i$ for notational simplicity. We first note that $\epsilon_i\mathbf{x}_i$ is also sub-exponential where $\mathbf{x}_i \overset{\mathrm{def}}{=} \mathrm{vec}(\mathbf{X}_i)$. In fact, by an equivalent definition of sub-exponentiality (see, e.g., Theorem 2.13 in Wainwright (2019)), the assumption that $\mathbf{x}$ is sub-exponential can be stated as

$$\sup_{|\mathbf{a}|_2=1} \mathrm{pr}(|\mathbf{a}^\top\mathbf{x}| > t) \leq c_1\exp(-c_2t),$$

for all $t > 0$. This, together with $|\epsilon| \leq 1$, implies that

$$\sup_{|\mathbf{a}|_2=1} \mathrm{pr}(|\mathbf{a}^\top\mathbf{x}\epsilon| > t) \leq \sup_{|\mathbf{a}|_2=1} \mathrm{pr}(|\mathbf{a}^\top\mathbf{x}| > t) \leq c_1\exp(-c_2t),$$

which shows the sub-exponentiality of $\epsilon\mathbf{x}$. Let $\mathbf{u}_j$ and $\mathbf{v}_k$ be the covering

of the unit sphere as in the proof of Lemma B.2 and denote $\mathbf{E} \stackrel{\text{def}}{=} \sum_i \mathbf{X}_i \epsilon_i / n$ for simplicity of notation. By Corollary 1.4 in Götze et al. (2021) and Lemma B.2, there exists an absolute constants η_0 and m_0 such that

$$\begin{aligned}
& \text{pr}\left(\|2\lambda_0 \mathbf{A}^* + \mathbf{E}\| > \lambda/2\right) \\
\leq\ & \text{pr}\left(\|\mathbf{E}\| > \lambda/4\right) \\
\leq\ & \sum_{j,k} \text{pr}\left(\mathbf{u}_j^\top \mathbf{E} \mathbf{v}_k > \lambda/8\right) \\
=\ & \sum_{j,k} \text{pr}\left\{(\mathbf{v}_k \otimes \mathbf{u}_j)^\top \text{vec}(\mathbf{E}) > \lambda/8\right\} \\
\leq\ & 2 \cdot 9^{p+q} \exp\left\{-1/\eta_0 \min\left(\frac{\lambda^2}{64 m_0^2}, \frac{\lambda}{8 m_0}\right) n\right\}.
\end{aligned}$$

This completes the proof of Lemma B.4. □

Lemma B.5. Under the condition of Theorem 1, we have

$$E\left(\left\|\sum_i \sigma_i Y_i \mathbf{X}_i\right\| + \left|\sum_i \sigma_i Y_i\right|\right) \leq 4c_1\{n(p+q)\log 9\}^{1/2},$$

for some $c_1 > 0$.

Proof of Lemma B.5. We simply show that $E(\|\sum_i \sigma_i Y_i \mathbf{X}_i\|) \leq 2c_1\{n(p+q)\log 9\}^{1/2}$ for some $c_1 > 0$. By similar arguments, one can show the same upper bound for $E(|\sum_i \sigma_i Y_i|)$.

Let $\epsilon_i \stackrel{\text{def}}{=} \sigma_i Y_i$ for notational simplicity. We first note that $2\epsilon_i \mathbf{x}_i$ is also sub-exponential where $\mathbf{x}_i \stackrel{\text{def}}{=} \text{vec}(\mathbf{X}_i)$. In fact, by an equivalent definition of sub-exponentiality (see, e.g., Theorem 2.13 in Wainwright (2019)), the assumption that $\mathbf{x}$ is sub-exponential can be stated as

$$\sup_{|\mathbf{a}|_2=1} \text{pr}(|\mathbf{a}^\top \mathbf{x}| > t) \leq c_2 \exp(-c_3 t),$$

for all $t > 0$. This, together with $|\epsilon| \leq 1$, implies that

$$\sup_{|\mathbf{a}|_2=1} \text{pr}(|2\mathbf{a}^\top \mathbf{x}\epsilon| > t) \leq \sup_{|\mathbf{a}|_2=1} \text{pr}(|2\mathbf{a}^\top \mathbf{x}| > t) \leq c_2 \exp(-c_3 t/2),$$

which shows the sub-exponentiality of $2\epsilon\mathbf{x}$, in particular (see, e.g., Proposition 2.7.1 (iv) in Vershynin (2018)), we have for any pq-vector $\mathbf{a}$, $0 < t < 1/c_1$ and some $c_1 > 0$,

$$E \exp(2t\mathbf{a}^\top \sum_i \mathbf{x}_i \epsilon_i) \leq \exp(c_1^2 t^2 |\mathbf{a}|_2^2 n).$$

