Nonconvex Penalised Regression and Post-Selection Least Squares Estimation under High Dimensions: a Local Asymptotic Perspective

Institute of Applied Mathematics, Shenzhen Polytechnic University

Department of Statistics and Actuarial Science, The University of Hong Kong

Supplementary Material

This supplement contains Appendix 1 and Appendix 2.

In Appendix 1, we generalise additional theoretical results and properties of adjusted effects estimation by including an additional heavy-tailed setting. The simulation results are included in Appendix 1.

All technical proofs are provided in Appendix 2.

S1 Appendix 1

S.1.1 Theory under more general tail conditions

In addition to the common sub-Gaussian and moderately heavy tail conditions typically assumed under high dimensions, we also consider a heavy-tailed setting $(\mathcal{T}_3)$ exemplified by an infinite variance and a kind of power law for the random error ϵ:

$(\mathcal{T}_3)$ $\mathbb{P}(|\epsilon| > t) \asymp t^{-\gamma}$ and $\mathbb{P}(|X^{(j)} - \mathbb{E}X^{(j)}| > t) \preceq t^{-2\gamma}$ as $t \to \infty$, for some $\gamma \in (1, 2)$.

The results provide fresh insights into the properties and applicability of nonconvex

penalised methods and their post-selection least squares estimators in a broader context. We shall generalise our theory in the main text by including $(\mathcal{T}_3)$. All technical proofs of the generalised version, which covers also the tail conditions considered in the main text, are given in Appendix 2.

Recalling that $\epsilon_i = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0$, we redefine $\boldsymbol{W} = [W_1, \ldots, W_p]^\top = T^{-1/2} \sum_{i=1}^n \epsilon_i \boldsymbol{X}_i$, where T is a scaling factor depending on n and rendering $\boldsymbol{W} \asymp 1$. Under tail conditions $(\mathcal{T}_1)$ or $(\mathcal{T}_2)$, we have $T = n$, with W_j converging in distribution to a normal random variable by the Central Limit Theorem. Under $(\mathcal{T}_3)$, we have $T = n^{2/\gamma}$ and that W_j has an asymptotically stable distribution by the Stable Central Limit Theorem (Hoffmann-Jørgensen, 1994, Section 5.25).

Remark S.1. *We generalise (A2) by assuming* $\lim_{n\to\infty} \dfrac{\lambda}{n(\alpha-1)\kappa} < 2\underline{\mathcal{C}}$ *and* $\sqrt{T} \prec \lambda \prec n$. *Note that for* $\lambda \succ \sqrt{T}$, *the signals captured by the index set* $\mathcal{A}_0$ *are not confounded with sampling noise.*

Remark S.2. *Assume the conditions (C1), (A1), (A3), (A4) and the generalised (A2). Lemma 1, Proposition 1 and Theorem 2 also hold under* $(\mathcal{T}_3)$ *and* $p \prec \lambda^\gamma / n$. *Thus, the growth of the dimension* p *cannot exceed a polynomial rate if* $(\boldsymbol{X}, Y)$ *has heavy tails with index* $\gamma \in (1, 2)$.

After introducing the new notations T, $\boldsymbol{W}$ and the generalised conditions, we extend our theorems and corollaries on the componentwise convergence rates and weak limits of sparse and consistent local minima $\hat{\boldsymbol{\beta}}$'s to accommodate the heavy tail

condition ($\mathcal{T}_3$). In the main text, theoretical statements and remarks not related to T carry over into the generalised version without change. They are therefore omitted hereafter for brevity.

Theorem S.1. *Suppose that* $|\mathcal{A}_0| \asymp 1$ *and the generalised conditions of Proposition 1 hold. Then, a consistent sparse local minimum* $\hat{\boldsymbol{\beta}}$ *exists with a componentwise convergence rate* $r_0 \succ n/\lambda$ *if and only if* $\psi \prec \lambda/n$. *In this case, we have* $r_0 \asymp (n/\sqrt{T}) \wedge \psi^{-1}$ *and, with probability converging to one,* $supp(\hat{\boldsymbol{\beta}}) = \mathcal{A}_0$ *and*

