Factorized Sparse Learning Models with Interpretable High Order Feature Interactions

Sanjay Purushotham*
University of Southern California
Los Angeles, CA, USA
spurusho@usc.edu

Martin Renqiang Min†
NEC Labs America
Princeton, NJ, USA
renqiang@nec-labs.com

C.-C. Jay Kuo
University of Southern California
Los Angeles, CA, USA
cckuo@sipi.usc.edu

Rachal Ostroff‡
SomaLogic, Inc.
Boulder, CO, USA
rostroff@somalogic.com

ABSTRACT

In this supplementary material, we will provide proofs for the Lemma 5.4 and Theorem 5.5 presented in section 5.2 of our paper. Our proofs is based on the papers [1], [2].

1. PROOFS

Please refer to our main paper for the regularity conditions (C4)-(C6) and the statements for the following Lemma 5.4 and Theorem 5.5.

PROOF OF LEMMA 5.4. Let $\eta_n = \sqrt{p_n}n^{-1/2} + a_n$ and $\{\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$ be the ball around $\boldsymbol{\theta}_n^*$, where $\boldsymbol{\delta} = (u_1, ..., u_p, v_{11}, ...v_{Kp})^T = (\mathbf{u}^T, \mathbf{v}^T)^T$.

Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}_n^*)$$

Let $-L_n$ and nP_n denote the first and second terms of Q_n. For any $\boldsymbol{\delta}$ that satisfies $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) =& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&+ nP_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) - nP_n(\boldsymbol{\theta}_n^*) \\
=& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&+ n\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}(|\beta_j^* + \eta_n u_j| - |\beta_j|) \\
&+ n\sum_{(k,l)\in\mathcal{A}_{n2}} \lambda_{nl}^{\alpha_k}(|\alpha_{k,l}^* + \eta_n v_{k,l})| - |\alpha_{k,l}^*| \\
\geq& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&- n\eta_n\Big\{\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}|u_j| + \\
&+ \sum_{(k,l)\in\mathcal{A}_{n2}} \lambda_{nl}^{\alpha_k}\alpha_{k,l}^*|v_{k,l})|\Big\} \\
\geq& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&- n\eta_n(\sum_{j\in\mathcal{A}_{n1}} a_n|u_j| + \sum_{(k,l)\in\mathcal{A}_{n2}} a_n|v_{k,l}|) \\
\geq& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n(\sqrt{s_n}a_n)d \\
=& - L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n^2 d \\
=& - [\nabla L_n(\boldsymbol{\theta}_n^*)]^T(\eta_n\boldsymbol{\delta}) - \frac{1}{2}(\eta_n\boldsymbol{\delta})^T[\nabla^2 L_n(\boldsymbol{\theta}_n^*)](\eta_n\boldsymbol{\delta}) \\
&- \frac{1}{6}\nabla^T\{\boldsymbol{\delta}^T[\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)]\boldsymbol{\delta}\}\boldsymbol{\delta}\eta_n^3 \\
&- n\eta_n^2 d \quad \text{(By Taylor's series expansion)}
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta})$ and $\boldsymbol{\theta}_n^*$. We split the above into four parts:

$$K_1 = -[\nabla L_n(\boldsymbol{\theta}_n^*)]^T(\eta_n\boldsymbol{\delta})$$

$$K_2 = -\frac{1}{2}(\eta_n\boldsymbol{\delta})^T[\nabla^2 L_n(\boldsymbol{\theta}_n^*)](\eta_n\boldsymbol{\delta})$$

$$K_3 = -\frac{1}{6}\nabla^T\{\boldsymbol{\delta}^T[\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)]\boldsymbol{\delta}\}\boldsymbol{\delta}\eta_n^3$$

$$K_4 = -n\eta_n^2 d$$

Then,

$$\begin{aligned}
|K_1| =& | - \eta_n[\nabla L_n(\boldsymbol{\theta}_n^*)]^T\boldsymbol{\delta}| \\
\leq& \eta_n||(\nabla L_n(\boldsymbol{\theta}_n^*))^T||||\boldsymbol{\delta}|| \\
=& O_p(\eta_n\sqrt{np_n})\boldsymbol{\delta} \\
=& O_p(n\eta_n^2)d
\end{aligned}$$