Let $\mathbf{u}_j$ and $\mathbf{v}_k$ be the covering of the unit sphere as in the proof of Lemma B.2 and denote $\mathbf{E} \stackrel{\text{def}}{=} 2\sum_i \epsilon_i \mathbf{X}_i$. By Jensen's inequality and Lemma B.2, we have

for any $0 < t < 1/c_1$,

$$\begin{aligned}
& \exp\{tE(\|\mathbf{E}/2\|)\} \\
\leq\ & \exp\left\{tE\left(\max_{j,k}\mathbf{u}_j^\top\mathbf{E}\mathbf{v}_k\right)\right\} \\
\leq\ & E\max_{j,k}\exp\left(t\mathbf{u}_j^\top\mathbf{E}\mathbf{v}_k\right) \\
=\ & E\max_{j,k}\exp\left\{t(\mathbf{v}_k\otimes\mathbf{u}_j)^\top\mathrm{vec}(\mathbf{E})\right\} \\
\leq\ & \sum_{j,k}E\exp\left\{t(\mathbf{v}_k\otimes\mathbf{u}_j)^\top\mathrm{vec}(\mathbf{E})\right\} \\
\leq\ & 9^{p+q}\exp(c_1^2t^2n).
\end{aligned}$$

It follows that for any $0 < t < 1/c_1$,

$$E(\|\sum_i\sigma_iY_i\mathbf{X}_i\|) \leq \{(p+q)\log 9\}/t + c_1^2tn.$$

By the condition of Theorem 1, we know

$$\{(p+q)\log 9\}^{1/2}/(c_1n^{1/2}) = o(1),$$

so for sufficiently large n,

$$\{(p+q)\log 9\}^{1/2}/(c_1n^{1/2}) \leq 1/c_1.$$

Thus, by setting $t = \{(p+q)\log 9\}^{1/2}/(c_1 n^{1/2})$, we obtain

$$E\left(\left\|\sum_i \sigma_i Y_i \mathbf{X}_i\right\|\right) \leq 2c_1\{n(p+q)\log 9\}^{1/2}.$$

This completes the proof of Lemma B.5. □

C. Proof of Theorem 2

Proof of Theorem 2. This is an immediate result of Theorem 3.1 in Hong and Luo (2017). □

D. Detailed Derivations for Algorithm 1

In the sequel, we discuss in detail the minimization problems in updates (4.5a) and (4.5b) for obtaining $\mathbf{A}_k$ and $\mathbf{r}_k$.

We first discuss how to find $\mathbf{A}_k$. The challenge of finding $\mathbf{A}_k$ is the minimization problem in (4.5a) has no closed-form solutions in general with a non-orthogonal design $\mathbb{X}$, so one needs to take multiple optimizing iterations to find an approximated minimizer (Zhu, 2017). Such a procedure, however, can be computationally intensive in large-scale problems. To overcome this challenge, we suggest employing a simple proximal gradient step to *inexactly* minimize (4.5a). In particular, we add a proximal term to the objective

function in (4.5a) and update $\mathbf{A}$ by minimizing

$$\mathbf{A}_k = \underset{\mathbf{A}\in\mathbb{R}^{p\times q}}{\arg\min}\ \lambda_0\|\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_* + \left\langle \mathbf{u}_{k-1}, \mathbf{y}\odot\left\{\mathbb{X}^\top \mathrm{vec}(\mathbf{A})\right\}\right\rangle + \frac{\tau}{2}\left|\mathbf{r}_{k-1} - \mathbf{y}\odot\left\{\mathbb{X}^\top\mathrm{vec}(\mathbf{A}) + a_{k-1}\mathbf{1}_n\right\}\right|_2^2 + \frac{1}{2}\left\|\mathrm{vec}(\mathbf{A}) - \mathbf{a}_{k-1}\right\|_{\tau\mathbf{S}}^2, \tag{D.6}$$