$$\begin{cases} \hat{\boldsymbol{\beta}}^{\mathcal{A}_0^c} = \mathbf{0}, \\ \hat{\boldsymbol{\beta}}^{\mathcal{A}_0} = \boldsymbol{\beta}_0^{\mathcal{A}_0} + \Big\{ \hat{C}_{\mathcal{A}_0\mathcal{A}_0} - \dfrac{\lambda}{2n(\alpha-1)\kappa} \Lambda_{\mathcal{A}_0\mathcal{A}_0} \Big\}^{-1} \Big(\dfrac{\sqrt{T}}{n} \boldsymbol{W}^{\mathcal{A}_0} + \hat{C}_{\mathcal{A}_0\mathcal{A}_0^c} \boldsymbol{\beta}_0^{\mathcal{A}_0^c} - \dfrac{\phi^{\mathcal{A}_0}}{r_0} \Big), \end{cases} \tag{S1.1}$$

where

$$\Lambda = \operatorname{diag}\big(\mathbf{1}\{ |\beta_0^{(j)}| < \alpha\kappa \} : j = 1, \ldots, p \big),$$
$$\phi = \frac{r_0 \lambda \, \alpha}{2n(\alpha-1)} \Big[\Big(1 - \frac{|\beta_0^{(j)}|}{\alpha\kappa} \Big)_+ \operatorname{sgn}(\beta_0^{(j)}) : j = 1, \ldots, p \Big].$$

If, in addition, $B_U \succ \lambda/n$, *then the above local minima are the only consistent sparse local minima selecting* $\mathcal{A}_0$ *with probability converging to one, while any other consistent sparse local minima converge at a slower rate* n/λ *and are less sparse with supports* $\supsetneq \mathcal{A}_0$ *asymptotically.*

If, in addition, $\lim_{n\to\infty} B_U/(\alpha\kappa) > 1$, *then the above minima yield an objective function (2.1) strictly smaller than that yielded by any other consistent sparse local*

minima.

Remark S.3. *Theorem S.1 suggests that if the group of strong signals is sufficiently distinct from the group of weak signals such that* $\psi \prec \lambda/n$*, then the set of consistent sparse local minima can be classified into two subsets, having a fast convergence rate* $(n/\sqrt{T}) \wedge \psi^{-1}$ *and a slow convergence rate* n/λ*, respectively, with the former being non-empty. With large probability, the fast converging local minima select the strong signal set* $\mathcal{A}_0$*, while the slow converging local minima, if any, select signal sets containing* $\mathcal{A}_0 \setminus \{j : |\beta_0^{(j)}| \asymp \lambda/n\}$*.*

Remark S.4. *Similar to our Proposition 1, Loh and Wainwright (2015) establish a conservative slow rate* n/λ *for the convergence of every local minimum as a statistical guarantee for the latter. Our results in Theorem S.1 clarify the conditions for the existence of local minima converging at a faster rate* $(n/\sqrt{T}) \wedge \psi^{-1}$*.*

Remark S.5. *If* $\alpha\kappa \succ \lambda/n$*, which excludes the common SCAD and MCP methods, then* $\psi \prec \lambda/n$ *implies* $B_U \succ \lambda/n$*, and* $\hat{\boldsymbol{\beta}}^{\mathcal{A}_0}$ *in (S1.1) reduces to* $\boldsymbol{\beta}_0^{\mathcal{A}_0} + \hat{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\big\{(\sqrt{T}/n)\boldsymbol{W}^{\mathcal{A}_0} + \hat{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} - r_0^{-1}\phi^{\mathcal{A}_0}\big\}$*.*

Remark S.6. *Theorem S.1 provides a more holistic picture of the selection and convergence properties of consistent sparse local minima* $\hat{\boldsymbol{\beta}}$ *from a local asymptotic perspective, which covers as a special case the weak oracle property introduced by Fan and Lv (2011), under weaker conditions on* B_0*,* B_U *and the covariate design than*

those assumed by the said paper. In particular, we see from (S1.1) that even when $\hat{\boldsymbol{\beta}}$ *selects* $\mathcal{A}_0$ *correctly and converges at a fast rate, it is not necessarily asymptotically equivalent to an ordinary least squares (OLS) estimator derived from* $\mathcal{A}_0$*, casting doubt on our conventional interpretation of oracle properties of nonconvex penalised estimators.*

Remark S.7. *In the special case* $\alpha\kappa = \infty$*, which holds for LASSO, the condition* $\psi \prec \lambda/n$ *fails. A contraposition of Theorem S.1 shows that the active* $\hat{\beta}_j$*'s have a slow convergence rate* n/λ*. A faster convergence rate may result under a fixed* p *if we set* $\lambda \asymp \sqrt{T}$*, as is typically adopted by LASSO. However, the latter condition fails to guarantee selection consistency in general, unless we impose further constraints on* $\mathcal{C}$*.*

We may deduce from (S1.1) a series of phase changes of the asymptotic behaviour of $U^{\mathcal{A}_0} = r_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$, when signal patterns undergo the following transition phases over the local asymptotic spectrum.