Next, since we have

$$\left|\left|\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}_n^*) + \mathbf{I}_n(\boldsymbol{\theta}_n^*)\right|\right| = o_p(1/p_n) \tag{1}$$

*Co-first author
†Co-first author, corresponding author
‡To whom data request should be sent

KDD '14 New York, USA

by Chebyshev's inequality and (C5) we can show that

$$\begin{aligned}K_2 =& -\frac{1}{2}(\eta_n\boldsymbol{\delta})^T[\nabla^2 L_n(\boldsymbol{\theta}_n^*)](\eta_n\boldsymbol{\delta})\\ =&\frac{1}{2}n\eta_n^2\boldsymbol{\delta}^T[\mathbf{I}_n(\boldsymbol{\theta}_n^*)]\boldsymbol{\delta}-\frac{1}{2}n\eta_n^2 d^2 o_p(1)\end{aligned}$$

Moreover, by Cauchy-Schwarz inequality, (C6) and the conditions $\sqrt{n}a_n \to 0$ and $p_n^5/n \to 0$,

$$\begin{aligned}|K_3| =&\Big|-\frac{1}{6}\nabla^T\{\boldsymbol{\delta}^T[\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)]\boldsymbol{\delta}\}\boldsymbol{\delta}\eta_n^3\Big|\\ =&\frac{1}{6}\eta_n^3\Big|\sum_{i=1}^{n}\sum_{j,l,m=1}^{p_n}\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}_n})}{\partial\theta_{nj}\partial\theta_{nl}\partial\theta_{nm}}\delta_j\delta_l\delta_m\Big|\\ \le&\eta_n^3\sum_{i=1}^{n}\Big(\sum_{j,l,m=1}^{p_n}M_{njlm}^2(\mathbf{V}_{ni})\Big)^{1/2}||\boldsymbol{\delta}||^3\\ =&n\eta_n^3 O_p(p_n^{3/2})(p_n O(1))^{1/2}||\boldsymbol{\delta}||^2\\ =&n\eta_n^2 O_p(\eta_n p_n^2)d^2\\ =&n\eta_n^2 o_p(1)d^2\end{aligned}$$

$$\begin{aligned}D_n(\boldsymbol{\delta}) \ge& K_1+K_2+K_3+K_4\\ =&-O_p(n\eta_n^2)\boldsymbol{\delta}-\frac{1}{2}n\eta_n^2\boldsymbol{\delta}^T[\mathbf{I}_n(\boldsymbol{\theta}_n^*)]\boldsymbol{\delta}\\ &-\frac{3}{2}n\eta_n^2 o_p(1)d^2-n\eta_n^2 d\end{aligned}$$

We see that K_2 dominates the rest of the terms and is positive since $I(\boldsymbol{\theta}_n)$ is positive definite at $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n^*$ from (C5). Therefore, for any given $\epsilon > 0$ there exists a large enough constant d such that

$$P\{\inf_{||\delta||=d}Q_n(\boldsymbol{\theta}_n^*+\eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}_n^*)\} \ge 1-\epsilon$$

This implies that with probability at-least $1-\epsilon$, there exists a local minimizer in the ball $\{\boldsymbol{\theta}_n^*+\eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \le d\}$. Thus, there exists a local minimizer of $Q_n(\boldsymbol{\theta}_n)$ such that $||\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}_n^*|| = O_p(\eta_n\sqrt{p_n})$. □

Proof of Theorem 5.5. [Proof of Sparsity]

First, we prove $P(\hat{\beta}_{\mathcal{A}_{n1}^c} = 0) \to 1$ as $n \to \infty$. It is sufficient to show that with probability tending to 1, for any $j \in \mathcal{A}_{n1}^c$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} < 0 \text{ for } -\epsilon_n < \hat{\beta}_{nj} < 0 \tag{2}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} > 0 \text{ for } \epsilon_n > \hat{\beta}_{nj} > 0 \tag{3}$$

where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (2), we consider Taylor expansion of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_n^*$.