where $\mathbf{a}_{k-1} = \mathrm{vec}(\mathbf{A}_{k-1})$, $\mathbf{S}$ is a positive semi-definite matrix in $\mathbb{R}^{pq\times pq}$, and $\|\mathbf{v}\|_{\mathbf{S}}^2 = \mathrm{tr}\left(\mathbf{v}^\top\mathbf{S}\mathbf{v}\right)$ for $\mathbf{v}\in\mathbb{R}^{pq}$. By taking $\mathbf{S} = \eta\mathbf{I}_{pq} - 2\lambda_0/\tau\mathbf{I}_{pq} - \mathbb{X}\mathbb{X}^\top$ with $\eta \geq 2\lambda_0/\tau + \Lambda_{\max}\left(\mathbb{X}\mathbb{X}^\top\right)$, we can write problem (D.6) as follows,

$$\mathbf{A}_k = \underset{\mathbf{A}\in\mathbb{R}^{p\times q}}{\arg\min}\ \frac{\tau\eta}{2}\|\mathbf{A} - \mathbf{G}_{k-1}\|_F^2 + \lambda\|\mathbf{A}\|_*,$$

where $\mathbf{G}_{k-1} = \mathrm{reshape}\left(\mathbb{X}[\mathbf{y}\odot(\mathbf{r}_{k-1} - \mathbf{u}_{k-1}/\tau) - a_{k-1}\mathbf{1}_n] + \mathbf{S}\mathbf{a}_{k-1}, p, q\right)/\eta$. Let $\mathbf{G}_{k-1} = \mathbf{U}\mathrm{diag}(\boldsymbol{\sigma})\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{G}_{k-1}$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{\min(p,q)})^\top$. Then, we have

$$\mathbf{A}_k = \mathbf{P}_\tau(\mathbf{G}_{k-1}, \lambda/(\tau\eta)) \overset{\text{def}}{=} \mathbf{U}\mathrm{diag}(\mathbf{g})\mathbf{V}^\top,$$

where $\mathbf{g} = (g_1, \ldots, g_{\min(p,q)})^\top$ and $g_i = \max\{\sigma_i - \lambda/(\tau\eta), 0\}$.

We then discuss seeking $\mathbf{r}_k$ in (4.5b). We suggest taking Newton–Raphson iterations to find $\mathbf{r}_k$ in (4.5b) for two reasons. First, the Newton–

Raphson iterations converge quadratically to the minimizer $\mathbf{r}_k$ when the initial value (e.g. $\mathbf{r}_{k-1}$) is close to the minimizer $\mathbf{r}_k$ under suitable conditions. In our numerical studies, one or two iterations suffice to yield a quite stable solution for obtaining $\mathbf{r}_k$. Second, the Hessian matrix of $\mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1})$ with respect to $\mathbf{r}$ is diagonal and enables fast computation of the Newton–Raphson iterations. In particular, the gradient and Hessian of $\mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1})$ respect to $\mathbf{r}$ are

$$\nabla_{\mathbf{r}}\mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1}) = \nabla_{\mathbf{r}} f(\mathbf{r}) - \mathbf{u}_{k-1} + \tau\big[\mathbf{r} - \mathbf{y} \odot (\mathbb{X}^\top \mathbf{a}_k + a_k \mathbf{1}_n)\big],$$
$$\nabla^2_{\mathbf{r}}\mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1}) = \nabla^2_{\mathbf{r}} f(\mathbf{r}) + \tau \mathbf{I}_n,$$

where $\nabla_{\mathbf{r}} f(\mathbf{r}) = (\mathcal{L}'_h(r_1), \ldots, \mathcal{L}'_h(r_n))^\top/n$ and $\nabla^2_{\mathbf{r}} f(\mathbf{r}) = \mathrm{diag}\{(\mathcal{L}''_h(r_1), \ldots, \mathcal{L}''_h(r_n))^\top/n\}$. Thus, from a starting point $\mathbf{r}_{k,0} = \mathbf{r}_{k-1}$, for $j = 1, 2, \ldots$, the Newton–Raphson iterations for solving (4.5b) take the form of,

$$\begin{aligned}\mathbf{r}_{k,j} \;=\; & \mathbf{r}_{k,j-1} - (\nabla^2_{\mathbf{r}} f(\mathbf{r}_{k,j-1}) + \tau \mathbf{I}_n)^{-1}\big\{\nabla_{\mathbf{r}} f(\mathbf{r}_{k,j-1}) - \mathbf{u}_{k-1} \\ & + \tau\big[\mathbf{r}_{k,j-1} - \mathbf{y} \odot (\mathbb{X}^\top \mathbf{a}_k + a_k \mathbf{1}_n)\big]\big\}.\end{aligned}$$

