Partial Consistency with Sparse Incidental Parameters

Jianqing Fan, Runlong Tang and Xiaofeng Shi

Princeton University, Johns Hopkins University and Princeton University

Supplementary Materials

In this supplementary file, we provide a supplement for each of Sections 1, 2 and 3 of the paper. The proofs of the theoretical results in Sections 2 and 3 of the paper are shown in Sections B and C, respetively. Furthermore, we also study an extension case where the number of covariates grows with but slower than the sample size in Section D.

A Supplement to Section 1

In this supplement, we first show that the method proposed by Neyman & Scott (1948) does not work for model (1.1) and then explain which assumptions or conditions for the consistency results of the penalized methods in Zhao & Yu (2006), Fan & Peng (2004) and Fan et al. (2011) are not satisfied for model (1.1).

Although the modified equations of maximum likelihood method proposed by Neyman & Scott (1948) could handle "a number of important cases" with incidental parameters, unfortunately, it does not work for model (1.1). More specifically, consider the simplest case of model (1.1) with $d = 1$:

$$Y_i = \mu_i^\star + X_i\beta^\star + \epsilon_i, \text{ for } i = 1, 2, \ldots, n,$$

where $\{\epsilon_i\}$ are i.i.d. copies of $N(0, \sigma^2)$. Using the notations of Neyman & Scott (1948), the likelihood function for (X_i, Y_i) is given by $p_i = p_i(\beta, \sigma, \mu_i | X_i, Y_i) = (\sqrt{2\pi}\sigma)^{-1} \exp\{-(2\sigma^2)^{-1}(Y_i - \mu_i^\star - X_i\beta)^2\}$, and the log-likelihood function is $\log p_i = -\log(\sqrt{2\pi}\sigma) - (2\sigma^2)^{-1}(Y_i - \mu_i^\star - X_i\beta)^2$. Then, the score functions are

$$\begin{aligned}
\phi_{i1} &= \frac{\partial \log p_i}{\partial \beta} = \frac{1}{\sigma^2}(Y_i - \mu_i^\star - X_i\beta)X_i, \\
\phi_{i2} &= \frac{\partial \log p_i}{\partial \sigma} = \frac{1}{\sigma} + \frac{1}{\sigma^3}(Y_i - \mu_i^\star - X_i\beta)^2, \\
\omega_i &= \frac{\partial \log p_i}{\partial \mu_i} = \frac{1}{\sigma^2}(Y_i - \mu_i^\star - X_i\beta).
\end{aligned}$$

From the equation $\omega_i = 0$, we have $\hat{\mu}_i = Y_i - X_i\beta$. Plugging this $\hat{\mu}_i$ into ϕ_{i1} and ϕ_{i2} (replacing μ_i with $\hat{\mu}_i$), we obtain $\phi_{i1} = 0$ and $\phi_{i2} = 1/\sigma$. Then, $E_{i1} = \mathbb{E}\phi_{i1} = 0$ and $E_{i2} = \mathbb{E}\phi_{i1} = 1/\sigma$. Thus, E_{i1} and E_{i2} do only depend on the structural parameters ($\boldsymbol{\beta}^\star$ and σ). However, we then have $\Phi_{i1} = \phi_{i1} - E_{i1} = 0$ and $\Phi_{i2} = \phi_{i2} - E_{i2} = 0$.

This means $F_{n1} = F_{n2} = 0$, independent of structural parameters! Consequently, the estimation equations degenerate to two $0 = 0$ equations, which means the modified equation of maximum likelihood method does not work for model (1.1).

Next, we explain which assumptions or conditions for the consistency results of the penalized methods in Zhao & Yu (2006), Fan & Peng (2004) and Fan et al. (2011) are not valid for model (1.1).

Zhao & Yu (2006) derive strong sign consistency for lasso estimator. However, their consistency results Theorems 3 and 4 do not apply to model (1.1), since the above specific design matrix $\boldsymbol{X}$ does not satisfy their regularity condition (6) on page 2546. More specifically, with model (1.1),

$$C_{11}^n = \frac{1}{n}\begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{X}_{1,s} \\ \boldsymbol{X}_{1,s}^T & \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T \end{pmatrix} \xrightarrow{a.s.} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_X \end{pmatrix},$$

where $\boldsymbol{\Sigma}_X$ is the covariance matrix of the covariates. This means that some of the eigenvalues of C_{11}^n goes to 0 as $n \to \infty$. Then the regularity condition (6), which is

$$\alpha^T C_{11}^n \alpha \geq \text{ a positive constant }, \text{ for all } \alpha \in \mathbb{R}^{s+d} \text{ such that } \|\alpha\|_2^2 = 1,$$

does not hold any more. Thus the consistency results Theorems 3 and 4 in Zhao & Yu (2006) is not applicable for model (1.1).

Fan & Peng (2004) show the consistency with Euclidean metric of a penalized likelihood estimator when the dimension of the sparse parameter increases with the sample size in Theorem 1 on Page 935. Under their framework, the log-likelihood function of the data point $V_i = (\boldsymbol{X}_i, Y_i)$ for each i from model (1.1) with random errors being i.i.d. copies of $N(0, \sigma^2)$ is given by

$$\log f_n(V_i, \mu_i, \boldsymbol{\beta}) \propto -\frac{1}{2\sigma^2}(Y_i - \mu_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2,$$

where $\propto$ means "proportional to". As we can see that log-likelihood functions with different i's might different since μ_i's might be different for different i's. This violates a condition that all the data points are i.i.d. from a structural density in Assumption (G) on Page 934.

This violation might not be essential, however, since we could consider the log-likelihood function for all the data directly. That is, we consider

$$L_n(\boldsymbol{\mu}, \boldsymbol{\beta}) = \sum_{i=1}^n \log f_n(V_i, \mu_i, \boldsymbol{\beta}) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2.$$

Then, the Fisher information matrix for $(\boldsymbol{\mu}, \boldsymbol{\beta})$ is given by

$$I_{n+d}(\boldsymbol{\mu}, \boldsymbol{\beta}) = \begin{pmatrix} \sigma^{-2}\boldsymbol{I}_n & 0 \\ 0 & n\sigma^{-2}\boldsymbol{\Sigma}_X^2 \end{pmatrix},$$

where I_n is the $n \times n$ identity matrix. Then, the Fisher information for one data point is

$$\frac{1}{n} I_{n+d}(\boldsymbol{\mu}, \boldsymbol{\beta}) = \begin{pmatrix} n^{-1}\sigma^{-2}\boldsymbol{I}_n & 0 \\ 0 & \sigma^{-2}\boldsymbol{\Sigma}_X^2 \end{pmatrix}.$$

It is clear that the minimal eigenvalue $\lambda_{\min}(I_{n+d}(\boldsymbol{\mu}, \boldsymbol{\beta})/n) = n^{-1}\sigma^2 \to 0$ as $n \to \infty$. This violates the condition that the minimal eigenvalue should be lower bounded from 0 in Assumption (F) on Page 934. Thus, the consistency result Theorem 1 in Fan & Peng (2004) can not be applied to model (1.1).

Fan et al. (2011) "consider the variable selection problem of nonpolynomial dimensionality in the context of generalized linear models" by taking the penalized likelihood approach with folded-concave penalties. Theorem 3 on page 5472 of Fan et al. (2011) shows that there exists a consistent estimator of the unknown parameters with the Euclidean metric under certain conditions. In Condition 4 on page 5472, there is a condition on a minimal eigenvalue

$$\min_{\boldsymbol{\delta} \in N_0} \lambda_{\min}[\boldsymbol{X}_I^T \boldsymbol{\Sigma}(\boldsymbol{X}_I \boldsymbol{\delta}) \boldsymbol{X}_I] \geq cn,$$

where $\boldsymbol{X}_I$ consists of the first $s + d$ columns of the design matrix $\boldsymbol{X}$. With model (1.1), this condition becomes

$$\lambda_{\min}[\boldsymbol{X}_I^T \boldsymbol{X}_I] \geq cn,$$

which is

$$\lambda_{\min}[(1/n)\boldsymbol{X}_I^T \boldsymbol{X}_I] = \lambda_{\min}[C_{11}^n] \geq c,$$

where C_{11}^n is the matrix defined in Zhao & Yu (2006) and c is a positive constant. Since the minimal eigenvalue $\lambda_{\min}[C_{11}^n]$ converges to 0, the above condition does not hold. Thus, the consistency result Theorem 3 of Fan et al. (2011) is not applicable for model (1.1).

B Supplement to Section 2

In this supplement, we provide two lemmas and one proposition with their proofs and two graphs Figures 1 and 2 illustrating the incidental parameters and the step of updating the responses in the iteration algorithm with $d = 1$.

Lemma B.1. *A necessary and sufficient condition for* $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ *to be a minimizer of* $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ *is that*

$$\begin{aligned} &\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}), \\ &Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}} = \lambda Sign(\hat{\mu}_i), \quad \textit{for } i \in \hat{I}_0^c, \\ &|Y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}| \leq \lambda, \quad \textit{for } i \in \hat{I}_0, \end{aligned}$$

where $Sign(\cdot)$ *is a sign function and* $\hat{I}_0 = \{1 \leq i \leq n : \hat{\mu}_i = 0\}$.

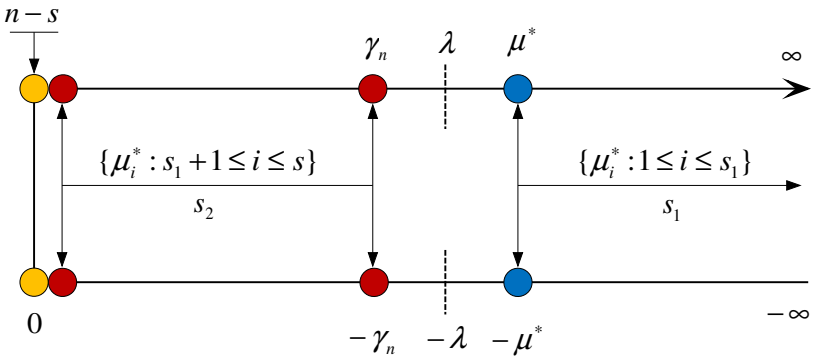

Figure 1: An illustration of three types of $\mu_i^\star$'s, that is, large $\boldsymbol{\mu}_1^\star$, bounded $\boldsymbol{\mu}_2^\star$ and zero $\boldsymbol{\mu}_3^\star$. The negative half of the real line is folded at 0 under the positive half for convenience. For the penalized least square method with a soft penalty function and under the assumption of fixed d, the specification of the regularization parameter λ is that $\kappa_n \ll \lambda$, $\alpha\gamma_n \leq \lambda$, and $\lambda \ll \min\{\mu^\star, \sqrt{n}\}$.

Proof of Lemma B.1. By subdifferential calculus (see, for example, Theorem 3.27 in Jahn (2007)), a necessary and sufficient condition for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ to be a minimizer of $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ is that zero is in the subdifferential of L at $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$, which means that, for each i,

$$\begin{aligned}
&\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}),\\
&Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} = \lambda \mathrm{Sign}(\hat{\mu}_i), \quad \text{if } \hat{\mu}_i \neq 0,\\
&|Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}| \leq \lambda, \quad \text{if } \hat{\mu}_i = 0.
\end{aligned}$$

Thus, the conclusion of Lemma B.1 follows. □

Proposition B.1. *Suppose Assumptions (A) and (B) hold and there exist positive constants C_1 and C_2 such that $\|\boldsymbol{\beta}^\star\|_2 < C_1$ and $\|\boldsymbol{\beta}^{(0)}\|_2 < C_2$ wpg1. If $s_1\lambda/n = O(1)$ and $s_2\gamma_n/n = o(1)$, then, for every $K \geq 1$ and $k \leq K$, wpg1 as $n \to \infty$,*

$$\|\boldsymbol{\beta}^{(K+1)} - \boldsymbol{\beta}^{(K)}\|_2 \leq O((s_1/n)^K), \quad \text{and} \quad \|\boldsymbol{\beta}^{(k)}\|_2 \leq 2\sqrt{d}C_1 + C_2.$$

Remark B.1. For any prespecified critical value in the stopping rule, Proposition B.1 implies that the algorithm stops at the second iteration wpg1. In practice, the sample size n might not be large enough for the two-iteration estimator to have a decent performance so that more iterations are usually needed to activate the stopping rule. By Proposition B.1, K iterations will make the distance $||\boldsymbol{\beta}^{(K+1)} - \boldsymbol{\beta}^{(K)}||_2$ of the small order $(s_1/n)^K$. When s_1/n is small, the algorithm converges quickly, which has been verified by our simulations.

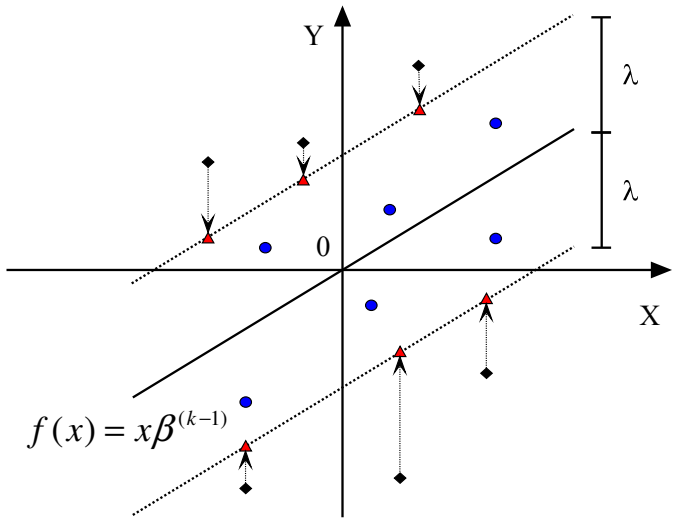

Figure 2: An illustration for the updating of responses with $d = 1$. The solid black line is a fitted regression line. The dashed black lines are the corresponding shifted regression lines. The circle and diamond points are the original data points. The circle and triangle points are the updated data points. That is, the diamond points are drawn onto the shifted regression lines.

Proof of Proposition B.1. First, we show that, wpg1, $\|\boldsymbol{\beta}^{(1)}\|_2$ is bounded by $2\sqrt{d}C_1+C_2$. For each $k \geq 1$, we have

$$\mathbb{S}\boldsymbol{\beta}^{(k)} = \mathbb{S}^{\mu}_{S_{11}} + \mathbb{S}^{\mu}_{S_{12}} + \mathbb{S}_{S_1}\boldsymbol{\beta}^{\star} + \mathbb{S}^{\epsilon}_{S_1} + \mathbb{S}_{S_2 \cup S_3}\boldsymbol{\beta}^{(k-1)} + \lambda(\mathcal{S}_{S_2} - \mathcal{S}_{S_3}),$$

where $S_i = \cup_{j=1}^{3} S_{ij}(\boldsymbol{\beta}^{(k-1)})$ for $i = 1, 2, 3$ and S_{ij}'s are defined at the end of Section 2. Denote $\mathcal{A}_{k-1}$ as the intersection of the events $\{S_{11}(\boldsymbol{\beta}^{(k-1)}) = \emptyset\}$, $\{S_{12}(\boldsymbol{\beta}^{(k-1)}) = S^{\star}_{12}\}$, $\{S_1(\boldsymbol{\beta}^{(k-1)}) = S^{\star}_{10} \cup S^{\star}_{12}\}$, $\{S_2(\boldsymbol{\beta}^{(k-1)}) = S^{\star}_{21}\}$ and $\{S_3(\boldsymbol{\beta}^{(k-1)}) = S^{\star}_{31}\}$, where $S^{\star}_{ij}$'s are defined at the beginning of Section 3.