$$\begin{aligned}\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} =& -\frac{\partial L_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}}+n\lambda_{nj}^{\beta}sgn(\hat{\beta}_{nj})\\ =&-\frac{\partial L_n(\boldsymbol{\theta}_n^*)}{\partial\beta_{nj}}-\sum_{k=1}^{p_n}\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial\beta_{nj}\partial\theta_{nk}}(\hat{\theta}_{nk}-\theta_{nk}^*)\\ &-\sum_{k=1}^{p_n}\sum_{k=1}^{p_n}\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}_n})}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}(\hat{\theta}_{nk}-\theta_{nk}^*)(\hat{\theta}_{nl}-\theta_{nl}^*)\\ &+n\lambda_{nj}^{\beta}sgn(\hat{\beta}_{nj})\end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ and $\boldsymbol{\theta}_n^*$. By (C4)-(C6), the lemma 5.4, and carefully solving the parts of above equation (4) using Cauchy Schwarz inequality, we have

$$\begin{aligned}\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} =&O_p(\sqrt{np_n})+n\lambda_{nj}^{\beta}sgn(\hat{\beta}_{nj})\\ =&\sqrt{np_n}\Big\{O_p(1)+\sqrt{n/p_n}\lambda_{nj}^{\beta}sgn(\hat{\beta}_{nj})\Big\}\end{aligned}$$

Since $\sqrt{n/p_n}b_n \to \infty$, $sgn(\hat{\beta}_{nj})$ dominates the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}}$ when n is large. Thus,

$$P\left(\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} > 0 \text{ for } 0 < \hat{\beta}_{nj} < \epsilon_n\right) \to 1 \text{ as } n \to \infty$$

(3) can be shown in the same way.
Also, $P(\hat{\boldsymbol{\alpha}}_{n\mathcal{A}_{n2}^c} = 0) \to 1$ can be proved similarly since in our model $\boldsymbol{\beta}_n$ and $\boldsymbol{\alpha}_n$ are independent of each other. □

Proof of Theorem 5.5. [Proof of Asymptotic normality]

We want to show that with probability tending to 1,

$$\begin{aligned}&\sqrt{n}\mathbf{A}_n\mathbf{I}_n^{1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}_{n\mathcal{A}_n}^*) = \sqrt{n}\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)\\ &\Big\{\frac{1}{n}\nabla L_n(\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*))+o_p(n^{-1/2})\Big\}\end{aligned} \tag{5}$$

Also, we need to show that probability tending to 1,

$$\begin{aligned}&\sqrt{n}\mathbf{A}_n\mathbf{I}_n^{1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}_{n\mathcal{A}_n}^*)\\ &=\frac{1}{\sqrt{n}}\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)\sum_{i=1}^{n}[\nabla L_{ni}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)]\\ &\quad+o_p(\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)\mathbf{1}_{s_n\times 1})\\ &=\frac{1}{\sqrt{n}}\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)\sum_{i=1}^{n}[\nabla L_{ni}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)]+o_p(1)\\ &\equiv\sum_{i=1}^{n}Y_{ni}+o_p(1)\\ &\to_d\quad N(\mathbf{0},\mathbf{G})\end{aligned} \tag{6}$$

where $\mathbf{Y}_{ni} = \frac{1}{\sqrt{n}}\mathbf{A}_n\mathbf{I}_n^{-1/2}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)[\nabla L_{ni}(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)]$. We will now prove (5) and (6) in (I) and (II) respectively.

(I) We want to show
$\mathbf{I}_n(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}_{n\mathcal{A}_n}^*) = \frac{1}{n}\nabla L_n(\mathbf{A}_n(\boldsymbol{\theta}_{n\mathcal{A}_n}^*)+o_p(n^{-1/2})$.

We know that with probability tending to 1,

$$\begin{aligned}\nabla_{\mathcal{A}_n}Q_n(\hat{\boldsymbol{\theta}_{n\mathcal{A}_n}}) =& -\nabla_{\mathcal{A}_n}L_n(\hat{\boldsymbol{\theta}_{n\mathcal{A}_n}})+n\nabla_{\mathcal{A}_n}P_{\lambda_n}(\hat{\boldsymbol{\theta}_{n\mathcal{A}_n}})\\ =&\mathbf{0}\end{aligned}$$