For two specific kernels considered in this paper, the Gaussian kernel $\mathcal{L}^G_h(v)$ and the Epanechnikov kernel $\mathcal{L}^E_h(v)$, the first- and second-order derivatives

are

$$\begin{aligned}
\mathcal{L}_h^{G'}(v) &= -\Phi\{(1-v)/h\}, \\
\mathcal{L}_h^{G''}(v) &= (2\pi h^2)^{-1/2} \exp\left\{-(1-v)^2/(2h^2)\right\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}_h^{E'}(v) &= -I(v \le 1-h) - (1-v+h)^2\{2h-(1-v)\}\Big/ \\
&\qquad (4h^3) I(1-h < v \le 1+h), \\
\mathcal{L}_h^{E''}(v) &= 3\{h^2-(1-v)^2\}/(4h^3) I(1-h < v \le 1+h),
\end{aligned}$$

respectively. It is easy to show that $0 < \mathcal{L}_h^{G''}(v) \le 1/\{(2\pi)^{1/2}h\}$ and $0 \le \mathcal{L}_h^{E''}(v) \le 3/(4h)$. It follows that $\mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1})$ is τ-strongly convex, and $\nabla_{\mathbf{r}} \mathcal{L}_\tau(a_k, \mathbf{A}_k, \mathbf{r}, \mathbf{u}_{k-1})$ is Lipschitz continuous with the Lipschitz constants c_h being $1/\{(2\pi)^{1/2}h\}$ and $3/(4h)$, respectively. Thus, the Newton–Raphson iterates converge to $\mathbf{r}_k$ quadratically (Nesterov and Nemirovskii, 1994) when $\mathbf{r}_{k-1}$ is close to $\mathbf{r}_k$.

We give two remarks. First, given the fast quadratic convergence, we suggest just taking one step Newton–Raphson iteration. This is similar to the Metropolis-within-Gibbs in Bayesian statistics. Second, we can majorize

the Hessian matrix with $(c_h + \tau)\mathbf{I}_n$ to further reduce the computational cost. Such a majorization procedure corresponds to a proximal gradient step.

E. Impletation Details of Algorithm 1

E.1 Stopping rule

First is the stopping rule. Following the recommendation by Boyd (2010), we adopt the following stopping criterion for the proximal ADMM Algorithm 1,

$$\begin{aligned} |R_k^p|_2 &= \left|\mathbf{r}_k - \mathbf{y} \odot \left(\mathbb{X}^\top \mathbf{a}_k + a_k \mathbf{1}_n\right)\right|_2 < \epsilon_p, \\ |R_k^d|_2 &= \tau \left|\mathbb{X}[\mathbf{y} \odot (\mathbf{r}_k - \mathbf{r}_{k-1}) - (a_k - a_{k-1})\mathbf{1}_n] + \mathbf{S}(\mathbf{a}_k - \mathbf{a}_{k-1})\right|_2 < \epsilon_d, \end{aligned} \tag{E.7}$$

where R_k^p is the primal residual, R_k^d is the dual residual, and $\epsilon_p > 0$ and $\epsilon_d > 0$ are feasibility tolerances. As discussed in Boyd (2010), we can choose these tolerances using absolute and relative criteria as follows

$$\begin{aligned} \epsilon_p &= \{(n+p)q\}^{1/2}\ \epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\left\{|\mathbf{r}_k|_2, \left|\mathbf{y} \odot \left(\mathbb{X}^\top \mathbf{a}_k\right)\right|_2, n^{1/2}|a_k|\right\}, \\ \epsilon_d &= \{pq\}^{1/2}\ \epsilon^{\text{abs}} + \tau\epsilon^{\text{rel}} \max\Big\{\left|\mathbb{X}(\mathbf{y} \odot \mathbf{r}_k)\right|_2, \left|\mathbb{X}(\mathbf{y} \odot \mathbf{r}_{k-1})\right|_2, \\ &\qquad |a_k| \cdot |\mathbb{X}\mathbf{1}_n|_2, |a_{k-1}| \cdot |\mathbb{X}\mathbf{1}_n|_2, |\mathbf{S}\mathbf{a}_k|_2, |\mathbf{S}\mathbf{a}_{k-1}|_2\Big\}, \end{aligned}$$

where $\epsilon^{\text{abs}} > 0$ and $\epsilon^{\text{rel}} > 0$ are chosen as 10^{-4}. In our simulations, we also stop the iterations in Algorithm 1 once the number of iterations exceeds some maximum number of iterations, e.g., 1000.