By Lemma 1 in the paper, $P(\mathcal{A}_0) \to 1$. Thus, wpg1,

$$\boldsymbol{\beta}^{(1)} = T_0^{-1}T_1 + T_0^{-1}T_2 + T_0^{-1}T_3 + T_0^{-1}T_4(\boldsymbol{\beta}^{(0)}) + T_0^{-1}T_5,$$

where $T_0 = \mathbb{S}/n$, $T_1 = \mathbb{S}^{\mu}_{S^{\star}_{12}}/n$, $T_2 = \mathbb{S}_{s_1+1,n}\boldsymbol{\beta}^{\star}/n$, $T_3 = \mathbb{S}^{\epsilon}_{s_1+1,n}/n$, $T_4(\boldsymbol{\beta}^{(0)}) = \mathbb{S}_{1,s_1}\boldsymbol{\beta}^{(0)}/n$ and $T_5 = (\mathbb{S}_{S^{\star}_{21}} - \mathbb{S}_{S^{\star}_{31}})\lambda/n$. We will show that, wpg1, $\|T_0^{-1}T_1\|_2 \leq C_2/4$, $\|T_0^{-1}T_2\|_2 \leq 2\sqrt{d}C_1$, $\|T_0^{-1}T_3\|_2 \leq C_2/4$, $\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 \leq C_2/4$ and $\|T_0^{-1}T_5\|_2 \leq C_2/4$. Then, wpg1,

$$\|\boldsymbol{\beta}^{(1)}\|_2 \leq \sum_{i=1}^{5} \|T_0^{-1}T_i\|_2 \leq 2\sqrt{d}C_1 + C_2.$$

On $T_0^{-1}T_1$. For $s_2\gamma_n/n = o(1)$, wpg1,

$$\|T_0^{-1}T_1\|_2 \leq \left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_F \|\frac{1}{n}\mathbb{S}^{\mu}_{S^{\star}_{12}}\|_2 \leq 4\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_F \mathbb{E}\|\boldsymbol{X}_0\|_2 \frac{s_2}{n}\gamma_n \to 0.$$

Thus, wpg1, $\|T_0^{-1}T_1\|_2 \leq C_2/4$.

On $T_0^{-1}T_2$. Wpg1,

$$\|T_0^{-1}T_2\|_2 \leq \left\|(\frac{1}{n}\mathbb{S})^{-1}\frac{1}{n}\mathbb{S}_{s_1+1,n}\right\|_F \|\boldsymbol{\beta}^{\star}\|_2 \leq 2\left\|\boldsymbol{I}_d\right\|_F C_1 = 2\sqrt{d}C_1.$$

On $T_0^{-1}T_3$. Wpg1,

$$\|T_0^{-1}T_3\|_2 \leq 2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_F \|\frac{1}{n}\mathbb{S}^{\epsilon}_{s_1+1,n}\|_2 \xrightarrow{P} 0.$$

Thus, wpg1, $\|T_0^{-1}T_3\|_2 \leq C_2/4$.

On $T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})$. For $s_1/n = o(1)$,

$$\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 \leq \frac{s_1}{n}\left\|(\frac{1}{n}\mathbb{S})^{-1}\frac{1}{s_1}\mathbb{S}_{1,s_1}\right\|_F \|\boldsymbol{\beta}^{(0)}\|_2 \leq \frac{s_1}{n}2\sqrt{d}C_2 \xrightarrow{P} 0.$$

Thus, wpg1, $\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 \leq C_2/4$.

On $T_0^{-1}T_5$. For $s_1\lambda/n = O(1)$, wpg1,

$$\|T_0^{-1}T_5\|_2 \leq 2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_F \frac{s_1\lambda}{n}(\|\frac{1}{s_1}\mathcal{S}_{S^{\star}_{21}}\|_2 + \|\frac{1}{s_1}\mathcal{S}_{S^{\star}_{31}}\|_2) \xrightarrow{P} 0.$$

Thus, wpg1, $\|T_0^{-1}T_5\|_2 \leq C_2/4$.

Next, consider $\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|_2$. Since $\boldsymbol{\beta}^{(1)}$ is bounded wpg1, by Lemma 1 in the paper, $\mathcal{A}_1$ occurs wpg1. Then,

$$\boldsymbol{\beta}^{(2)} = T_0^{-1}T_1 + T_0^{-1}T_2 + T_0^{-1}T_3 + T_0^{-1}T_4(\boldsymbol{\beta}^{(1)}) + T_0^{-1}T_5,$$

where $T_4(\boldsymbol{\beta}^{(1)}) = (1/n)\mathbb{S}_{1,s_1}\boldsymbol{\beta}^{(1)}$. Thus, wpg1,

$$\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)} = \mathbb{S}^{-1}\mathbb{S}_{1,s_1}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}).$$

It follows that, for $s_1 = o(n)$, wpg1,

$$\|\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)}\|_2 \leq \left\|\mathbb{S}^{-1}\mathbb{S}_{1,s_1}\right\|_F \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}\|_2 \leq (2\sqrt{d}s_1/n)(4\sqrt{d}C_1 + 2C_2) \to 0.$$

Then, wpg1, $\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)}$, which means that, wpg1, the iteration algorithm stops at the second iteration.

Finally, for any $K \geq 1$, repeat the above arguments. Then, with at least probability $p_{n,K} = P(\bigcap_{k=0}^{K}\mathcal{A}_k)$, which increases to one by Lemma 1 in the paper, we have

$$\|\boldsymbol{\beta}^{(K+1)} - \boldsymbol{\beta}^{(K)}\|_2 \leq (2\sqrt{d}s_1/n)^K(4\sqrt{d}C_1 + 2C_2) = O((s_1/n)^K) \to 0,$$

and $\|\boldsymbol{\beta}^{(k)}\|_2 \leq 2\sqrt{d}C_1 + C_2$ for all $k \leq K$. □

Lemma B.2. *A necessary and sufficient condition for* $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ *to be a minimizer of* $L(\boldsymbol{\mu}, \boldsymbol{\beta})$ *is that it is a solution to equations (2.5) and (2.6).*

Proof of Lemma B.2. First, we show a solution of (2.5) and (2.6) satisfies the necessary and sufficient condition in Lemma B.1. Denote a solution of (2.5) and (2.6) as $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$. Then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, which is exactly the first condition in Lemma B.1, and, for each $i = 1, 2, \ldots, n$, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies one of three cases: $|Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}| \leq \lambda$ and $\hat{\mu}_i = 0$; $Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} > \lambda$ and $\hat{\mu}_i = Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} - \lambda$; $Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} < -\lambda$ and $\hat{\mu}_i = Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} + \lambda$. If $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the first case, it satisfies the third condition in Lemma B.1. If $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the second case, then $\hat{\mu}_i > 0$ and $Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} = \lambda = \lambda\text{Sign}(\hat{\mu}_i)$, which means that the second case satisfies the second condition in Lemma B.1. Similarly, the third case also satisfies the second condition in Lemma B.1. Thus $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the necessary and sufficient condition in Lemma B.1.

In the other direction, suppose $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the necessary and sufficient condition in Lemma B.1. Then, the first condition in Lemma B.1 exactly (2.5). For each i, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies one of three cases: $\hat{\mu}_i = 0$ and $|Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}| \leq \lambda$; $\hat{\mu}_i > 0$ and $Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} = \lambda$; $\hat{\mu}_i < 0$ and $Y_i - \hat{\mu}_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} = -\lambda$. If $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the first case, it satisfies the first case in (2.6). If $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the second case, then $\hat{\mu}_i = Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} - \lambda$ and $Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}} > \lambda$, which means that $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the second case of (2.6). Similarly, If $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies the third case, then it satisfies the third case of (2.6). Thus, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ satisfies (2.5) and (2.6). □

C Supplement to Section 3

In this supplement, we provide the proofs of the theoretical results in Section 3. Before that, we point out that those two different sufficient conditions in Theorem 1 in the paper come from the different analysis on the term $\mathbb{S}^{\mu}_{S^{\star}_{12}}$. Each of the two different sufficient conditions does not imply the other. Specifically, on one hand, suppose the absolute values of $\mu_i^{\star}$'s are all equal for $i = s_1 + 1, s_2 + 2, \ldots, s$. Then, $\|\boldsymbol{\mu}_2^{\star}\|_2^{2+\delta} = s_2^{(2+\delta)/2}|\mu_s^{\star}|^{2+\delta}$ and $\sum_{i=s_1+1}^{s}|\mu_i^{\star}|^{2+\delta} = s_2|\mu_s^{\star}|^{2+\delta}$. Thus Assumption (C) holds automatically since $s_2 \to \infty$. This means that Assumption (C) holds at least when the absolute magnitudes of $\mu_i^{\star}$'s are similar to each other. For this case, there still exists a consistent estimator even if $n/(\kappa_n\gamma_n) \ll s_2 \ll n$. On the other hand, suppose $\mu_s^{\star} = \gamma_n$ and the other $\mu_i^{\star}$'s are all equal to a constant $c > 0$. Then, $\|\boldsymbol{\mu}_2^{\star}\|_2^{2+\delta} = [\gamma_n^2 + (s_2 - 1)c^2]^{(2+\delta)/2}$ and $\sum_{i=s_1+1}^{s}|\mu_i^{\star}|^{2+\delta} = \gamma_n^{2+\delta} + (s_2 - 1)c^{2+\delta}$. If $s_2 \ll \gamma_n^2 \ll n/(\kappa_n\gamma_n)$, the previous two terms are both asymptotically equivalent to $\gamma_n^{2+\delta}$. Thus Assumption (C) fails but the other sufficient condition holds.

Proof of Lemma 1 in the paper. The proof is the similar to that of Lemma D.1 and omitted. □

Proof of Theorem 1 in the paper. By Lemma 1 in the paper, wpg1, the solution $\hat{\boldsymbol{\beta}}_n$ to $\varphi_n(\boldsymbol{\beta}) = 0$ on $\mathcal{B}_C(\boldsymbol{\beta}^\star)$ is explicitly given by

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^\star + T_0^{-1}(T_1 + T_2 + T_3 - T_4),$$

where $T_0 = (1/n)\mathbb{S}_{s_1+1,n}$, $T_1 = (1/n)\mathbb{S}^{\mu}_{S^\star_{12}}$, $T_2 = (1/n)\mathbb{S}^{\epsilon}_{s_1+1,n}$, $T_3 = (\lambda/n)\mathcal{S}_{S^\star_{21}}$ and $T_4 = (\lambda/n)\mathcal{S}_{S^\star_{31}}$. We will show that $T_0 \xrightarrow{P} \boldsymbol{\Sigma}_X^{-1} > 0$ with the Frobenius norm and $T_i \xrightarrow{P} 0$ with the Euclidean norm for $i = 1, 2, 3, 4$. Thus, by Slutsky's lemma (see, for example, Lemma 2.8 on page 11 of van der Vaart (1998)), $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}^\star$.

On T_0^{-1}. By law of large number, $T_0 \xrightarrow{P} \boldsymbol{\Sigma}_X > 0$. Then, by continuous mapping theorem, $T_0^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_X^{-1} > 0$.

On T_1: Approach One. Suppose $s_2 = o(n/(\kappa_n\gamma_n))$. Then,

$$\|T_1\|_2 \le \frac{1}{n}\sum_{i=s_1+1}^{s} \|\boldsymbol{X}_i\mu_i^\star\|_2 = \frac{1}{n}\sum_{i=s_1+1}^{s} \|\boldsymbol{X}_i\|_2 \cdot |\mu_i^\star| \le s_2\kappa_n\gamma_n/n = o(1).$$

On T_1: Approach Two. Under Assumption (C), it follows

$$\left(\boldsymbol{\Sigma}_X \sum_{i=s_1+1}^{s} \mu_i^{\star 2}\right)^{-1/2} \mathbb{S}^{\mu}_{S^\star_{12}} \xrightarrow{d} N(0, I_d).$$

In fact, Assumption (C) implies the Lyapunov condition for sequence of random vectors (see, e.g. Proposition 2.27 on page 332 of (van der Vaart 1998)). More specifically, recall the Lyapunov condition is that there exists some constant $\delta > 0$ such that

$$\sum_{i=s_1+1}^{s} \mathbb{E}\|(\boldsymbol{\Sigma}_X \sum_{j=s_1+1}^{s} \mu_j^{\star 2})^{-1/2}\boldsymbol{X}_i\mu_i^\star\|_2^{2+\delta} \to 0.$$

Then, by Assumption (C),

$$\begin{aligned}
&\sum_{i=s_1+1}^{s} \mathbb{E}\|(\boldsymbol{\Sigma}_X \sum_{j=s_1+1}^{s} \mu_j^{\star 2})^{-\frac{1}{2}}\boldsymbol{X}_i\mu_i^\star\|_2^{2+\delta} \\
&\le (\sum_{j=s_1+1}^{s} \mu_j^{\star 2})^{-\frac{2+\delta}{2}} \sum_{i=s_1+1}^{s} |\mu_i^\star|^{2+\delta}\lambda_{\min}^{-\frac{2+\delta}{2}}\mathbb{E}\|X_0\|_2^{2+\delta} \longrightarrow 0,
\end{aligned}$$

where $\lambda_{\min} > 0$ is the minimum eigenvalue of $\boldsymbol{\Sigma}_X$. Then,

$$\begin{aligned}
\|T_1\|_2 &= \|\frac{1}{n}\mathbb{S}^{\mu}_{S^\star_{12}}\|_2 \le \frac{1}{n}\left\|(\boldsymbol{\Sigma}_X \sum_{i=s_1+1}^{s} \mu_i^{\star 2})^{1/2}\right\|_F \|(\boldsymbol{\Sigma}_X \sum_{i=s_1+1}^{s} \mu_i^{\star 2})^{-1/2}\mathbb{S}^{\mu}_{S^\star_{12}}\|_2 \\
&= \frac{1}{n}(\sum_{i=s_1+1}^{s} \mu_i^{\star 2})^{1/2}\left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_F O_P(1) \le \frac{1}{n}(s_2\gamma_n^2)^{1/2}O_P(1) \le \frac{1}{\sqrt{n}}\gamma_n O_P(1) = o_P(1).
\end{aligned}$$

On T_2. By law of large number, $T_2 = o_P(1)$.
On T_3 and T_4. By noting $\lambda \ll \sqrt{n}$,

$$\|T_3\|_2 = \|\lambda \frac{1}{n} \mathcal{S}_{S_{21}^\star}\|_2 = \lambda \frac{\sqrt{s_1}}{n} \| \frac{1}{\sqrt{s_1}} \mathcal{S}_{S_{21}^\star}\|_2 \leq \frac{\lambda}{\sqrt{n}} O_P(1) = o_P(1).$$

Thus $T_3 = o_P(1)$. In the same way, we can show that $T_4 = o_P(1)$ holds. □

In Theorem 2 of the paper, one condition is $D_n/n = o(1)$. In fact, we can consider other conditions on D_n and derive more possible asymptotic distributions for $\hat{\boldsymbol{\beta}}_n$.