By Taylor's expansion of $\nabla_{\mathcal{A}_n}L_n(\hat{\boldsymbol{\theta}_{n\mathcal{A}_n}})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_{n\mathcal{A}_n}^*$, and

substituting it in (5), we get

$$\begin{aligned}
&\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&= -\frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&= \frac{1}{n}\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad - \frac{1}{2n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^T[\nabla^2_{\mathcal{A}_n}(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}))] \\
&\quad (\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) - \nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})
\end{aligned}$$

Therefore, it is sufficient to show that

$$\begin{aligned}
&- \frac{1}{2n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^T[\nabla^2_{\mathcal{A}_n}(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}))] \\
&(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) - \nabla_{n\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&+ \Big\{\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\equiv A_1 + A_2 + A_3 \\
&= o_p(n^{-1/2})
\end{aligned}$$

Now, using Cauchy-Schwartz inequality and (C6), we can show that

$$||A_1||^2 = o_p(1/n)$$

Since $a_n = o(1/\sqrt{np_n})$ from the condition in the theorem,

$$\begin{aligned}
||A_2||^2 =& \left|\left|(\lambda^{\beta}_{n1} sgn(\beta^*_{n1}), ..., \lambda^{\boldsymbol{\alpha}_k}_{n,Kp} sgn(\boldsymbol{\alpha}^*_{k,Kp}))^T\right|\right|^2 \\
\leq& s_n[\max\{\lambda^{\beta}_{nj}, \lambda^{\boldsymbol{\alpha}_k}_{n,l} : j \in \mathcal{A}_{n1}, (k,l) \in \mathcal{A}_{n2}\}]^2 \\
=& s_n a_n^2 = s_n o(1/np_n) \\
=& o(1/n)
\end{aligned}$$

Now, from equation (1), we can show that

$$\begin{aligned}
||A_3||^2 \leq& ||\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})||^2 \\
& ||(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})||^2 \\
=& o_p(1/p_n^2) O_p(p_n/n) = o_p(1/np_n) \\
=& o_p(1/n)
\end{aligned}$$

Therefore, we get,

$$A_1 + A_2 + A_3 = o_p(n^{-1/2})$$

(II) Now, we show $\sum_{i=1}^n Y_{ni} + o_p(1) \to_d N(\mathbf{0}, \mathbf{G})$ where

$$\mathbf{Y}_{ni} = \frac{1}{\sqrt{n}}\mathbf{A}_n \mathbf{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})[\nabla_{\mathcal{A}_n} L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})]$$

Now, we need to show that $\mathbf{Y}_{ni}$ satifies the conditions for Lindeberg-Feller central limit theorem. For any $\epsilon > 0$, by Cauchy-Schwartz inequality, we get

$$\begin{aligned}
\sum_i^n E[||\mathbf{Y}_{ni}||^2 I\{||\mathbf{Y}_{ni}|| > \epsilon\}] &= nE[||\mathbf{Y}_{n1}||^2 I\{||\mathbf{Y}_{n1}|| > \epsilon\}] \\
&\leq n[E[||\mathbf{Y}_{n1}||^4]^{1/2}[E(1\{||\mathbf{Y}_{n1}|| > \epsilon\})]^{1/2} \\
&= nA_4^{1/2} A_5^{1/2}
\end{aligned}$$

Now, solving for A_4 we get,

$$\begin{aligned}
A_4 =& \frac{1}{n^2} E||\mathbf{A}_n \mathbf{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})[\nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})]||^4 \\
\leq& \frac{1}{n^2}||\mathbf{A}_n^T \mathbf{A}_n||^2 ||\mathbf{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})||^2 \\
& E[\nabla^T_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})]^2 \\
=& \frac{1}{n^2}\lambda^2_{\max}(\mathbf{A}_n^T \mathbf{A}_n)\lambda^2_{\max}(\mathbf{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}))O(s_n^2) \\
=& O(p_n^2/n^2)
\end{aligned}$$