E.2 The choice of the penalty parameter τ

Second is the choice of the Lagrangian penalty parameter τ. The penalty parameter τ is also the step size in updating the dual variable $\mathbf{u}$. Tuning τ faces a trade-off. On the one hand, τ needs to be small enough to satisfy $\tau < c/\sigma_1^2(\mathbb{X})$ to ensure the linear convergence rate according to the proof of Theorem 2, where $c > 0$ is a constant. According to Yin et al. (1988) and Allen and Perry (2024), when $n \asymp pq$, one can get that $\sigma_1^2(\mathbb{X}) \asymp n$ and hence $\tau = O(1/n)$ almost surely. On the other hand, smaller τ will impose less penalty on primal feasibility and inflate the constant within the linear convergence rate. In the optimization community, a widely used technique is the use of adaptive step size to guarantee the decrease of the sum of the primal and dual optimality gap (He et al., 2000; Wang and Liao, 2001; Boyd, 2010). In particular, we use the following scheme to adjust the value of τ at

iteration $k+1$

$$\tau_{k+1} = \begin{cases} \alpha^{\text{incr}} \tau_k, & |R_k^p|_2 > \mu |R_k^d|_2, \\ \tau_k / \alpha^{\text{decr}}, & |R_k^d|_2 > \mu |R_k^p|_2, \\ \tau_k, & \text{otherwise}, \end{cases}$$

where $\mu > 1$, $\alpha^{\text{incr}} > 1$ and $\alpha^{\text{decr}} > 1$ are constants. The above scheme ensures the primal and dual residual norms $|R_k^p|_2 / |R_k^d|_2$ and $|R_k^d|_2 / |R_k^p|_2$ vary within a factor of μ of one another as they both converge to zero. In our implementation, we take the typical choice $\mu = 10$ and $\alpha^{\text{incr}} = \alpha^{\text{decr}} = 2$ (Boyd, 2010).

E.3 Computational complexity analysis

Third is the computational complexity. There are two types of computational costs involved. One arises from calculating the matrix-vector products, while the other is from the singular value decomposition to obtain $\mathbf{A}_k$. The computational complexity for calculating the matrix-vector products is $O(npq)$. The computational complexity of performing singular value decomposition depends on the optimization algorithm used. In general, its computational complexity is upper bounded by $O\{pq \min(p, q)\}$. One can further reduce the cost by applying advanced algorithms that only involve matrix-vector products such as power iteration (Larsen, 2004), Lanczos

bidiagonalization (Baglama and Reichel, 2005; Cai and Osher, 2013) and shift-and-inverse iteration (Allen-Zhu and Li, 2016). Accordingly, the overall computational complexity for one iteration is $O(npq)$. Thanks to the superior linear convergence rate of Algorithm 1, to achieve the desired prefixed precision $\epsilon > 0$, the number of iterations required is $O\{\log(1/\epsilon)\}$. In summary, the entire computational complexity of Algorithm 1 is of order $O\{npq \log(1/\epsilon)\}$ for matrix inputs and reduces to $O\{np \log(1/\epsilon)\}$ for vector inputs with $q = 1$.

We can compare our Algorithm 1 with the classical DWD method, which also uses a smooth and convex loss function. The classical DWD method focuses on vector inputs with $q = 1$ in our setting. By Marron et al. (2007) and Egashira et al. (2021), DWD is often formulated as a second-order cone programming (SOCP) problem with $n + p$ second-order cone constraints. Solving SOCP is usually computationally extensive due to the complicated second-order cone constraints. To the best of our knowledge, the state-of-the-art optimization algorithm for solving SOCP holds a computational complexity of order $O\{(n + p)^{\omega} \log(1/\epsilon)\}$ up to some polynomic terms of $\log\{(n + p)/\epsilon\}$ to achieve a desired precision ϵ (Nesterov and Nemirovskij, 2001; Wei and Ye, 2023). Here, ω is the exponent of matrix multiplication, and the current value is about 2.37. In this regard, our proposed Algorithm

1 is computationally more efficient than DWD.