Theorem C.1 (Asymptotic Distributions on $\hat{\boldsymbol{\beta}}_n$: more cases)**.** *Under Assumptions (A), (B) and (C), for all constants $b, c \in \mathbb{R}^+$,*

(1) when $s_1 \ll n/\lambda^2$ and $D_n^2/n = o(1)$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{\Sigma}_X^{-1})$; **[main case]**
(2) when $s_1 \ll n/\lambda^2$ and $D_n^2/n \sim c$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (c+\sigma^2) \boldsymbol{\Sigma}_X^{-1})$;
(3) when $s_1 \ll n/\lambda^2$ and $D_n^2/n \to \infty$, $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$, where $r_n \sim n/D_n \ll \sqrt{n}$;
(4) when $s_1 \sim bn/\lambda^2$ and $D_n^2/n = o(1)$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (b+\sigma^2) \boldsymbol{\Sigma}_X^{-1})$;
(5) when $s_1 \sim bn/\lambda^2$ and $D_n^2/n \sim c$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (b+c+\sigma^2) \boldsymbol{\Sigma}_X^{-1})$;
(6) when $s_1 \sim bn/\lambda^2$ and $D_n^2/n \to \infty$, $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$, where $r_n \sim n/D_n \ll \sqrt{n}$;
(7) when $s_1 \gg n/\lambda^2$ and $D_n^2/n = o(1)$ or $D_n^2/n \sim c$, $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$, where $r_n \sim n/(\lambda\sqrt{s_1}) \ll \sqrt{n}$;
(8) when $s_1 \gg n/\lambda^2$ and $D_n^2/n \to \infty$, letting $r_n \sim \min\{\sqrt{b}n/(\lambda\sqrt{s_1}), n/D_n\} \ll \sqrt{n}$,

(8a) if $\sqrt{b}n/(\lambda\sqrt{s_1}) \gg n/D_n$, then $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$;
(8b) if $\sqrt{b}n/(\lambda\sqrt{s_1}) \sim n/D_n$, then $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (1+b) \boldsymbol{\Sigma}_X^{-1})$;
(8c) if $\sqrt{b}n/(\lambda\sqrt{s_1}) \ll n/D_n$, then $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, b\boldsymbol{\Sigma}_X^{-1})$.

Theorem 2 in the paper groups the results according to the asymptotic magnitude of s_1 given an upper bound of the diverging speed of s_2. Alternatively, Theorem C.1 groups the results according to the asymptotic magnitudes of s_1 and D_n^2. Since both s_1 and D_n^2 have three cases, Theorem C.1 basically contains nine cases. For the last case, there are further three cases on the relationship between $\sqrt{b}n/(\lambda\sqrt{s_1})$ and n/D_n. As in Theorem 2 in the paper, the first case of Theorem C.1 is denoted as the *main case* since for this case the incidental parameters are sparse in the sense that the size and magnitude of the nonzero incidental parameters $\boldsymbol{\mu}_1^\star$ and $\boldsymbol{\mu}_2^\star$ are well controlled. Note that $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$ implies $D_n^2/n = o(1)$. which means that, under Assumption (C), the cases *(1)*, *(4)* and *(7)* of Theorem C.1 actually imply the three results of Theorem 2 in the paper. As in Theorem 2 in the paper, the convergence rate of $\hat{\boldsymbol{\beta}}_n$ becomes less than $\sqrt{n}$ when $s_1 \gg n/\lambda^2$ or $D_n^2/n \to \infty$, that is, when the size and magnitude of the nonzero incidental parameters are large; the boundary phenomenon also appears.

Proof of Theorems 2 in the paper and C.1. It is sufficient to provide the proof for the case where the sizes of index sets $S_{21}^\star = \{1 \le i \le s_1 : \mu_i^\star > 0\}$ and $S_{31}^\star = \{1 \le i \le s_1 : \mu_i^\star < 0\}$ are both asymptotically $s_1/2$ and $b = 2$.

From the proof of Theorem 1 in the paper, wpg1,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^\star + \mathbb{S}^{-1}_{s_1+1,n}[\mathbb{S}^\mu_{S_{12}^\star} + \mathbb{S}^\epsilon_{s_1+1,n} + \lambda(\mathcal{S}_{S_{21}^\star} - \mathcal{S}_{S_{31}^\star})].$$

Let r_n be a sequence going to infinity. Then, $r_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) = T_0^{-1}(V_1 + V_2 + V_3 - V_4)$, where $V_1 = r_n T_1$, $V_2 = r_n T_2$, $V_3 = r_n T_3$, $V_4 = r_n T_4$ and T_i's are defined in the proof of Theorem 1 in the paper. Next we derive the asymptotic properties of T_0 and V_i's, from which the desired results follow by Slutsky's lemma.

On T_0. By the proof of Theorem 1 in the paper, $T_0^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_X^{-1}$

On V_1: Approach One. If $r_n = \sqrt{n}$ and $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$, then

$$\|T_1\|_2 = \|r_n \frac{1}{n}\mathbb{S}^\mu_{S_{12}^\star}\|_2 \le r_n \frac{1}{n}\sum_{i=s_1+1}^{s} \|\boldsymbol{X}_i\|_2 \cdot |\mu_i^\star| \le r_n \frac{1}{n} s_2\kappa_n\gamma_n = \frac{1}{\sqrt{n}} s_2\kappa_n\gamma_n = o(1).$$

Thus, if $r_n = \sqrt{n}$ or $r_n \ll \sqrt{n}$ and $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$, then $T_1 = o_P(1)$.

On V_1: Approach Two. If $r_n = \sqrt{n}$, then

$$T_1 = r_n \frac{1}{n}\mathbb{S}^\mu_{S_{12}^\star} = r_n \frac{D_n}{n}\frac{1}{D_n}\mathbb{S}^\mu_{S_{12}^\star} = \frac{D_n}{\sqrt{n}}\frac{1}{D_n}\mathbb{S}^\mu_{S_{12}^\star},$$

where $D_n = \|\boldsymbol{\mu}_2^\star\|_2 = (\sum_{i=s_1+1}^{s} \mu_i^{\star 2})^{1/2}$. There are three cases on $D_n/\sqrt{n}$ or D_n^2/n. If $D_n^2/n \to 0$, then $T_1 \xrightarrow{P} 0$. If $D_n^2/n \to 1$, then $T_1 \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X)$. If $D_n^2/n \to \infty$, it means that $r_n = \sqrt{n}$ is too fast. Let $r_n \sim n/D_n = \sqrt{n}\sqrt{n/D_n^2} \ll \sqrt{n}$. Then $T_1 \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X)$;

On V_2. If $r_n = \sqrt{n}$, then $T_2 \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X)$. Thus, if $r_n \ll \sqrt{n}$, $T_2 \xrightarrow{P} 0$; if $r_n \gg \sqrt{n}$; $T_2 \xrightarrow{P} \infty$.

On V_3 and V_4. First consider T_3. Denote $\#(\cdot)$ as the size function. If $r_n = \sqrt{n}$, then

$$T_3 = \lambda r_n \frac{1}{n}\mathcal{S}_{S_{21}^\star} = \lambda\sqrt{\frac{s_1/2}{n}}\frac{1}{\sqrt{\#(S_{21}^\star)}}\mathcal{S}_{S_{21}^\star}.$$

Note that $\#(S_{21}^\star) = s_1/2$. There are three cases on $\lambda\sqrt{s_1/(2n)}$. If $\lambda\sqrt{s_1/(2n)} \to 0$, then $T_3 \xrightarrow{P} 0$. Note that $\lambda\sqrt{s_1/(2n)} \to 0$ is equivalent to $s_1 = o(2n/\lambda^2)$. If $\lambda\sqrt{s_1/(2n)} \to 1$, then $T_3 \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X)$. Note that $\lambda\sqrt{s_1/(2n)} \to 1$ is equivalent to $s_1 \sim 2n/\lambda^2$. If $\lambda\sqrt{s_1/(2n)} \to \infty$, it means $r_n = \sqrt{n}$ is too large. Let $r_n \sim n/(\lambda\sqrt{(s_1/2)}) = \sqrt{n}\sqrt{2n}/(\lambda\sqrt{s_1}) \ll \sqrt{n}$. With this rate r_n, $T_3 \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X)$. Note that $\lambda\sqrt{s_1/2n} \to \infty$ is equivalent to $s_1 \gg O(2n/\lambda^2)$. In the same way, T_4 can be analyzed and parallel results can be obtained. □

Proof of Theorem 3 in the paper. The proof is similar to that of Theorem 7 in the paper and omitted. □

C.1 Supplement for Subsection 3.1

The following Theorem implies Theorem 4 in the paper since it contains more details.

Theorem C.2 (Consistency and Asymptotic Normality on $\tilde{\boldsymbol{\beta}}$)**.** *Suppose Assumptions (A) and (B) hold. If either $s_2 = o(n/(\kappa_n\gamma_n))$ or Assumption (C) holds, then $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$. If $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$, then $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$. On the other hand, under Assumption (C),*

(1) if $D_n^2/n = o(1)$, then $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{\Sigma}_X^{-1})$; **[main case]**

(2) if $D_n^2/n \sim c$, then $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, (c+\sigma^2)\boldsymbol{\Sigma}_X^{-1})$, for every constant $c \in \mathbb{R}^+$;

(3) if $D_n^2/n \to \infty$, then $r_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_X^{-1})$ where $r_n \sim n/D_n \ll \sqrt{n}$.

Proof of Theorem C.2. Denote $I_0 = \{s_1+1, s_1+2, \ldots, s = s_1+s_2, s+1, \ldots, n\}$. Note that $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$ ensures that $\hat{\boldsymbol{\beta}}$ is consistent by Theorem 1 in the paper. By Theorem 3 in the paper, $P\{\hat{I}_0 = I_0\}$ goes to 1. Then,

$$\tilde{\boldsymbol{\beta}} = R_1 + R_2 + T_0^{-1}(T_1 + T_2),$$

where $R_1 = (\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{X}_{\hat{I}_0})^{-1}\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{Y}_{\hat{I}_0}\{\hat{I}_0 \neq I_0\}$ and $R_2 = -(\boldsymbol{X}_{I_0}^T\boldsymbol{X}_{I_0})^{-1}\boldsymbol{X}_{I_0}^T\boldsymbol{Y}_{I_0}\{\hat{I}_0 \neq I_0\}$ and T_i's are defined in the proof of Theorem 1 in the paper. The proof for the consistency is similar to that of Theorem 1 in the paper and is omitted. Next we show the asymptotic normality. We have,

$$r_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) = r_nR_1 + r_nR_2 + T_0^{-1}(V_1 + V_2),$$

where V_i's are defined in the proof of Theorem C.1. Since $P(\sqrt{n}R_1 = 0) \geq P\{\hat{I}_0 = I_0\} \to 1$, we have $\sqrt{n}R_1 = o_P(1)$. Similarly, $\sqrt{n}R_2 = o_P(1)$. From the analysis on V_i's in the proof of Theorem C.1, the asymptotic distributions follows by Slutsky's lemma. □

Lemma C.1 (Consistency on $\hat{\sigma}$)**.** *Suppose Assumptions (A) and (B) hold and either $s_2 = o(n/(\kappa_n\gamma_n))$ or Assumption (C) holds. If $s_2 = o(n/\gamma_n^2)$, then $\hat{\sigma} \xrightarrow{P} \sigma$.*

Proof of Lemma C.1. When Assumption (C) or $s_2 = o(n/(\kappa_n\gamma_n))$ holds, the penalized estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent estimators of $\boldsymbol{\beta}^\star$ by Theorem 1 in the paper and 4 in the paper. Denote $\mathcal{C} = \{\hat{I}_0 = I_0\}$. By Theorem 3 in the paper, $\mathcal{C}$ occurs wpg1. Then, $\hat{\sigma}^2 = T\mathcal{C} + \hat{\sigma}^2\mathcal{C}^c$, where $T = a_n\|\boldsymbol{Y}_{I_0} - \boldsymbol{X}_{I_0}^T\tilde{\boldsymbol{\beta}}\|_2^2$ and $a_n = 1/(n-s_1)$. It is sufficient to show $T \xrightarrow{P} \sigma^2$. We have $T = \sum_{i=1}^6 T_i$, where $T_1 = a_n\sum_{i=s_1+1}^n[\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})]^2$, $T_2 = a_n\sum_{i=s_1+1}^n\epsilon_i^2$, $T_3 = 2a_n\sum_{i=s_1+1}^n\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})\epsilon_i$, $T_4 = a_n\sum_{i=s_1+1}^s\mu_i^{\star 2}$, $T_5 = 2a_n\sum_{i=s_1+1}^s\mu_i\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})$ and $T_6 = 2a_n\sum_{i=s_1+1}^s\mu_i^\star\epsilon_i$. It is straightforward to show that $T_2 \xrightarrow{P} \sigma^2$ and each other $T_i \xrightarrow{P} 0$ under the condition $s_2 = o(n/\gamma_n^2)$ and by noting that $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$. Then $\hat{\sigma}$ is a consistent estimator of σ. □

C.2 Supplement for Subsection 3.2

In this supplement, we consider a special case with exponentially tailed covariates and errors. For convenience, we first introduce the definition of Orlicz norm and related inequalities. For a strictly increasing and convex function ψ with $\psi(0) = 0$, the Orlicz norm of a random variable Z with respect to ψ is defined as

$$\|Z\|_\psi = \inf\{C > 0 : \mathbb{E}\psi(|Z|/C) \le 1\}.$$

Then, for each $x > 0$,

$$P(|Z| > x) \le 1/\psi(x/\|Z\|_\psi). \tag{1}$$

(See Page 96 of van der Vaart & Wellner (1996)). Next, we introduce a lemma on Orlicz norm with ψ_1. Suppose $\{Z_i\}_{i=1}^n$ is a sequence of random variables and $\{\boldsymbol{Z}_i\}_{i=1}^n$ is a sequence of d-dimensional random vectors with $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{id})^T$. From Lemma 8.3 on Page 131 of Kosorok (2008), we have the following extension.

Lemma C.2. *If for each $1 \le i \le n$ and $1 \le j \le d$,*

$$P(|Z_i| > x) \le c\exp\{-\frac{1}{2}\cdot\frac{x^2}{ax+b}\} \text{ and } P(|Z_{ij}| > x) \le c\exp\{-\frac{1}{2}\cdot\frac{x^2}{ax+b}\},$$

with $a, b \ge 0$ and $c > 0$, then

$$\|\max_{1\le i\le n}|Z_i|\|_{\psi_1} \le K\{a(1+c)\log(1+n) + \sqrt{b(1+c)}\sqrt{\log(1+n)}\},$$
$$\|\max_{1\le i\le n}\|\boldsymbol{Z}_i\|_2\|_{\psi_1} \le K\{a\sqrt{d}(1+cd)\log(1+n) + \sqrt{bd(1+cd)}\sqrt{\log(1+n)}\}.$$

where K is a universal constant which is independent of a, b, c, $\{Z_i\}$ and $\{\boldsymbol{Z}_i\}$.