Now, by Markov inequality,

$$\begin{aligned}
A_5 =& P(||\mathbf{Y}_{n1}|| > \epsilon) \\
\leq& \frac{E(||\mathbf{Y}_{n1}||^2)}{\epsilon^2} \\
=& O(p_n/n)
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\sum_i^n E[||\mathbf{Y}_{ni}||^2 I\{||\mathbf{Y}_{ni}|| > \epsilon\}] =& nO(p_n/n)O(\sqrt{p_n/n}) \\
=& o(1)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\sum_{i=1}^n Cov(\mathbf{Y}_{ni}) =& nCov(\mathbf{Y}_{n1}) \\
=& \mathbf{A}_n \mathbf{A}_n^T \to G
\end{aligned}$$

Since $\mathbf{Y}_{ni}, i = 1, ..., n$ satisfies the conditions for Lindeberg-Feller central limit theorem, we have

$$\sum_{i=1}^n Y_{ni} + o_p(1) \to_d N(\mathbf{0}, \mathbf{G})$$

□

References

[1] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

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

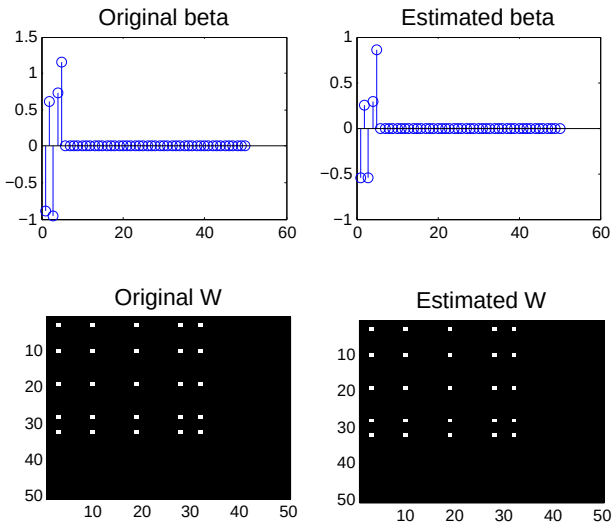

Figure 1: Support Recovery of β (90 % sparse) and W (99 % sparse) for synthetic data Case 1: $n > p$ and $q > n$ where $n = 1000, p = 50, q = 1275$.

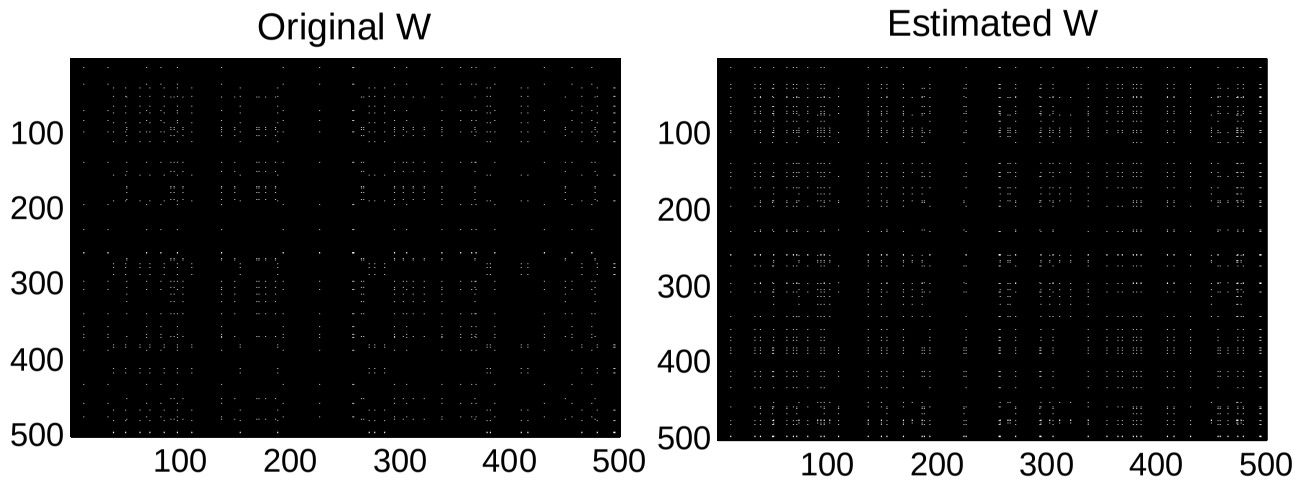

Figure 2: Support Recovery of W (99.5 % sparse) for synthetic data Case 2: $p > n$ where p = 500, n = 100.