F. Extension to Tensor Data

Notations. We begin by introducing standard notations and operations commonly employed in tensor data analysis, following Kolda and Bader (2009). For positive integers $M \geq 2$ and $p_1, \ldots, p_M$, an M-dimensional array $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_M}$ is referred to as an M-way tensor. The vectorization of a tensor $\mathcal{A}$, denoted as $\mathrm{vec}(\mathcal{A})$, is a $\prod_{m=1}^{M} p_m$-dimensional vector where the jth element corresponds to $\mathcal{A}_{i_1,\ldots,i_M}$, with $j = 1 + \sum_{m=1}^{M} (i_m - 1) \prod_{m'=1}^{m-1} p_{m'}$. The inner product of two tensors of identical dimensions is defined as $\langle \mathcal{X}, \mathcal{A} \rangle = \{\mathrm{vec}(\mathcal{X})\}^\top \mathrm{vec}(\mathcal{A})$. The outer product of M vectors $\boldsymbol{\alpha}_1 \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{\alpha}_M \in \mathbb{R}^{p_M}$ is expressed as $\boldsymbol{\alpha}_1 \circ \ldots \circ \boldsymbol{\alpha}_M$, which forms a $p_1 \times \ldots \times p_M$ tensor with the $(j_1, \ldots, j_M)$th element equal to $\prod_{m=1}^{M} \alpha_{m j_m}$.

Methodology for Tensor Data. The convolution-type smoothing technique we propose is applicable to more general and practical tensor inputs, often encountered in fields such as image processing, social network analysis, and digital marketing. Consider a dataset $\{(Y_i, \mathcal{X}_i)\}_{i=1}^{n}$, where $Y_i \in \{-1, 1\}$ represents a binary label and $\mathcal{X}_i \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_M}$ is an M-way tensor-valued predictor. A natural approach is to identify a hyperplane that separates the

two classes of data points by solving the following optimization problem:

$$\min_{a\in\mathbb{R}^1,\mathcal{A}\in\mathbb{R}^{p_1\times p_2\times\cdots\times p_M}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_h\{Y_i(\langle\mathcal{X}_i,\mathcal{A}\rangle+a)\}+\lambda_0\sum_{i_1=1,\ldots,i_M=1}^{p_1,\ldots,p_M}\mathcal{A}^2_{i_1,\ldots,i_M},$$

where $\mathcal{L}_h$ denotes our newly introduced smoothed hinge loss function. The primary challenge here is that $\mathcal{A}$ contains $\prod_m p_m$ parameters, which could grow speedily. To effectively reduce the dimensionality of the classifier, we recommend employing canonical decomposition or parallel factors (CANDECOMP/PARAFAC, or CP) decomposition (Kolda and Bader, 2009) on the tensor coefficient $\mathcal{A}$:

$$\mathcal{A}=\sum_{k=1}^{K}\boldsymbol{\alpha}^{(1k)}\circ\boldsymbol{\alpha}^{(2k)}\circ\cdots\circ\boldsymbol{\alpha}^{(Mk)},$$

where $\boldsymbol{\alpha}^{(mk)}\in\mathbb{R}^{p_m}$ for each m and k, and K is the rank of the CP decomposition. For simplicity, we represent the above CP decomposition as

$$\mathcal{A}=[\![\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_M]\!],$$

where $\mathbf{A}_m=(\boldsymbol{\alpha}^{(m1)},\ldots,\boldsymbol{\alpha}^{(mK)})\in\mathbb{R}^{p_m\times K}$ for each $m=1,\ldots,M$. With the CP decomposition, the parameter count in $\mathcal{A}$ is significantly reduced from $\prod_m p_m$ to $K\sum_m p_m$. To ensure the identifiability of $\boldsymbol{\alpha}^{(mk)}$ in the presence

of scaling and permutation ambiguities (Zhou et al., 2013), the parameter space is typically restricted to

$$\mathcal{S}_{\mathcal{A}} = \{[\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M]\!] : \alpha_1^{(mk)} = 1, m = 1, \ldots, M-1, k = 1, \ldots, K, \\ \alpha_1^{(M1)} \geq \alpha_1^{(M2)} \geq \ldots \geq \alpha_1^{(MK)}\}.$$

Regularized Estimation. In typical neuroimaging studies, the sample size is often limited. Even with a low-rank tensor SVM model, the number of parameters frequently exceeds the sample size. Consequently, the scenario where $p_m > n$ is commonplace in neuroimaging analysis. To address this challenge, imposing structural assumptions and regularization is crucial for stabilizing the classification task. Different penalties cater to various structures and objectives. Here, we focus on the widely adopted *sparsity* structure and its associated penalties. Leveraging the dimension reduction afforded by CP decomposition, we consider the following optimization problem:

$$\min_{a \in \mathbb{R}^1, \mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_M}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_h\{Y_i(\langle \mathcal{X}_i, [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M]\!] \rangle + a)\} \\ + \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{p_m} \mathcal{P}_\tau(|\alpha_i^{(mk)}|, \lambda_0),$$