Proof of Lemma C.2. The proof for random variables $\{Z_i\}$ is the same to the proof of Lemma 8.3 on Page 131 of Kosorok (2008). For random vectors $\{\boldsymbol{Z}_i\}$,

$$P(\|\boldsymbol{Z}_i\|_2 \ge x) \le P(\max_{1\le j\le d}|Z_{ij}| > x/\sqrt{d}) \le \sum_{j=1}^d P(|Z_{ij}| > x/\sqrt{d}) \le c'\exp\{-\frac{1}{2}\frac{x^2}{a'x+b'}\},$$

where $a' = a\sqrt{d}$, $b' = bd$ and $c' = cd$. Then, by the result on random variables, the desired result on random vectors follows. □

Now, suppose, for every $x > 0$,

$$P(|\epsilon_i| > x) \le c_1\exp\{-\frac{1}{2}\cdot\frac{x^2}{a_1x+b_1}\} \text{ and } P(|X_{ij}| > x) \le c_2\exp\{-\frac{1}{2}\cdot\frac{x^2}{a_2x+b_2}\}, \tag{2}$$

with $a_i, b_i \ge 0$ and $c_i > 0$ for $i = 1, 2$. By Lemma C.2, it follows

$$\|\max_{1\le i\le n}|\epsilon_i|\|_{\psi_1} \le K\{a_1(1+c_2)\log(1+n) + \sqrt{b_1(1+c_1)}\sqrt{\log(1+n)}\},$$
$$\|\max_{1\le i\le n}\|\boldsymbol{X}_i\|_2\|_{\psi_1} \le K\{a_2\sqrt{d}(1+c_2d)\log(1+n) + \sqrt{b_2d(1+c_2d)}\sqrt{\log(1+n)}\}.$$

Thus, from the inequality (1), if $a_1 > 0$, let $\gamma_n \gg \log(n)$; otherwise, let $\gamma_n \gg \sqrt{\log(n)}$. Similarly, if $a_2 > 0$, let $\kappa_n \gg \log(n)$; otherwise, let $\kappa_n \gg \sqrt{\log(n)}$. Then, such γ_n and κ_n satisfy the condition (2.2). Suppose both a_1 and a_2 are positive, which means both ϵ_i and X_{ij}'s have exponential tails. As before, set $\kappa_n = \gamma_n = \log(n)\tau_n$. For this case, the regularization parameter specification (2.4) becomes $\log(n)\tau_n \ll \lambda \ll \min\{\mu^\star, \sqrt{n}\}$.

At the end of this supplement, we simply list explicit expressions of κ_n under different assumptions on the covariates for the case with a diverging number of covariates, which are the extension of the results in Section 3.2. The magnitude of κ_n becomes larger than that for the case with d fixed while γ_n keeps the same. Specifically, if $\boldsymbol{X}_0$ is bounded with $C_X > 0$, then $\kappa_n = \sqrt{d}C_X$. If $\boldsymbol{X}_0$ follows a Gaussian distribution $N(0, \boldsymbol{\Sigma}_X)$, then $\kappa_n = \sqrt{2d\sigma_X^2[(3/2)\log(d) + \log(n)]}$. If the Orlicz norm $\|X_{0j}\|_\psi$ exists for $1 \le j \le d$ and their average $(1/d)\sum_{j=1}^d \|X_{0j}\|_\psi$ is bounded, then $\kappa_n \gg d\psi^{-1}(n)$; for instance, if $\psi = \psi_p$ with $p \ge 1$, then $\kappa_n \gg d(\log(n))^{1/p}$. Finally, if the data $\{\boldsymbol{X}_i\}$ satisfies the right inequality of (2) with $a_2 > 0$, that is, each component of $\boldsymbol{X}_i$ is sub-exponentially tailed, then $\kappa_n \gg d^{3/2}\log(n)$. It is worthwhile to note that these expressions of κ_n depend on a factor involving the diverging number of covariates d, which will influence the specification of the regularization parameter and the sufficient conditions of all the theoretical results in Section 4.

D Supplement on an Extension with a Diverging number of structural parameters

In Sections 2 and 3 of the paper, we have considered model (2.1) under the setting that the number of covariates d is a fixed integer. However, when there are a moderate or large number of covariates, it is appropriate to assume that d diverges to infinity with the sample size. In this section, we consider model (2.1) with the assumption that $d \to \infty$ and $d \ll n$.

Since the number of covariates grows orderly slower than the sample size, we can continue to use the penalized estimation (2.3) for $(\boldsymbol{\mu}^\star, \boldsymbol{\beta}^\star)$ and the penalized two-step estimation (3.3) for $\boldsymbol{\beta}^\star$. The corresponding estimators are still denoted as $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ and $\tilde{\boldsymbol{\beta}}$, but we should note that their dimensions diverge to infinity with n. The characterizations of $\hat{\boldsymbol{\beta}}$ in Lemmas B.1 and B.2 are still valid since they are finite-sample results. The iteration algorithm also wpg1 stops at the second iteration, which is shown by Proposition D.1.

As before, it is critical to properly specify the regularization parameter λ for the case with a diverging number of covariates.

Assumption (B'): The regularization parameter λ satisfies

$$\sqrt{d}\kappa_n \ll \lambda, \ \alpha\gamma_n \le \lambda, \text{ and } \lambda \ll \mu^\star, \tag{3}$$

where κ_n and γ_n are defined in (2.2) and $\alpha > 2$.

Comparing Assumption (B') with Assumption (B) of the paper, the main difference in formation is that κ_n is changed to $\sqrt{d}\kappa_n$. In fact, κ_n in (3) also depends on d, which is shown in Supplement C. This difference is caused by the assumption that d diverges to ∞.

Lemma D.1 (On Index Sets S_{ij}'s)**.** *Under Assumptions (A) and (B'), the conclusion of Lemma 1 of the paper holds.*

Thus, wpg1, still valid is the crucial analytic expression of $\hat{\boldsymbol{\beta}}$ (3.1), from which we derive its theoretical properties. They are similar to those in the previous section, with additional technical complexity caused by the diverging dimension d.

Denote $\|\cdot\|_{F,d} = d^{-1/2}\|\cdot\|_F$, where $\|\cdot\|_F$ is the Frobenius norm. Let the average of the square root of the fourth marginal moments of $\boldsymbol{X}_0$ be $\kappa_X = d^{-1}\sum_{j=1}^{d}(\mathbb{E}[X_{0j}^4])^{1/2}$. We make the following assumptions on $\boldsymbol{\Sigma}_X$ and κ_X.
Assumption (D): $\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}$ is bounded.
Assumption (E): κ_X is bounded.

Theorem D.1 (Existence and Consistency on $\hat{\boldsymbol{\beta}}$)**.** *Suppose Assumptions (A), (B'), (D) and (E) hold. If there exists r_d, a sequence of positive numbers depending on d, such that $d^3/n \to 0$, $(r_d d)^2/n \to 0$, $s_1 = o(n/(r_d\sqrt{d}\kappa_n\lambda))$ and $s_2 = o(n/(r_d\sqrt{d}\kappa_n\gamma_n))$, then, for every fixed $C > 0$, wpg1, there exists a unique estimator $\hat{\boldsymbol{\beta}} \in B_C(\boldsymbol{\beta}^\star)$ such that $\psi_n(\hat{\boldsymbol{\beta}}) = 0$ and $r_d\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \xrightarrow{P} 0$.*

Next, we consider the asymptotic distribution on $\hat{\boldsymbol{\beta}}$. Since the dimension of $\hat{\boldsymbol{\beta}}$ diverges to infinity, following Fan et al. (2011), it is more appropriate to study its linear maps. Let $\boldsymbol{A}_n$ be a $q \times d$ matrix, where q is a fixed integer, $\boldsymbol{G}_n = \boldsymbol{A}_n\boldsymbol{A}_n^T$ with the largest eigenvalue $\lambda_{\max}(\boldsymbol{G}_n)$, and $\boldsymbol{G}_{X,n} = \boldsymbol{A}_n\boldsymbol{\Sigma}_X^{-1}\boldsymbol{A}_n^T$. Denote by $\lambda_{\min}(\boldsymbol{\Sigma}_X)$ the smallest eigenvalue of $\boldsymbol{\Sigma}_X$, $\sigma_{X,\max}^2 = \max_{1\le j\le d}\mathrm{Var}[X_{0j}]$, $\sigma_{X,\min}^2 = \min_{1\le j\le d}\mathrm{Var}[X_{0j}]$ and $\gamma_{X,\max} = \max_{1\le j\le d}\mathbb{E}|X_{0j}|^3$. Abbreviate "with respect to" by "wrt". We assume further
Assumption (D'): $\lambda_{\min}(\boldsymbol{\Sigma}_X)$ is bounded away from zero, which implies Assumption (D).
Assumption (D"): $\|\boldsymbol{\Sigma}_X\|_{F,d}$ is bounded.
Assumption (F): $\|\boldsymbol{A}_n\|_F$ and $\lambda_{\max}(\boldsymbol{G}_n)$ are bounded and $\boldsymbol{G}_{X,n}$ converges to a $q \times q$ symmetric matrix $\boldsymbol{G}_X$ wrt $\|\cdot\|_F$.
Assumption (G): $\sigma_{X,\max} > 0$; $\sigma_{X,\max}$ and $\gamma_{X,\min}$ are bounded from above and $\sigma_{X,\min}$ is bounded away from zero.

Similar to the main case of Theorem 2, $\hat{\boldsymbol{\beta}}$ is asymptotically Gaussian.

Theorem D.2 (Asymptotic Distribution on $\hat{\boldsymbol{\beta}}$)**.** *Suppose Assumptions (A), (B'), (D'), (D"), (E), (F) and (G) hold. If $s_1 = o(\sqrt{n}/(\lambda\sqrt{d}\kappa_n))$, $s_2 = o(\sqrt{n}/(\sqrt{d}\kappa_n\gamma_n))$ and $d^5\log d = o(n)$, then $\sqrt{n}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}_X)$.*

The penalized estimator $\hat{\boldsymbol{\mu}}$ obtained by (3.2) has partial selection consistency.

Theorem D.3 (Partial Selection Consistency on $\hat{\boldsymbol{\mu}}$)**.** *Suppose Assumptions (A) and (B') hold and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^\star$ wrt $r_d\|\cdot\|_2$. If $r_d \geq 1/\sqrt{d}$, then $P(\mathcal{E}) \to 1$.*

We can construct the penalized two-step estimator $\tilde{\boldsymbol{\beta}}$ through (3.3) with $\hat{\boldsymbol{\mu}}$. This two-step estimator is consistent by Theorem D.2 and its asymptotic distribution, as an extension of the main case in Theorem 4, is given as follows.

Theorem D.4 (Asymptotic Distribution on $\tilde{\boldsymbol{\beta}}$)**.** *Suppose all the assumptions and conditions of Theorem D.2 hold except that the condition on s_1 is not required. Then $\sqrt{n}\boldsymbol{A}_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{G}_X)$.*

From Theorems D.2 and D.4, Wald-type asymptotic confidence regions of $\boldsymbol{\beta}^\star$ are availabe. For example, a confidence region based on $\tilde{\boldsymbol{\beta}}$ with asymptotic confidence level $1-\alpha$ is given by

$$\{\boldsymbol{\beta} \in \mathbb{R}^d : \sigma^{-1}\sqrt{n}\|\boldsymbol{G}_{X,n}^{-1/2}\boldsymbol{A}_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \leq q_\alpha(\chi_q)\}. \tag{4}$$

Since $\boldsymbol{G}_{X,n}$ involves the unknown $\boldsymbol{\Sigma}_X$, we estimate it by $\hat{\boldsymbol{G}}_{X,n} = \boldsymbol{A}_n \hat{\boldsymbol{\Sigma}}_X^{-1} \boldsymbol{A}_n^T$. On the other hand, σ is estimated by $\hat{\sigma}$ in (3.6) as in the paper. After plugging $\hat{\boldsymbol{G}}_{X,n}$ and $\hat{\sigma}$ into (4), we obtain

$$\{\boldsymbol{\beta} \in \mathbb{R}^d : \hat{\sigma}^{-1}\sqrt{n}\|\hat{\boldsymbol{G}}_{X,n}^{-1/2}\boldsymbol{A}_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \leq q_\alpha(\chi_q)\}. \tag{5}$$

By Lemma D.5, the consistency of $\hat{\sigma}$ is assured. Then, Theorem D.1 guarantees the asymptotic validity of the confidence region (5).

Next we provide the proofs of the above theoretical results.

Let $\mathbb{S}_{k,l} = \mathbb{S}_{\{k,k+1,\ldots,l\}}$, $\mathbb{S}^\epsilon_{k,l} = \mathbb{S}^\epsilon_{\{k,k+1,\ldots,l\}}$, $\mathcal{B} = \{\max_{s+1\leq i\leq n}\|\boldsymbol{X}_i\|_2 \leq \kappa_n\}$ and $\mathcal{D} = \bigcap_{i=1}^n\{-\gamma_n \leq \epsilon_i \leq \gamma_n\}$. Then $P(\mathcal{B}) \to 1$ and $P(\mathcal{D}) \to 1$ by (2.2) of the paper.

Proof of Lemma D.1. We first consider S_{i0}'s, then S_{i1}'s, and finally S_{i2}'s with $i = 1, 2, 3$. Consider S_{10}, S_{20} and S_{30}. Let $\mathcal{A} = \{S_{10} = S^\star_{10}\}$. Note that $P(\mathcal{A}) \geq P(\mathcal{A}|\mathcal{B})P(\mathcal{B})$ and $P(\mathcal{B}) \to 1$. It suffices to show that $P(\mathcal{A}|\mathcal{B}) \to 1$. By $\lambda \gg \sqrt{d}\kappa_n$, it follows $P(\mathcal{A}|\mathcal{B}) \geq P(\{s+1 \leq i \leq n : -\lambda + \max_{s+1\leq i\leq n}\|\boldsymbol{X}_i\|_2\sqrt{d}C \leq \epsilon_i \leq \lambda - \max_{s+1\leq i\leq n}\|\boldsymbol{X}_i\|_2\sqrt{d}C\} \supset S^\star_{10}|\mathcal{B}) \geq P(\{s+1 \leq i \leq n : -\lambda+\kappa_n\sqrt{d}C \leq \epsilon_i \leq \lambda-\kappa_n\sqrt{d}C\} \supset S^\star_{10}) \geq P(\mathcal{D}) \to 1$. Thus, wpg1, $S_{10} = S^\star_{10}$. From $S_{10} \cup S_{20} \cup S_{30} = S^\star_{10}$, it follows that, wpg1, $S_{20} = S_{30} = \emptyset$. Consider S_{21}, S_{31} and S_{11}. Recall that $\mu^\star = \min\{|\mu^\star_i| : 1 \leq i \leq s_1\}$ and note that $\lambda - \mu^\star + \sqrt{d}C\kappa_n < -\gamma_n$ when n is large. Let $S_{211} = S_{21}S^\star_{21}$ and $S_{212} = S_{21}S^{\star c}_{21}$. We will show $P(S_{211} = S^\star_{21}) \to 1$ and $P(S_{212} = \emptyset) \to 1$. Then $P(S_{21} = S^\star_{21}) \to 1$. Denote $\mathcal{A}_1 = \{S_{211} \supset S^\star_{21}\}$. On the event $\mathcal{B}$, $S_{211} \supset \{1 \leq i \leq s_1 : \epsilon_i > \lambda-\mu^\star+\sqrt{d}C\kappa_n$ and $\mu^\star_i >$