where $\mathcal{P}_\iota(|\alpha|, \lambda_0)$ represents a scalar penalty function, λ_0 is the penalty tuning parameter, and ι is an index for the penalty family. Notable penalties include the power family (Frank and Friedman, 1993), where $\mathcal{P}_\iota(|\alpha|, \lambda_0) = \lambda_0|\alpha|^\iota$ with $\iota \in (0, 2]$, notably including lasso (Tibshirani, 1996) ($\iota = 1$) and ridge ($\iota = 2$); elastic net (Zou and Hastie, 2005), defined as $\mathcal{P}_\iota(|\alpha|, \lambda_0) = \lambda_0\{(1-\iota)|\alpha| + \iota\alpha^2\}$, with $\iota \in [0, 1]$; and SCAD (Fan and Li, 2001), where $\partial/\partial|\alpha|\mathcal{P}_\iota(|\alpha|, \lambda_0) = \lambda_0[I(|\alpha| \leq \lambda_0) + (\iota\lambda_0 - |\alpha|)_+/\{(\iota-1)\lambda_0\}I(|\alpha| > \lambda_0)]$ for $\iota > 2$, among others. While it is conceptually feasible to disregard the tensor structure and directly apply the penalty to the full coefficient array $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_M}$–effectively treating $\text{vec}(\mathcal{X}_i)$–this approach may lead to an explosion in dimensionality, particularly in brain imaging applications with tensor structures. For instance, the dimensionality could reach $64^3 = 262,144$ or even $256^3 = 16,777,216$ variables, which could severely degrade both statistical and computational performance. One can directly extend our proximal ADMM Algorithm 1 for handling these penalties with convexity and integrate Algorithm 1 with a local linear approximation for handling nonconvex penalties.

We emphasize that the CP decomposition is only one of several methods for modeling the low-rank structure of tensors, chosen here for illustrative purposes. Other choices include the Tucker decomposition.

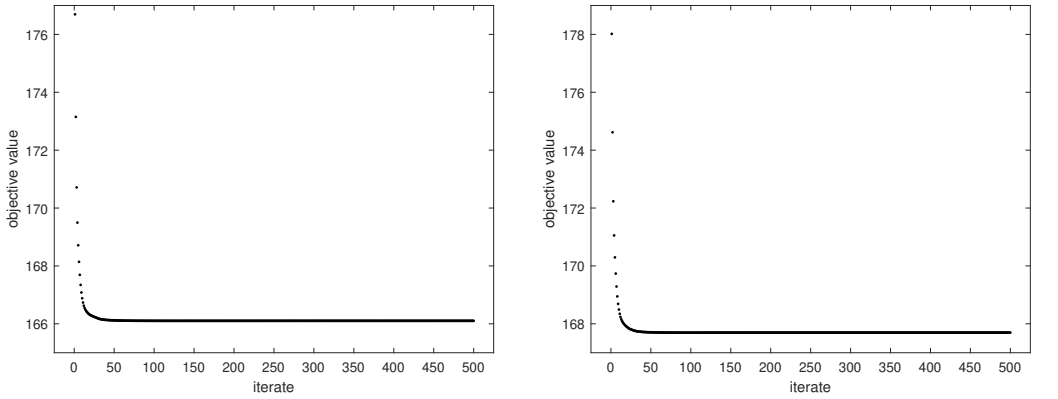
G. Additional Numerical Experiments

We study the impact of different kernels and penalties on our proposed CSMM. We fix $p = 50$, $q = 30$ and the sample size $n = 300$. We consider LRCSMM-G, LRCSMM-E, EnetCSMM-G, and EnetCSMM-E in this study. Table G.1 collects the results. From Table G.1, we can see that the CSMM with Gaussian kernel performs slightly better than CSMM with Epanechnikov kernel. Thus, we only consider CSMM with Gaussian kernel in the following numerical studies.

To visualize the fast linear convergence of our pADMM algorithm for solving the penalized CSMM, we present a visualization of the objective values in Figure G.1. These values are obtained by applying pADMM to the penalized CSMM with the validated tuning parameters in Example 1 with $d = -1$. Figure G.1 clearly demonstrates the rapid decrease in the objective value throughout the optimization process.

Table G.1: The prediction error (in percentages) and runtime (in seconds) of low-rank convoluted SMM with Gaussian and Epanechnikov kernels, and elastic-net convoluted SMM with Gaussian and Epanechnikov kernels. Under each simulation setting, the method with the lowest prediction error is marked by a black box. All the results are averaged over 50 independent runs.

	LRCSMM-G		LRCSMM-E		EnetCSMM-G		EnetCSMM-E	
d	err (%)	time	err (%)	time	err (%)	time	err (%)	time
Example 1								
-0.5	**15.67**	0.11	18.33	0.36	**15.67**	0.11	21.34	0.15
-1	7.67	0.1	**7.00**	0.23	**7.00**	0.14	7.67	0.11
-1.5	**2.33**	0.14	4.67	0.19	**2.33**	0.14	5.67	0.14
Example 2								
-0.5	**31.00**	0.09	37.00	0.18	31.33	0.09	36.00	0.16
-1	**19.33**	0.09	23.00	0.25	**19.33**	0.09	20.33	0.16
-1.5	**14.33**	0.13	18 .00	0.14	14.67	0.12	18.00	0.15



(a) Low-rank CSMM with Gaussian kernel (b) Elastic-net CSMM with Gaussian kernel

Figure G.1: Convergence plot for solving penalized CSMM with validated tuning parameters using Algorithm 1 under the simulation setting in Example 1 with $d = -1$ based on 50 independent runs.

References

Allen, G. I. and P. O. Perry (2024). Singular Value Decomposition and High-Dimensional Data. `https://ptrckprry.com/reports/svd-hdd.pdf`.

Allen-Zhu, Z. and Y. Li (2016). LazySVD: Even faster SVD decomposition yet without agonizing pain. *Advances in Neural Information Processing Systems 29*.

Baglama, J. and L. Reichel (2005). Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing 27*(1), 19–42.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Cai, J.-F. and S. Osher (2013). Fast singular value thresholding without singular value decomposition. *Methods and Applications of Analysis 20*(4), 335–352.

Egashira, K., K. Yata, and M. Aoshima (2021). Asymptotic properties of distance-weighted discrimination and its bias correction for high-dimension, low-sample-size data. *Japanese Journal of Statistics and Data Science 4*(2), 821–840.

Fan, J. and R. Li (2001, December). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Frank, l. E. and J. H. Friedman (1993, May). A Statistical View of Some Chemometrics Regression Tools. *Technometrics 35*(2), 109–135.

Götze, F., H. Sambale, and A. Sinulis (2021, January). Concentration inequalities for polyno-

mials in α-sub-exponential random variables. *Electronic Journal of Probability 26*(none).

He, B.-S., H. Yang, and S. Wang (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications 106*, 337–356.

Hong, M. and Z.-Q. Luo (2017, March). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming 162*(1), 165–199.

Kolda, T. G. and B. W. Bader (2009, August). Tensor Decompositions and Applications. *SIAM Review 51*(3), 455–500.

Larsen, R. M. (2004, August). Propack-software for large and sparse svd calculations. Available at http://sun.stanford.edu/rmunk/PROPACK..

Ledoux, M. and M. Talagrand (2011). *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Berlin ; London: Springer.

Marron, J. S., M. J. Todd, and J. Ahn (2007). Distance-weighted discrimination. *Journal of the American Statistical Association 102*(480), 1267–1271.

Nesterov, J. E. and A. S. Nemirovskij (2001). *Interior-Point Polynomial Algorithms in Convex Programming* (3. printing ed.). Number 13 in SIAM Studies in Applied Mathematics. Philadelphia, Pa: SIAM.

Nesterov, Y. and A. Nemirovskii (1994, January). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

van der Vaart, A. W. and J. A. Wellner (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.

Vershynin, R. (2018). *High-Dimensional Probability, An Introduction with Applications in Data Science*. Cambridge: Cambridge University Press.

Wainwright, M. J. (2019, February). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (First ed.). Cambridge University Press.

Wang, S. and L. Liao (2001). Decomposition method with a variable parameter for a class of monotone variational inequality problems. *Journal of optimization theory and applications 109*, 415–429.

Wei, M. and G. Ye (2023). Solving Second-Order Cone Programs Deterministically in Matrix Multiplication Time. `https://shorturl.at/l1Q4G`.

Yin, Y. Q., Z. D. Bai, and P. R. Krishnaiah (1988, August). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields 78*(4), 509–521.

Zhou, H., L. Li, and H. Zhu (2013, June). Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association 108*(502), 540–552.

Zhu, Y. (2017, January). An Augmented ADMM Algorithm With Application to the General-

ized Lasso Problem. *Journal of Computational and Graphical Statistics 26*(1), 195–204.

Zou, H. and T. Hastie (2005, April). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.