$0\} \supset \{1 \le i \le s_1 : \epsilon_i > -\gamma_n \text{ and } \mu_i^\star > 0\}$. Then, $P(\mathcal{A}_1) \ge P(\mathcal{A}_1|\mathcal{B})P(\mathcal{B}) \ge P(\{1 \le i \le s_1 : \epsilon_i > -\gamma_n \text{ and } \mu_i^\star > 0\} \supset S_{21}^\star)P(\mathcal{B}) \to 1 \cdot 1 = 1$. It follows that, wpg1, $S_{211} \supset S_{21}^\star$. Note that $S_{211} \subset S_{21}^\star$. Then, wpg1, $S_{211} = S_{21}^\star$. Denote $\mathcal{A}_2 = \{S_{212} = \emptyset\}$. On the event $\mathcal{B}$, $S_{212} \subset \{1 \le i \le s_1 : \epsilon_i > \lambda + \mu^\star - \sqrt{d}C\kappa_n \text{ and } \mu_i^\star < 0\}$, which contains $\{1 \le i \le s_1 : \epsilon_i > \gamma_n\}$. Then, $P(\mathcal{A}_2) \ge P(\mathcal{A}_2|\mathcal{B})P(\mathcal{B}) \ge P(\{1 \le i \le s_1 : \epsilon_i > \gamma_n\} = \emptyset)P(\mathcal{B}) \to = 1$. Then, wpg1, $S_{212} = \emptyset$. Thus, $P(S_{21} = S_{21}^\star) \to 1$. Similarly, we can show, wpg1, $S_{31} = S_{31}^\star$. Note that S_{11}, S_{21} and S_{31} are disjoint and their union is $S_{21}^\star \cup S_{31}^\star$. Then, wpg1, $S_{11} = \emptyset$. Consider S_{12}, S_{22} and S_{32}. Denote $\mathcal{A} = \{S_{12} = S_{12}^\star\}$. Note that $-\lambda - \mu_i^\star + \sqrt{d}C\kappa_n < -\gamma_n$ and $\lambda - \mu_i^\star - \sqrt{d}C\kappa_n > \gamma_n$ when n is large for $s_1 + 1 \le i \le s$. On the event $\mathcal{B}$, $S_{12} \supset \{s_1+1 \le i \le s : -\lambda-\mu_i^\star+\sqrt{d}C\kappa_n \le \epsilon_i \le \lambda-\mu_i^\star-\sqrt{d}C\kappa_n\}$, which contains $\{s_1 + 1 \le i \le s : -\gamma_n \le \epsilon_i \le \gamma_n\}$. Then, $P(\mathcal{A}) \ge P(\mathcal{A}|\mathcal{B})P(\mathcal{B}) \ge P(\{s_1 + 1 \le i \le s : -\gamma_n \le \epsilon_i \le \gamma_n\} = S_{12}^\star)P(\mathcal{B}) \to 1$. Thus, wpg1, $S_{12} = S_{12}^\star$. Note that S_{12}, S_{22} and S_{32} are disjoint and their union is $S_{12}^\star$. Then, wpg1, $S_{22} = S_{32} = \emptyset$. □

Let $\bar{\sigma}_X^2 = d^{-1}\sum_{j=1}^d \mathrm{Var}[X_{0j}]$ and $\bar{\sigma}_{XX}^2 = d^{-2}\sum_{k=1}^d\sum_{l=1}^d \mathrm{Var}[X_{0k}X_{0l}]$. We make the following assumptions.

Assumption (E1): $\bar{\sigma}_X^2$ is bounded.

Assumption (E2): $\bar{\sigma}_{XX}^2$ is bounded.

Assumption (E) in Section 4 implies Assumptions (E1) and (E2) by Cauchy-Schwartz inequality. For simplicity, we adopt the notation $\lesssim$, which means the left hand side is bounded by a constant times the right, where the constant does not affect related analysis.

Below are three lemmas needed for proving Theorem D.1. Suppose that $\boldsymbol{M}$ and $\boldsymbol{E}$ are matrices and $\|\cdot\|$ is a matrix norm and that $\{\boldsymbol{A}_n\}$ is a sequence of random $d \times d$ matrices and $\boldsymbol{A}$ a deterministic $d \times d$ matrix, and denote $\hat{\boldsymbol{\Sigma}}_n = (1/n)\mathbb{S}_n$, the sample covariance matrix.

Lemma D.2 (Stewart (1969))**.** *If* $\|\boldsymbol{I}\| = 1$ *and* $\left\|\boldsymbol{M}^{-1}\right\| \|\boldsymbol{E}\| < 1$*, then*

$$\frac{\left\|(\boldsymbol{M}+\boldsymbol{E})^{-1} - \boldsymbol{M}^{-1}\right\|}{\left\|\boldsymbol{M}^{-1}\right\|} \le \frac{\left\|\boldsymbol{M}^{-1}\right\| \|\boldsymbol{E}\|}{1 - \left\|\boldsymbol{M}^{-1}\right\| \|\boldsymbol{E}\|}.$$

Lemma D.3. *If* $\left\|\boldsymbol{A}^{-1}\right\|_{F,d}$ *is bounded,* $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{A}$*, and* $r_d \ge 1/\sqrt{d}$*, then* $\boldsymbol{A}_n^{-1} \xrightarrow{P} \boldsymbol{A}^{-1}$*, where the convergence in probability is wrt* $r_d \|\cdot\|_F$*.*

Proof of Lemma D.3. Let $\boldsymbol{E} = \boldsymbol{A}_n - \boldsymbol{A}$. Note that $r_d \ge 1/\sqrt{d}$. Then, $r_d \|\boldsymbol{E}\|_F \xrightarrow{P} 0$ implies $\|\boldsymbol{E}\|_{F,d} \xrightarrow{P} 0$. Thus, wpg1, $\|\boldsymbol{E}\|_{F,d}$ is bounded by a constant $C > 0$. By Lemma D.2,

$$\left\|\boldsymbol{A}_n^{-1} - \boldsymbol{A}^{-1}\right\|_{F,d} \le \left\|\boldsymbol{A}^{-1}\right\|_{F,d} \frac{\left\|\boldsymbol{A}^{-1}\right\|_{F,d} \|\boldsymbol{E}\|_{F,d}}{1 - \left\|\boldsymbol{A}^{-1}\right\|_{F,d} \|\boldsymbol{E}\|_{F,d}} \le C^2 \frac{\|\boldsymbol{E}\|_{F,d}}{1 - C\|\boldsymbol{E}\|_{F,d}}.$$

Therefore,

$$r_d \left\| \boldsymbol{A}_n^{-1} - \boldsymbol{A}^{-1} \right\|_F \leq C^2 \frac{r_d \left\| \boldsymbol{E} \right\|_F}{1 - C \left\| \boldsymbol{E} \right\|_{F,d}} \xrightarrow{P} 0.$$

This completes the proof. □

Lemma D.4. *If Assumption (E2) holds and $r_d^2 d^4/n \to 0$, then $\hat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}_X$ wrt $r_d \left\| \cdot \right\|_F$.*

Proof of Lemma D.4. For any $\delta > 0$, we have

$$P(\left\| \hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X \right\|_F > \delta) \leq \sum_{k=1}^{d} \sum_{l=1}^{d} \frac{d^2}{\delta^2} P(\frac{1}{n} \sum_{i=1}^{n} X_{ik} X_{il} - \sigma_{kl})^2 \leq \frac{d^4}{n} \frac{1}{\delta^2} \bar{\sigma}_{XX}^2.$$

Thus, $P(r_d \left\| \hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X \right\|_F > \delta) \leq \bar{\sigma}_{XX}^2 r_d^2 d^4/(n\delta^2) = o(1)$ by Assumption (E2) and for $r_d^2 d^4/n \to 0$. Thus, $\hat{\boldsymbol{\Sigma}}_n$ is a consistent estimator of $\boldsymbol{\Sigma}_X$ wrt $r_d \left\| \cdot \right\|_F$. □

Proof of Theorem D.1. By the proof of Lemma 2 in the paper, wpg1, the solution $\hat{\boldsymbol{\beta}}_n$ to $\varphi_n(\boldsymbol{\beta}) = 0$ on $\mathcal{B}_C(\boldsymbol{\beta}^\star)$ is explicitly given by $\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^\star + T_0^{-1}(T_1 + T_2 + T_3 - T_4)$, where $T_0 = (1/n)\mathbb{S}_{s_1+1,n}$, $T_1 = (1/n)\mathbb{S}^{\mu}_{S^\star_{12}}$, $T_2 = (1/n)\mathbb{S}^{\epsilon}_{s_1+1,n}$, $T_3 = (\lambda/n)\mathcal{S}_{S^\star_{21}}$ and $T_4 = (\lambda/n)\mathcal{S}_{S^\star_{31}}$. Then, $r_d \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star\|_2 \leq \left\| T_0^{-1} \right\|_{F,d} \sum_{i=1}^{4} r_d \sqrt{d} \|T_i\|_2$. We will show that $\left\| T_0^{-1} \right\|_{F,d}$ is bounded by a positive constant wpg1 and $r_d \sqrt{d} \|T_i\|_2 \xrightarrow{P} 0$ for $i = 1, 2, 3, 4$. Then, $r_d \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star\|_2 = o_P(1)$. Consider T_0. By Lemma D.4, $\|T_0 - \boldsymbol{\Sigma}_X\|_{F,d} \xrightarrow{P} 0$ under Assumption (E2) and the condition $d^3/n \to 0$. Then, by Lemma D.3, together with Assumption (D), $\left\| T_0^{-1} - \boldsymbol{\Sigma}_X^{-1} \right\|_{F,d} \xrightarrow{P} 0$. This implies that, wpg1, $\left\| T_0^{-1} \right\|_{F,d}$ is bounded by a positive constant. Consider T_1. Wpg1, $r_d \sqrt{d} \|T_1\|_2 \leq r_d \sqrt{d} s_2 \kappa_n \gamma_n / n = o(1)$ for $s_2 = o(n/(r_d \sqrt{d} \kappa_n \gamma_n))$. Consider T_2. For any $\delta > 0$, $P(\|T_2\|_2 > \delta) \leq (1/\delta^2) P \|(1/n) \sum_{i=s_1+1}^{n} \boldsymbol{X}_i \epsilon_i\|_2^2 \leq d\sigma^2 \bar{\sigma}_X^2/(n\delta^2)$, where $\bar{\sigma}_X^2 = (1/d) \sum_{j=1}^{d} \sigma_j^2$. Thus, $P(r_d \sqrt{d} \|T_2\|_2 > \delta) \leq r_d^2 d^2 \sigma^2 \bar{\sigma}_X^2/(n\delta^2) \to 0$ by Assumption **(E1)** and $(r_d d)^2/n \to 0$. Consider T_3 and T_4. Wpg1, $r_d \sqrt{d} \|T_3\|_2 \leq r_d \sqrt{d} \lambda s_1 \kappa_n / n = o(1)$ for $s_1 = o(n/(r_d \sqrt{d} \lambda \kappa_n))$. Similarly, $r_d \sqrt{d} \|T_4\|_2 = o_P(1)$. □

The next lemma is needed for proving Theorem D.2. Suppose $\{\boldsymbol{\xi}_i\}$ are i.i.d. copies of $\boldsymbol{\xi}_0$, a d-dimensional random vector with mean zero. Denote $\sigma^2_{\xi,\max} = \max_{1 \leq j \leq d} \mathrm{Var}[\xi_{0j}]$, $\sigma^2_{\xi,\min} = \min_{1 \leq j \leq d} \mathrm{Var}[\xi_{0j}]$ and $\gamma_{\xi,\max} = \max_{1 \leq j \leq d} \mathbb{E}|\xi_{0j}|^3$.

Lemma D.5. *Suppose $\sigma_{\xi,\max}$ and $\gamma_{\xi,\max}$ are bounded from above and $\sigma_{\xi,\max}$ is bounded from zero. If $d = o(\sqrt{n})$, then $(1/\sqrt{n}) \sum_{i=1}^{n} \boldsymbol{\xi}_i = O_P(\sqrt{d \log d})$ wrt $\|\cdot\|_2$.*

Proof of Lemma D.5. Let $\alpha_d = \sqrt{d \log d}$ and $C_1 \geq \sqrt{2} \sigma_{\xi,\max}$. Then

$$P(\|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\xi}_i\|_2 > \alpha_d C_1) \leq \sum_{j=1}^{d} P(|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\xi_{ij}}{\sigma_j}| > \frac{\alpha_d C_1}{\sigma_j \sqrt{d}}),$$

where σ_j is the standard deviation of ξ_{0j}. By Berry and Esseen Theorem (see, for example, P375 in Shiryaev (1995)), there exists a constant $C_2 > 0$ such that

$$P(\|(1/\sqrt{n})\sum_{i=1}^{n}\boldsymbol{\xi}_i\|_2 > \alpha_d C_1) \le T_1 + 2T_2,$$

where

$$T_1 = \sum_{j=1}^{d} P(|N(0,1)| > \frac{\alpha_d C_1}{\sigma_j\sqrt{d}}), \quad T_2 = \sum_{j=1}^{d}\frac{C_2\mathbb{E}|\xi_{0j}|^3}{\sigma_j^3\sqrt{n}}.$$

By noting $d^2 = o(n)$,

$$T_1 \le \sum_{j=1}^{d} P(|N(0,1)| > \frac{\alpha_d C_1}{\sigma_{\xi,\max}\sqrt{d}}) < 2d\frac{\sigma_{\xi,\max}\sqrt{d}}{\alpha_d C_1}\phi(\frac{\alpha_d C_1}{\sigma_{\xi,\max}\sqrt{d}}) \to 0,$$
$$T_2 \le \sum_{j=1}^{d}\frac{C_2\gamma_{\xi,\max}}{\sigma_{\xi,\min}^3\sqrt{n}} = d\frac{C_2\gamma_{\xi,\max}}{\sigma_{\min}^3\sqrt{n}} \to 0.$$

Therefore, $\|(1/\sqrt{n})\sum_{i=1}^{n}\boldsymbol{\xi}_i\|_2 = O_P(\alpha_d)$. □

Proof of Theorem D.2. We reuse the notations T_i's in the proof of Theorems 5 in the paper, from which, $\sqrt{n}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) = V_1 + V_2 + V_3 - V_4$, where $V_i = \boldsymbol{B}_n T_i$ for $i = 1,2,3,4$ and $\boldsymbol{B}_n = \sqrt{n}\boldsymbol{A}_n T_0^{-1}$. It is sufficient to show that $V_2 \xrightarrow{d} N(0,\sigma^2\boldsymbol{G}_X)$ and other V_i's are $o_P(1)$. Consider V_1. We have $\|V_1\|_2 \le \sqrt{nd}\,\|\boldsymbol{A}_n\|_F\,\|T_0^{-1}\|_{F,d}\,\|T_1\|_2$. By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is bounded. By Lemmas D.3 and D.4 and Assumption (D), for $d = o(n^{1/3})$, wpg1, $\|T_0^{-1}\|_{F,d}$ is bounded. We have, wpg1, $\|T_1\|_2 \le s_2\kappa_n\gamma_n/n$. Then, $\|V_1\|_2 \lesssim \sqrt{d/n}s_2\kappa_n\gamma_n$, Thus, $\|V_1\|_2 = o_P(1)$ for $s_2 = o(\sqrt{n}/(\sqrt{d}\kappa_n\gamma_n))$. Consider V_2. We have $V_2 = V_{21} + V_{22}$, where $V_{21} = \sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{-1}T_2$ and $V_{22} = \sqrt{n}\boldsymbol{A}_n(T_0^{-1} - \boldsymbol{\Sigma}_X^{-1})T_2$. First, note that $V_{21} = \sqrt{(n-s_1)/n}\sum_{i=s_1+1}^{n}\boldsymbol{Z}_{n,i}$, where $\boldsymbol{Z}_{n,i} = (1/\sqrt{n-s_1})\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{-1}\boldsymbol{X}_i\epsilon_i$. On one hand, for every $\delta > 0$, $\sum_{i=s_1+1}^{n}\mathbb{E}\|\boldsymbol{Z}_{n,i}\|_2^2\{\|\boldsymbol{Z}_{n,i}\|_2 > \delta\} \le (n-s_1)\mathbb{E}\|\boldsymbol{Z}_{n,0}\|_2^4/\delta^2$, and $\mathbb{E}\|\boldsymbol{Z}_{n,0}\|_2^4 = (n-s_1)^{-2}\mathbb{E}\epsilon_0^4\mathbb{E}(\boldsymbol{X}_0^T\boldsymbol{\Sigma}_X^{-1}\boldsymbol{A}_n^T\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{-1}\boldsymbol{X}_0)^2$, which is less than or equal to

$$\frac{d^2}{(n-s_1)^2}\mathbb{E}\epsilon_0^4\lambda_{\max}(\boldsymbol{G}_n)\lambda_{\min}^{-2}(\boldsymbol{\Sigma}_X)\kappa_X^2.$$

Then, by Assumptions (D'), (E) and (F) and for $d = o(\sqrt{n})$,

$$\sum_{i=s_1+1}^{n}\mathbb{E}\|\boldsymbol{Z}_{n,i}\|_2^2\{\|\boldsymbol{Z}_{n,i}\|_2 > \delta\} \to 0.$$

On the other hand, $\sum_{i=s_1+1}^{n}\mathrm{Cov}(\boldsymbol{Z}_{n,i}) = \sigma^2\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{-1}\boldsymbol{A}_n^T \to \sigma^2\boldsymbol{G}_X$ by Assumption (F). Thus, by central limit theorem (see Proposition 2.27 in van der Vaart (1998)), $V_{21} \xrightarrow{d} N(0,\sigma^2\boldsymbol{G}_X)$.

Next, consider V_{22}. Note that

$$\|V_{22}\|_2 \le \|\boldsymbol{A}_n\|_F \left(d\log(d)\right)^{1/2} \left\|T_0^{-1} - \boldsymbol{\Sigma}_X^{-1}\right\|_F \left(d\log(d)\right)^{-1/2} \|\sqrt{n}T_2\|_2.$$

By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is $O(1)$; by Lemmas D.3 and D.4, for $d^5\log(d) = o(n)$, $\left(d\log(d)\right)^{1/2} \left\|T_0^{-1} - \boldsymbol{\Sigma}_X^{-1}\right\|_F = o_P(1)$; by Lemma D.5, for $d = o(\sqrt{n})$,

$$\left(d\log(d)\right)^{-1/2} \|\sqrt{n}T_2\|_2 = \left(d\log(d)\right)^{-1/2} \|\frac{1}{\sqrt{n}}\mathbb{S}^{\epsilon}_{s_1+1,n}\|_2 = O_P(1).$$

Then, $V_{22} \xrightarrow{P} 0$. Thus, by Slutsky's lemma, $V_2 \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}_X)$. Consider V_3 and V_4. First consider V_3. By noting that $s_1 = o(\sqrt{n}/(\lambda\sqrt{d}\kappa_n))$, wpg1,

$$\|V_3\|_2 \le \sqrt{nd}\,\|\boldsymbol{A}_n\|_F \left\|T_0^{-1}\right\|_{F,d} \|T_3\|_2 \lesssim \sqrt{d}\lambda s_1\kappa_n/\sqrt{n} \to 0.$$

Thus, $\|V_3\|_2 = o_P(1)$. In the same way, $\|V_4\|_2 = o_P(1)$. □

Proof of Theorem D.3. By the definition of $\mathcal{E}$, we have $P(\mathcal{E}) = T_1T_2T_3$, where $T_1 = P(\bigcap_{i=1}^{s_1}\{|\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) + \epsilon_i| > \lambda\})$, $T_2 = P(\bigcap_{i=s_1+1}^{s}\{|\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) + \epsilon_i| \le \lambda\})$ and $T_3 = P(\bigcap_{i=s+1}^{n}\{|\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) + \epsilon_i| \le \lambda\})$. We will show that each T_i converges to one. Then, $P(\mathcal{E}) \to 1$. Denote $\mathcal{C} = \{r_d\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le 1\}$. Then $P(\mathcal{C}) \to 1$ since $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^\star$ wrt $r_d\|\cdot\|_2$. Consider T_1. We have $1 - T_1 \le T_{11} + T_{12}$, where $T_{11} = P(\bigcup_{i\in S_{21}^\star}\{|\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) + \epsilon_i| \le \lambda\})$ and $T_{12} = P(\bigcup_{i\in S_{31}^\star}\{|\mu_i^\star + \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) + \epsilon_i| \le \lambda\})$. It is sufficient to show that both T_{11} and T_{12} converge to zero. By $\sqrt{d}\kappa_n \ll \lambda \ll \mu^\star$, $T_{11} \le P(\bigcup_{i\in S_{21}^\star}\{\epsilon_i \le \lambda - \mu^\star + \|\boldsymbol{X}_i\|_2 \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2\}, \mathcal{C}) + P(\mathcal{C}^c)$, which is $\le P(\bigcup_{i\in S_{21}^\star}\{\epsilon_i \le \lambda - \mu^\star + \sqrt{d}\kappa_n\}) + P(\mathcal{C}^c) \le s_1 P\{\epsilon_0 \le -\gamma_n\} + P(\mathcal{C}^c) \longrightarrow 0$. Similarly, $T_{12} \to 0$. Thus $T_1 \to 1$. Consider T_2 and T_3. By $\alpha\gamma_n \le \lambda$ and $\sqrt{d}\kappa_n \ll \lambda$, $T_2 \ge P(\bigcap_{i=s_1}^{s}\{-\lambda - \mu_i^\star + (1/r_d)\kappa_n \le \epsilon_i \le \lambda - \mu_i^\star - (1/r_d)\kappa_n\}, \mathcal{C})$, which is $\ge P(\bigcap_{i=s_1}^{s}\{-\lambda - \mu_i^\star + \sqrt{d}\kappa_n \le \epsilon_i \le \lambda - \mu_i^\star - \sqrt{d}\kappa_n\}, \mathcal{C}) \ge P(\bigcap_{i=s_1}^{s}\{-\gamma_n \le \epsilon_i \le \gamma_n\}, \mathcal{C}) \to 1$. Then, $T_2 \to 1$. Similarly, $T_3 \to 1$. □

Proof of Theorem D.4. Note that $\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) = \tilde{R}_1 + \tilde{R}_2 + V_1 + V_2$, where $\tilde{R}_1 = \sqrt{n}\boldsymbol{A}_nR_1$, $\tilde{R}_2 = \sqrt{n}\boldsymbol{A}_nR_2$, $R_1 = (\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{X}_{\hat{I}_0})^{-1}\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{Y}_{\hat{I}_0}\{\hat{I}_0 \ne I_0\}$ and $R_2 = -(\boldsymbol{X}_{I_0}^T\boldsymbol{X}_{I_0})^{-1}\boldsymbol{X}_{I_0}^T\boldsymbol{Y}_{I_0}\{\hat{I}_0 \ne I_0\}$, and V_i's are defined in the proof of Theorem D.2. Since $P(\|\tilde{R}_1\|_2 = 0) \ge P\{\hat{I}_0 = I_0\} \to 1$, we have $\tilde{R}_1 = o_P(1)$. Similarly, $\tilde{R}_2 = o_P(1)$. By the proof of Theorem D.2, $V_1 = o_P(1)$ and $V_2 \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}_X)$. Therefore, the desired result follows by Slutsky's lemma. □

Lemma D.6 (Consistency on $\hat{\sigma}$)**.** *Suppose the assumptions and conditions of Theorem D.1 hold with $r_d \ge \sqrt{d}$. If $s_2 = o(n/\gamma_n^2)$, then $\hat{\sigma} \xrightarrow{P} \sigma$.*

Proof of Lemma D.6. Since the assumptions and conditions of Theorem D.1 hold with $r_d \ge \sqrt{d}$, the penalized estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent estimators of $\boldsymbol{\beta}^\star$ wrt $\sqrt{d}\|\cdot\|_2$

by Theorems 5 in the paper and D.6 in Supplement **??**. Let $\mathcal{A} = \{\hat{I}_0 = I_0\}$. Then $\mathcal{A}$ occurs wpg1 by Theorem 7 in the paper.

Note that $\hat{\sigma}^2 = T\mathcal{A} + \hat{\sigma}^2\mathcal{A}^c$, where $T = (n-s_1)^{-1}\|\boldsymbol{Y}_{I_0} - \boldsymbol{X}_{I_0}^T\tilde{\boldsymbol{\beta}}\|_2^2$. It suffices to show that $T \xrightarrow{P} \sigma^2$. Note that $T = \sum_{i=1}^6 T_i$, where $T_1 = (n-s_1)^{-1}\sum_{i=s_1+1}^n [\boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})]^2$, $T_2 = (n-s_1)^{-1}\sum_{i=s_1+1}^n \epsilon_i^2$, $T_3 = 2(n-s_1)^{-1}\sum_{i=s_1+1}^n \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})\epsilon_i$, $T_4 = (n-s_1)^{-1}\sum_{i=s_1+1}^s \mu_i^{\star 2}$, $T_5 = 2(n-s_1)^{-1}\sum_{i=s_1+1}^s \mu_i \boldsymbol{X}_i^T(\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}})$ and $T_6 = 2(n-s_1)^{-1}\sum_{i=s_1+1}^s \mu_i^\star\epsilon_i$. It is clear that $T_2 \xrightarrow{P} \sigma^2$. Thus, it is sufficient to show other T_i's are $o_P(1)$. For every $\eta > 0$, wpg1, $\sqrt{d}\|\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}}\|_2 \le \eta$. By Assumption (E1), wpg1, $|T_1| \le d^{-1}(n-s_1)^{-1}\sum_{i=s_1+1}^n \|\boldsymbol{X}_i^T\|_2^2(\sqrt{d}\|\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}}\|_2)^2 \le 2\eta^2 d^{-1}\mathbb{E}\|\boldsymbol{X}_0^T\|_2^2 = 2\eta^2\bar{\sigma}_X^2 \lesssim \eta^2$. For every $\eta > 0$, wpg1, $|T_3| \le 2d^{-1/2}(n-s_1)^{-1}\sum_{i=s_1+1}^n \|\boldsymbol{X}_i^T\epsilon_i\|_2\sqrt{d}\|\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}}\|_2 \le 4\eta d^{-1/2}\mathbb{E}\|\boldsymbol{X}_0^T\epsilon_0\|_2 = 4\sigma\eta\bar{\sigma}_X \lesssim \eta$. For $s_2 = o(n/\gamma_n^2)$, $|T_4| \le (n-s_1)^{-1}s_2\gamma_n^2 \to 0$. For $s_2 = o(\sqrt{d}n/(\gamma_n\kappa_n))$, $|T_5| \le 2d^{-1/2}(n-s_1)^{-1}s_2\gamma_n\kappa_n\sqrt{d}\|\boldsymbol{\beta}^\star - \tilde{\boldsymbol{\beta}}\|_2 \le 2\eta d^{-1/2}(n-s_1)^{-1}s_2\gamma_n\kappa_n \xrightarrow{P} 0$. For $s_2 = o(n/\gamma_n)$, wpg1, $|T_6| \le 4(n-s_1)^{-1}\gamma_n s_2\mathbb{E}|\epsilon_0| \to 0$. □

Theorem D.5 (Asymptotic Distributions on $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{G}}_{X,n}$)**.** *Under the assumptions and conditions of Theorem D.2, if* $d^8(\log(d))^2 = o(n)$*, then* $\sqrt{n}\hat{\boldsymbol{G}}_{X,n}^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{I}_q)$*. Similarly, under the assumptions and conditions of Theorem 8 in the paper, If* $d^8(\log(d))^2 = o(n)$*, then* $\sqrt{n}\hat{\boldsymbol{G}}_{X,n}^{-1/2}\boldsymbol{A}_n(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{I}_q)$.

Note that a stronger requirement on d is required to handle $\hat{\boldsymbol{G}}_{X,n}^{-1/2}$ in Theorem D.5.. Below is a lemma needed for proving Theorem D.5.

Lemma D.7 (Wihler (2009))**.** *Suppose* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are* $m \times m$ *symmetric positive semidefinite matrices. Then, for* $p > 1$*,*

$$\left\|\boldsymbol{A}^{1/p} - \boldsymbol{B}^{1/p}\right\|_F^p \le m^{(p-1)/2}\|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

Specifically, for $p = 2$*,*

$$\left\|\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2}\right\|_F \le (m^{1/2}\|\boldsymbol{A} - \boldsymbol{B}\|_F)^{1/2}.$$

Proof of Theorem D.5. We only show the result on $\hat{\boldsymbol{\beta}}$. since the result on $\tilde{\boldsymbol{\beta}}$ can be obtained in a similar way. We reuse the notations T_i's in the proof of Theorems 5 in the paper, from which, $\sqrt{n}\hat{\boldsymbol{G}}_{X,n}^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) = M + R$, where $M = \sqrt{n}\boldsymbol{G}_{X,n}^{-1/2}\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star)$ and $R = \sqrt{n}(\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2})\boldsymbol{A}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star)$. By Theorem D.2, $M \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}_X)$. Then, it is sufficient to show that $R \xrightarrow{P} 0$ wrt $\|\cdot\|_2$. We have $R = R_1 + R_2 + R_3 - R_4$, where $R_i = \boldsymbol{B}_n T_i$ for $i = 1, 2, 3, 4$ and $\boldsymbol{B}_n = \sqrt{n}(\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2})\boldsymbol{A}_n T_0^{-1}$. We will show each R_i converges to zero in probability, which finishes the proof. Before that, we first establish an inequality for $\left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F$. By Lemma D.7, $\left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F \le (\sqrt{q}\left\|\hat{\boldsymbol{G}}_{X,n}^{-1} - \boldsymbol{G}_{X,n}^{-1}\right\|_F)^{1/2}$. Note that, by Lemma D.4, $\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F \xrightarrow{P} 0$ for $d^4 = o(n)$.

Then, by Lemma D.3,

$$\left\|\hat{\boldsymbol{G}}_{X,n} - \boldsymbol{G}_{X,n}\right\|_F \leq \|\boldsymbol{A}_n\|_F^2 \left\|\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_X^{-1}\right\|_F \lesssim \|\boldsymbol{A}_n\|_F^2 \left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F \xrightarrow{P} 0.$$

Thus, by Lemma D.3,

$$\left\|\hat{\boldsymbol{G}}_{X,n}^{-1} - \boldsymbol{G}_{X,n}^{-1}\right\|_F \lesssim \left\|\hat{\boldsymbol{G}}_{X,n} - \boldsymbol{G}_{X,n}\right\|_F \lesssim \|\boldsymbol{A}_n\|_F^2 \left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F.$$

Since q is a fixed integer, it follows

$$\left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F \lesssim \|\boldsymbol{A}_n\|_F \,(\sqrt{q}\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F)^{1/2} \lesssim \|\boldsymbol{A}_n\|_F \,(\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F)^{1/2}.$$

Consider R_1. Note that

$$\|R_1\|_2 \leq \sqrt{n}\sqrt{d}\left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F \|\boldsymbol{A}_n\|_F \left\|T_0^{-1}\right\|_{F,d} \|T_1\|_2,$$

which is $\lesssim \sqrt{n}(d\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F)^{1/2} \|\boldsymbol{A}_n\|_F^2 \left\|T_0^{-1}\right\|_{F,d} \|T_1\|_2$. By Lemmas D.3 and D.4, $d\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F = o_P(1)$ for $d^6 = o(n)$. By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is bounded. By Lemmas D.3 and D.4 and Assumption (D), for $d = o(n^{1/3})$, wpg1, $\left\|T_0^{-1}\right\|_{F,d}$ is bounded. Also note that, wpg1, $\|T_1\|_2 \leq s_2\kappa_n\gamma_n/n$. Then, $\|R_1\|_2 \lesssim s_2\kappa_n\gamma_n/\sqrt{n}$. Thus, $\|R_1\|_2 = o_P(1)$ for $s_2 = o(\sqrt{n}/(\kappa_n\gamma_n))$. Consider R_2. Note that

$$\|R_2\|_2 \leq \left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F \|\boldsymbol{A}_n\|_F \left\|T_0^{-1}\right\|_F \|\sqrt{n}T_2\|_2,$$

which is $\lesssim (d^2\log(d)\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F)^{1/2} \|\boldsymbol{A}_n\|_F^2 \left\|T_0^{-1}\right\|_{F,d} (d\log(d))^{-1/2}\|\sqrt{n}T_2\|_2$. By Lemmas D.3 and D.4, $d^2\log(d)\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F$ is $o_P(1)$ for $d^8(\log(d))^2 = o(n)$. By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is $O(1)$. By Lemmas D.3 and D.4 and Assumption (D), for $d = o(n^{1/3})$, wpg1, $\left\|T_0^{-1}\right\|_{F,d}$ is bounded. By Lemma D.5, $(d\log(d))^{-1/2}\|\sqrt{n}T_2\|_2 = (d\log(d))^{-1/2}\|\frac{1}{\sqrt{n}}\mathbb{S}^{\epsilon}_{s_1+1,n}\|_2$ is $O_P(1)$ for $d = o(\sqrt{n})$. Thus, $R_2 \xrightarrow{P} 0$. Consider R_3 and R_4. By $s_1 = o(\sqrt{n}/(\lambda\kappa_n))$, wpg1, $\|R_3\|_2 \leq \sqrt{n}\left\|\hat{\boldsymbol{G}}_{X,n}^{-1/2} - \boldsymbol{G}_{X,n}^{-1/2}\right\|_F \|\boldsymbol{A}_n\|_F \left\|T_0^{-1}\right\|_F \|T_3\|_2$, which is
$\lesssim \sqrt{n}(d\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_X\right\|_F)^{1/2} \|\boldsymbol{A}_n\|_F^2 \left\|T_0^{-1}\right\|_{F,d} \|T_3\|_2 \lesssim \lambda s_1\kappa_n/\sqrt{n} \to 0$. Thus, $\|R_3\|_2 = o_P(1)$. In the same way, $\|R_4\|_2 = o_P(1)$. □

Next, we extend Proposition B.1 to the case with $d \to \infty$ and $d \ll n$. Before that, we list three simple lemmas for a diverging d. Suppose $\{\boldsymbol{\xi}_i\}$ is a sequence of i.i.d. copies of $\boldsymbol{\xi}_0$, a d-dimensional random vector with mean zero. Denote $\bar{\sigma}_\xi^2 = (1/d)\sum_{j=1}^d \mathrm{Var}[\xi_{0j}]$.

Lemma D.8. *Suppose $\bar{\sigma}_\xi^2$ is bounded. If $d/n = o(1)$, then*

$$\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\|_2 \xrightarrow{P} 0.$$

Lemma D.9. *Suppose $\bar{\sigma}_\xi^2$ is bounded. If $d/n = o(1)$, then*

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\xi}_i\|_2 - P\|\boldsymbol{\xi}_0\|_2 \xrightarrow{P} 0.$$

Lemma D.10. *Suppose $\boldsymbol{a}$ is a vector. Then*

$$\left\|\boldsymbol{a}\boldsymbol{a}^T\right\|_F = \|\boldsymbol{a}\|_2^2.$$

Suppose the specification of the regularization parameter is given by

$$d\kappa_n \ll \lambda,\ \alpha\gamma_n \leq \lambda,\ \text{and}\ \lambda \ll \mu^\star, \tag{6}$$

where α is a constant greater than 2.

Proposition D.1. *Suppose assumptions (D) and (G) hold and the regularization parameter satisfies (6). Suppose there exist constants C_1 and C_2 such that $\|\boldsymbol{\beta}^\star\|_2 < C_1\sqrt{d}$ and $\|\boldsymbol{\beta}^{(0)}\|_2 < C_2\sqrt{d}$ wpg1. If the regularization parameter satisfies (??), $s_1\lambda\kappa_n/(n\sqrt{d}) = o(1)$ and $s_2\kappa_n\gamma_n/(n\sqrt{d}) = o(1)$, then, for every $K \geq 1$, with at least probability $p_{n,K}$ which increases to one as $n \to \infty$, $\|\boldsymbol{\beta}^{(K+1)} - \boldsymbol{\beta}^{(K)}\|_2 \leq O((\sqrt{d}s_1\kappa_n^2/n)^K d)$ and $\|\boldsymbol{\beta}^{(k)}\|_2 \leq (2C_1 + C_2)d$ for all $k \leq K$. Specifically, wpg1, the iterative algorithm stops at the second iteration.*

Proof of Proposition D.1. Reuse the notations in the proof of Lemma B.1. First, we show that, wpg1, $\|\boldsymbol{\beta}^{(1)}\|_2 \leq (2C_1 + C_2)d$. For each $k \geq 1$,

$$\mathbb{S}\boldsymbol{\beta}^{(k)} = \mathbb{S}_{S_{11}}^{\mu} + \mathbb{S}_{S_{12}}^{\mu} + \mathbb{S}_{S_1}\boldsymbol{\beta}^\star + \mathbb{S}_{S_1}^{\epsilon} + \mathbb{S}_{S_2\cup S_3}\boldsymbol{\beta}^{(k-1)} + \lambda(\mathcal{S}_{S_2} - \mathcal{S}_{S_3}),$$

Since the regularization parameter satisfies (6), it is easy to check that the conclusion of Lemma 2 in the paper continues to hold, which implies $P(\mathcal{A}_0) \to 1$.

Thus, wpg1,

$$\boldsymbol{\beta}^{(1)} = T_0^{-1}T_1 + T_0^{-1}T_2 + T_0^{-1}T_3 + T_0^{-1}T_4(\boldsymbol{\beta}^{(0)}) + T_0^{-1}T_5.$$

We will show that, wpg1,

$$\begin{aligned}
&\|T_0^{-1}T_1\|_2 \leq (C_2/4)d,\\
&\|T_0^{-1}T_2\|_2 \leq 2C_1 d,\\
&\|T_0^{-1}T_3\|_2 \leq (C_2/4)d,\\
&\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 \leq (C_2/4)d,\\
&\|T_0^{-1}T_5\|_2 \leq (C_2/4)d.
\end{aligned}$$

Thus, wpg1,

$$\|\boldsymbol{\beta}^{(1)}\|_2 \leq \sum_{i=1}^{5}\|T_0^{-1}T_i\|_2 \leq (2C_1 + C_2)d.$$

On $T_0^{-1}T_1$. Under Assumption (D), for $s_2\kappa_n\gamma_n/(n\sqrt{d}) = o(1)$, wpg1,

$$\|T_0^{-1}T_1\|_2 \leq \left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_F \|\frac{1}{n}\mathbb{S}^{\mu}_{S^\star_{12}}\|_2 \leq 2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\frac{s_2}{n\sqrt{d}}\kappa_n\gamma_n d \to 0.$$

Thus, wpg1, $\|T_0^{-1}T_1\|_2 \leq C_2 d/4$.

On $T_0^{-1}T_2$. Wpg1,

$$\begin{aligned}\|T_0^{-1}T_2\|_2 &\leq \left\|(\frac{1}{n}\mathbb{S})^{-1}\frac{1}{n}\mathbb{S}_{s_1+1,n}\right\|_F \|\boldsymbol{\beta}^\star\|_2 \\ &\leq \|\boldsymbol{I}_d\|_F C_1\sqrt{d} + \left\|(\frac{1}{n}\mathbb{S})^{-1}\frac{1}{n}\mathbb{S}_{1,s_1}\right\|_F C_1\sqrt{d} \\ &\leq C_1 d + \left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_F \left\|\frac{1}{n}\mathbb{S}_{1,s_1}\right\|_F C_1\sqrt{d},\end{aligned}$$

and

$$\left\|\frac{1}{n}\mathbb{S}_{1,s_1}\right\|_F = \frac{1}{n}\sum_{i=1}^{s_1}\|\boldsymbol{X}_i\|_2^2 \leq \frac{s_1}{n}\kappa_n^2.$$

Thus, Under Assumption (D), for $s_1\kappa_n^2/n = o(1)$, wpg1,

$$\|T_0^{-1}T_2\|_2 \leq C_1 d + 2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\frac{\sqrt{d}s_1}{n}\kappa_n^2 C_1\sqrt{d} \leq 2C_1 d.$$

On $T_0^{-1}T_3$. Under assumptions (D) and (G), for $\log(d)/n = o(1)$, wpg1,

$$\begin{aligned}\|T_0^{-1}T_3\|_2 &= \sqrt{d}\frac{1}{\sqrt{n}}\sqrt{d\log(d)}\left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_{F,d}(d\log(d))^{-1/2}\|\frac{1}{\sqrt{n}}\mathbb{S}^{\epsilon}_{s_1+1,n}\|_2 \\ &\leq \frac{d\sqrt{\log(d)}}{\sqrt{n}}2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}O_P(1) \xrightarrow{P} 0.\end{aligned}$$

Thus, wpg1, $\|T_0^{-1}T_3\|_2 \leq C_2 d/4$.

On $T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})$. Under Assumption (D), for $s_1\kappa_n^2/n$, wpg1,

$$\begin{aligned}\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 &\leq \sqrt{d}\left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_{F,d}\left\|\frac{1}{n}\mathbb{S}_{1,s_1}\right\|_F\|\boldsymbol{\beta}^{(0)}\|_2 \\ &\leq \sqrt{d}2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\frac{s_1}{n}\kappa_n^2 C_2\sqrt{d} \xrightarrow{P} 0.\end{aligned}$$

Thus, wpg1, $\|T_0^{-1}T_4(\boldsymbol{\beta}^{(0)})\|_2 \leq C_2 d/4$.

On $T_0^{-1}T_5$. Under Assumption (D), for $s_1\kappa_n\lambda/(n\sqrt{d}) = o(1)$, wpg1,

$$\begin{aligned}\|T_0^{-1}T_5\|_2 &\leq \sqrt{d}\left\|(\frac{1}{n}\mathbb{S})^{-1}\right\|_{F,d}\frac{\lambda}{n}(\|\mathcal{S}_{S^\star_{21}}\|_2 + \|\mathcal{S}_{S^\star_{31}}\|_2) \\ &\leq \sqrt{d}2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\frac{\lambda}{n}s_1\kappa_n \leq C_2 d/4.\end{aligned}$$

Next, consider $\|\boldsymbol{\beta}_2-\boldsymbol{\beta}_1\|_2$. Since $\boldsymbol{\beta}^{(1)} \le (2C_1+C_2)d$ wpg1, the conclusion of Lemma 2 in the paper holds, which implies $\mathcal{A}_1$ occurs wpg1.

Then,

$$\boldsymbol{\beta}^{(2)} = T_0^{-1}T_1 + T_0^{-1}T_2 + T_0^{-1}T_3 + T_0^{-1}T_4(\boldsymbol{\beta}^{(1)}) + T_0^{-1}T_5,$$

where

$$T_4(\boldsymbol{\beta}^{(1)}) = \frac{1}{n}\mathbb{S}_{1,s_1}\boldsymbol{\beta}^{(1)}.$$

Thus, wpg1,

$$\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)} = \mathbb{S}^{-1}\mathbb{S}_{1,s_1}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}).$$

Thus, for $d^{3/2}s_1\kappa_n^2/n = o(1)$, wpg1,

$$\begin{aligned}\|\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)}\|_2 &\le \sqrt{d}\left\|\frac{1}{n}\mathbb{S}^{-1}\right\|_{F,d}\left\|\frac{1}{n}\mathbb{S}_{1,s_1}\right\|_F \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}\|_2 \\ &\le 2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\sqrt{d}\frac{s_1}{n}\kappa_n^2(2C_1+C_2)d \lesssim d^{3/2}s_1\kappa_n^2/n \to 0.\end{aligned}$$

Thus, wpg1, $\boldsymbol{\beta}^{(2)} = \boldsymbol{\beta}^{(1)}$, which means that, wpg1, the iteration algorithm stops at the second iteration.

For any $K \ge 1$, repeating the above arguments, with at least probability $p_{n,K} = P(\bigcap_{k=0}^K \mathcal{A}_k)$, which increases to one, we have $\boldsymbol{\beta}^{(k)} \le (2C_1+C_2)d$ for $k \le K$ and

$$\|\boldsymbol{\beta}^{(K+1)} - \boldsymbol{\beta}^{(K)}\|_2 \le (2\left\|\boldsymbol{\Sigma}_X^{-1}\right\|_{F,d}\sqrt{d}\frac{s_1}{n}\kappa_n^2)^K(2C_1+C_2)d \lesssim (\sqrt{d}s_1\kappa_n^2/n)^K d \to 0.$$

This completes the proof. □

Next result is on the consistency of the penalized two-step estimator $\tilde{\boldsymbol{\beta}}$.

Theorem D.6 (Consistency on $\tilde{\boldsymbol{\beta}}$)**.** *Suppose the assumptions and conditions of Theorem D.1 hold. If $r_d \ge 1/\sqrt{d}$, then $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$ wrt $r_d\|\cdot\|_2$.*

Proof of Theorems D.6. By Theorem D.1, $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$ wrt $r_d\|\cdot\|_2$. By Theorem 7 in the paper, $P\{\hat{I}_0 = I_0\} \to 1$ for $r_d \ge 1/\sqrt{d}$, where $I_0 = \{s_1+1, s_1+2, \ldots, s = s_1+s_2, s+1, \ldots, n\}$. Then, wpg1,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star = R_1 + R_2 + T_0^{-1}T_1 + T_0^{-1}T_2,$$

where $R_1 = (\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{X}_{\hat{I}_0})^{-1}\boldsymbol{X}_{\hat{I}_0}^T\boldsymbol{Y}_{\hat{I}_0}\{\hat{I}_0 \ne I_0\}$, $R_2 = -(\boldsymbol{X}_{I_0}^T\boldsymbol{X}_{I_0})^{-1}\boldsymbol{X}_{I_0}^T\boldsymbol{Y}_{I_0}\{\hat{I}_0 \ne I_0\}$ and T_i's are defined in the proof of Theorem D.1. Then,

$$r_d\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \le r_d\|R_1\|_2 + r_d\|R_2\|_2 + \left\|T_0^{-1}\right\|_{F,d} r_d\sqrt{d}\|T_1\|_2 + \left\|T_0^{-1}\right\|_{F,d} r_d\sqrt{d}\|T_2\|_2.$$

Since $P(\|R_1\|_{2,d} = 0) \ge P\{\hat{I}_0 = I_0\} \to 1$, we have $R_1 = o_P(1)$. Similarly, $R_2 = o_P(1)$. By the proof of Theorem D.1, $\left\|T_0^{-1}\right\|_{F,d}$ is bounded and $r_d\sqrt{d}\|T_i\|_2 \xrightarrow{P} 0$ for $i = 1, 2$. Thus, $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^\star$ wrt $r_d\|\cdot\|_2$ and $r_d \ge 1/\sqrt{d}$. □

Finally, we provide some additional results on the asymptotic distributions of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ with a different scaling. Specifically, the scaling in Section 4 is $\sqrt{n}\boldsymbol{A}_n$. Next, we consider another natural scaling $\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}$.

Theorem D.7 (Asymptotic Distribution on $\hat{\boldsymbol{\beta}}$)**.** *Suppose assumptions (D'), (D"), (E), (F) and (G) hold. If* $d^6 \log d = o(n)$, $s_1 = o(\sqrt{n}/(\lambda d\kappa_n))$ *and* $s_2 = o(\sqrt{n}/(d\kappa_n\gamma_n))$, *then*

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}).$$

Theorem D.8 (Asymptotic Distribution on $\tilde{\boldsymbol{\beta}}$)**.** *Suppose the assumptions and conditions of Theorem D.7 hold except the condition* $s_1 = o(\sqrt{n}/(\lambda d\kappa_n))$. *Then*

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}).$$

By Theorems D.7 and D.8, Wald-type confidence regions can be constructed. In order to validate these confidence regions with estimated σ and $\boldsymbol{\Sigma}_X$, we need Lemma D.6 and the following result.

Theorem D.9 (Asymptotic Distributions on $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\Sigma}}_n$)**.** *Suppose the assumptions and conditions of Theorem D.7 hold. If* $d^9(\log(d))^2 = o(n)$, *then*

$$\sqrt{n}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}).$$

Similarly, suppose the assumptions and conditions of Theorem D.8 hold. If $d^9(\log(d))^2 = o(n)$, *then*

$$\sqrt{n}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \xrightarrow{d} N(0, \sigma^2\boldsymbol{G}).$$

Remark D.1. A comparison of the assumptions and conditions of Theorem D.9 with those of Theorems D.7 and D.8 reveals that a much stronger requirement on d is needed to ensure $\hat{\boldsymbol{\Sigma}}_n$ is a good estimator of $\boldsymbol{\Sigma}_X$. Precisely, the former require that $d^9(\log(d))^2 = o(n)$ and the latter $d^6\log(d) = o(n)$. This stronger requirement on d is a price paid for estimating $\boldsymbol{\Sigma}_X$.

Remark D.2. The condition on the dimension d in Theorems 6 in the paper and 8 in the paper is $d^5\log(d) = o(n)$, slightly weaker than the condition $d^6\log(d) = o(n)$ in Theorems D.7 and D.8. Accordingly, The condition on the dimension d in Theorem D.5 is $d^8(\log(d))^2 = o(n)$, slightly weaker than the condition $d^9(\log(d))^2 = o(n)$ in Theorem D.9. This means that the scaling $\sqrt{n}\boldsymbol{A}_n$ is slightly better than the scaling $\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}$ in terms of the condition on d. Further, the former scaling is more suitable for constructing confidence regions for some entries of $\boldsymbol{\beta}^\star$.

At the end of this supplement, we provide the proofs of the above theorems.

Proof of Theorems D.7. Reuse the notations T_i's in the proof of Theorems 5 in the paper, from which,

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^\star) = V_1 + V_2 + V_3 - V_4,$$

where $V_i = \boldsymbol{B}_n T_i$ for $i = 1, 2, 3, 4$ and $\boldsymbol{B}_n = \sqrt{n}\boldsymbol{A}_n \boldsymbol{\Sigma}_X^{1/2} T_0^{-1}$. We will show $V_2 \xrightarrow{d} N(0, \sigma^2 \boldsymbol{G})$ and other V_i's are $o_P(1)$, from which the desired result follows by applying Slutsky's lemma.

On V_1. We have $\|V_1\|_2 \leq \sqrt{n} d \|\boldsymbol{A}_n\|_F \left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d} \left\|T_0^{-1}\right\|_{F,d} \|T_1\|_2$. By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is bounded. By Assumption (D"), $\left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d}$ is bounded. By Lemmas D.3 and D.4 and Assumption (D), for $d = o(n^{1/3})$, wpg1, $\left\|T_0^{-1}\right\|_{F,d}$ is bounded. Further, wpg1, $\|T_1\|_2 \leq \frac{1}{n} s_2 \kappa_n \gamma_n$. Then, $\|V_1\|_2 \lesssim \frac{1}{\sqrt{n}} s_2 d \kappa_n \gamma_n$, where $\lesssim$ means that the left side is bounded by a constant times the right side, as noted at the beginning of the appendix. Thus, $\|V_1\|_2 = o_P(1)$ for $s_2 = o(\sqrt{n}/(d\kappa_n\gamma_n))$.

On V_2. We have $V_2 = V_{21} + V_{22}$, where

$$V_{21} = \sqrt{n}\boldsymbol{A}_n \boldsymbol{\Sigma}_X^{-1/2} T_2, \quad V_{22} = \sqrt{n}\boldsymbol{A}_n \boldsymbol{\Sigma}_X^{1/2} (T_0^{-1} - \boldsymbol{\Sigma}_X^{-1}) T_2.$$

First, consider V_{21}. We have $V_{21} = \sqrt{(n-s_1)/n} \sum_{i=s_1+1}^n \boldsymbol{Z}_{n,i}$, where

$$\boldsymbol{Z}_{n,i} = \frac{1}{\sqrt{n-s_1}} \boldsymbol{A}_n \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{X}_i \epsilon_i.$$

On one hand, for every $\delta > 0$, $\sum_{i=s_1+1}^n \mathbb{E}\|\boldsymbol{Z}_{n,i}\|_2^2 \{\|\boldsymbol{Z}_{n,i}\|_2 > \delta\} \leq (n-s_1)\mathbb{E}\|\boldsymbol{Z}_{n,0}\|_2^4/\delta^2$ and

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{Z}_{n,0}\|_2^4 &= \frac{1}{(n-s_1)^2} \mathbb{E}\epsilon_0^4 \mathbb{E}(\boldsymbol{X}_0^T \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{A}_n^T \boldsymbol{A}_n \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{X}_0)^2 \\
&\leq \frac{1}{(n-s_1)^2} \mathbb{E}\epsilon_0^4 \lambda_{\max}(\boldsymbol{G}_n) \lambda_{\min}(\boldsymbol{\Sigma}_X)^{-1} \mathbb{E}(\boldsymbol{X}_0^T \boldsymbol{X}_0)^2 \\
&\leq \frac{d^2}{(n-s_1)^2} \mathbb{E}\epsilon_0^4 \lambda_{\max}(\boldsymbol{G}_n) \lambda_{\min}(\boldsymbol{\Sigma}_X)^{-1} (\frac{1}{d} \sum_{j=1}^d (\mathbb{E}X_{0j}^4)^{1/2})^2.
\end{aligned}$$

Thus, by assumptions (D'), (E) and (F), $\sum_{i=s_1+1}^n \mathbb{E}\|\boldsymbol{Z}_{n,i}\|_2^2 \{\|\boldsymbol{Z}_{n,i}\|_2 > \delta\} \to 0$ for $d = o(\sqrt{n})$. On the other hand, $\sum_{i=s_1+1}^n \mathrm{Cov}(\boldsymbol{Z}_{n,i}) = \sigma^2 \boldsymbol{A}_n \boldsymbol{A}_n^T \to \sigma^2 \boldsymbol{G}$. Thus, by central limit theorem (see, for example, Proposition 2.27 in van der Vaart (1998)), $V_{21} \xrightarrow{d} N(0, \sigma^2 \boldsymbol{G})$. Next, consider V_{22}. We have

$$\|V_{22}\|_2 \leq \|\boldsymbol{A}_n\|_F \left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d} d(\log(d))^{1/2} \left\|T_0^{-1} - \boldsymbol{\Sigma}_X^{-1}\right\|_F (d\log(d))^{-1/2} \|\sqrt{n} T_2\|_2.$$

By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is $O(1)$; By Assumption (D"), $\left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d}$ is $O(1)$; by Lemmas D.3 and D.4, $d(\log(d))^{1/2} \left\|T_0^{-1} - \boldsymbol{\Sigma}_X^{-1}\right\|_F$ is $o_P(1)$ for $d^6 \log(d) = o(n)$; By Lemma D.5, together with Assumption (G), $(d\log(d))^{-1/2} \|\sqrt{n} T_2\|_2 = (d\log(d))^{-1/2} \|\frac{1}{\sqrt{n}} \mathbb{S}_{s_1+1,n}^{\epsilon}\|_2$ is $O_P(1)$ for $d = o(\sqrt{n})$. Thus, $V_{22} \xrightarrow{P} 0$. By slutsky's lemma, $V_2 \xrightarrow{d} N(0, \sigma^2 \boldsymbol{G})$.

On V_3 and V_4. First consider V_3. By noting that $s_1 = o(\sqrt{n}/(\lambda d \kappa_n))$, wpg1, $\|V_3\|_2 \leq d\sqrt{n} \|\boldsymbol{A}_n\|_F \left\|\boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d} \left\|T_0^{-1}\right\|_{F,d} \|T_3\|_2 \lesssim d\lambda s_1 \kappa_n / \sqrt{n} \to 0$. Thus, $\|V_3\|_2 = o_P(1)$. In the same way, $\|V_4\|_2 = o_P(1)$. This completes the proof. □

Proof of Theorem D.8. From the proof of Theorem D.6, we have $\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star) = \tilde{R}_1+\tilde{R}_2+V_1+V_2$, where $\tilde{R}_1 = \sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}R_1$, $\tilde{R}_2 = \sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}R_2$, and R_i's and V_i's are defined in the proofs of Theorems D.6 and D.7. Since $P(\|\tilde{R}_1\|_2 = 0) \geq P\{\hat{I}_0 = I_0\} \to 1$, we have $\tilde{R}_1 = o_P(1)$. Similarly, $\tilde{R}_2 = o_P(1)$. By the proof of Theorem D.7, $V_1 = o_P(1)$ and $V_2 \xrightarrow{d} N(0, \sigma^2\boldsymbol{G})$. Thus, the asymptotic distribution of $\tilde{\boldsymbol{\beta}}$ is Gaussian by Slutsky's lemma. □

Proof of Theorem D.9. We only show the result on $\hat{\boldsymbol{\beta}}$. since the result on $\tilde{\boldsymbol{\beta}}$ can be obtained in a similar way. We reuse the definitions of T_i's in the proof of Theorems 5 in the paper, from which,

$$\sqrt{n}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_n^{1/2}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^\star) = M+R,$$

where $M = \sqrt{n}\boldsymbol{A}_n\boldsymbol{\Sigma}_X^{1/2}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^\star)$ and $R = \sqrt{n}\boldsymbol{A}_n(\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2})(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^\star)$. By Theorem D.7, $M \xrightarrow{d} N(0,\sigma^2\boldsymbol{G})$. Then, it is sufficient to show that $R \xrightarrow{P} 0$ wrt $\|\cdot\|_2$. We have

$$R = R_1+R_2+R_3-R_4,$$

where $R_i = \boldsymbol{B}_nT_i$ for $i = 1,2,3,4$ and $\boldsymbol{B}_n = \sqrt{n}\boldsymbol{A}_n(\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2})T_0^{-1}$. We will show each R_i converges to zero in probability, which finishes the proof.

On R_1. By Lemma D.7, $\left\|\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2}\right\|_F \leq (d^{1/2}\left\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_X\right\|_F)^{1/2}$. Then,

$$\begin{aligned}\|R_1\|_2 &\leq \sqrt{n}\,\|\boldsymbol{A}_n\|_F\left\|\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2}\right\|_F\left\|T_0^{-1}\right\|_F\|T_1\|_2\\ &\leq \sqrt{n}d\,\|\boldsymbol{A}_n\|_F\,(\left\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_X\right\|_{F,d})^{1/2}\left\|T_0^{-1}\right\|_{F,d}\|T_1\|_2.\end{aligned}$$

By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is bounded. By Lemma D.4, $\left\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_X\right\|_{F,d} = o_P(1)$ for $d = o(n^{1/3})$. By Lemmas D.3 and D.4 and Assumption (D), for $d = o(n^{1/3})$, wpg1, $\left\|T_0^{-1}\right\|_{F,d}$ is bounded. We have, wpg1, $\|T_1\|_2 \leq \frac{1}{n}s_2\kappa_n\gamma_n$. Then, $\|R_1\|_2 \lesssim \frac{1}{\sqrt{n}}s_2d\kappa_n\gamma_n$. Thus, $\|R_1\|_2 = o_P(1)$ for $s_2 = o(\sqrt{n}/(d\kappa_n\gamma_n))$.

On R_2. We have

$$\|R_2\|_2 \leq \|\boldsymbol{A}_n\|_F\,d(\log(d))^{1/2}\left\|\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2}\right\|_F\left\|T_0^{-1}\right\|_{F,d}(d\log(d))^{-1/2}\|\sqrt{n}T_2\|_2,$$

and

$$d(\log(d))^{1/2}\left\|\hat{\boldsymbol{\Sigma}}_n^{1/2}-\boldsymbol{\Sigma}_X^{1/2}\right\|_F \leq (d^{5/2}\log(d)\left\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_X\right\|_F)^{1/2}.$$

By Assumption (F), $\|\boldsymbol{A}_n\|_F$ is $O(1)$; by Lemma D.4, $d^{5/2}\log(d)\left\|\hat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}_X\right\|_F = o_P(1)$ for $d^9(\log(d))^2 = o(n)$; by Lemmas D.3 and D.4, $d(\log(d))^{1/2}\left\|T_0^{-1}-\boldsymbol{\Sigma}_X^{-1}\right\|_F$ is $o_P(1)$ for $d^6\log(d) = o(n)$; by Lemma D.5, $(d\log(d))^{-1/2}\|\sqrt{n}T_2\|_2 = (d\log(d))^{-1/2}\|\frac{1}{\sqrt{n}}\mathbb{S}^\epsilon_{s_1+1,n}\|_2$ is $O_P(1)$ for $d = o(\sqrt{n})$. Thus, $R_2 \xrightarrow{P} 0$.

On R_3 and R_4. First consider R_3. By noting that $s_1 = o(\sqrt{n}/(\lambda d \kappa_n))$, wpg1,

$$\|R_3\|_2 \leq d\sqrt{n}\,\|\boldsymbol{A}_n\|_F\,(\left\|\hat{\boldsymbol{\Sigma}}_n^{1/2} - \boldsymbol{\Sigma}_X^{1/2}\right\|_{F,d})^{1/2}\,\left\|T_0^{-1}\right\|_{F,d}\|T_3\|_2 \lesssim d\lambda s_1 \kappa_n/\sqrt{n} \to 0.$$

Thus, $\|R_3\|_2 = o_P(1)$. In the same way, $\|R_4\|_2 = o_P(1)$. □

References

Fan, J., Liao, Y. & Mincheva, M. (2011), 'High-dimensional covariance matrix estimation in approximate factor models', *Ann. Statist.* **39**(6), 3320–3356.

Fan, J. & Peng, H. (2004), 'On non-concave penalized likelihood with diverging number of parameters', *The Annals of Statistics* **32**, 928–961.

Jahn, J. (2007), *Introduction to the Theory of Nonlinear Optimization*, Springer Berlin Heidelberg.

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer New York.

Neyman, J. & Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica* **16**, 1–32.

Shiryaev, A. N. (1995), *Probability*, second edn, Springer-Verlag.

Stewart, G. W. (1969), 'On the continuity of the generalized inverse', *SIAM Journal on Applied Mathematics* **17**, 33–45.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Wihler, T. P. (2009), 'On the holder continuity of matrix functions for normal matrices', *Journal of inequalities in pure and applied mathematics* **10**.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *The Journal of Machine Learning Research* **7(Nov)**, 2541–2563.