(a) If $\psi = (\lambda/n)\big(1 - B_U/(\alpha\kappa)\big)_+ \succ B_0$, setting $r_0 = (n/\sqrt{T})\big\{1 \wedge (\sqrt{T}/\lambda)\big(1 - B_U/(\alpha\kappa)\big)_+^{-1}\big\}$ gives

$$U^{\mathcal{A}_0} = \Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1}\Big[\big\{1\wedge(\sqrt{T}/\lambda)\big(1-B_U/(\alpha\kappa)\big)_+^{-1}\big\}\boldsymbol{W}^{\mathcal{A}_0} - \phi^{\mathcal{A}_0}\Big] + o_p(1),$$

which has a non-random leading term

$$-\Big\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\Big\}^{-1}\phi^{\mathcal{A}_0}$$

if and only if $\left(1 - B_U/(\alpha\kappa)\right)_+ \succ \sqrt{T}/\lambda$.

(b) If $\psi = B_0 \succ (\lambda/n)\left(1 - B_U/(\alpha\kappa)\right)_+$, setting $r_0 = (n/\sqrt{T}) \wedge B_0^{-1}$ gives

$$U^{\mathcal{A}_0} = \left\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\right\}^{-1}\left(1 \wedge \frac{\sqrt{T}}{nB_0}\right)\left\{\boldsymbol{W}^{\mathcal{A}_0} + (n/\sqrt{T})\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\right\} + o_p(1),$$

which has a non-random leading term

$$\left\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\right\}^{-1} B_0^{-1}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}$$

if and only if $B_0 \succ \sqrt{T}/n$, or reduces to

$$\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\left(1 \wedge \frac{\sqrt{T}}{nB_0}\right)\left\{\boldsymbol{W}^{\mathcal{A}_0} + (n/\sqrt{T})\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\right\} + o_p(1)$$

if $B_U \geq \alpha\kappa$.

(c) If $\psi \prec \sqrt{T}/n$, setting $r_0 = n/\sqrt{T}$ gives

$$U^{\mathcal{A}_0} = \left\{\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0} - \frac{\lambda}{2n(\alpha-1)\kappa}\Lambda_{\mathcal{A}_0\mathcal{A}_0}\right\}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1),$$

which has a random leading term.

Given its prominence in the literature, the conventional oracle property, generalised under our local asymptotic framework, is given below as an immediate corollary to Theorem S.1, which is obtained by intersecting the above phases (b) and (c) into a final oracle phase.

Corollary S.1. *(Generalised oracle property) Assume the generalised conditions of Proposition 1, $|\mathcal{A}_0| \asymp 1$, $\psi \prec \sqrt{T}/n$, and that either $\lambda/n \prec \alpha\kappa$ or $B_U \geq \alpha\kappa$. Then a*

generalised oracle estimator $\hat{\boldsymbol{\beta}}_{go}$ *exists with*

$$\mathbb{P}\big(\hat{\boldsymbol{\beta}}_{go}^{\mathcal{A}_0^c} = 0\big) \to 1 \quad \textit{and} \quad nT^{-1/2}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1).$$

If, in addition, $B_U \succ \lambda/n$*, then* $\hat{\boldsymbol{\beta}}_{go}$ *is the only consistent sparse local minimum selecting* $\mathcal{A}_0$ *with probability converging to one. Any other consistent sparse local minima necessarily converge at a slower rate* n/λ *and are less sparse with supports* $\hat{\mathcal{A}} \supsetneq \mathcal{A}_0$ *asymptotically.*

Remark S.8. *The generalised oracle estimator* $\hat{\boldsymbol{\beta}}_{go}$ *estimates the coefficients of weak signals (in* $\mathcal{A}_0^c$*) to be zero and those of strong signals (in* $\mathcal{A}_0$*) by ordinary least squares. If* $(\boldsymbol{X}, Y)$ *satisfies tail conditions* ($\mathcal{T}_1$) *or* ($\mathcal{T}_2$)*, then* $n^{1/2}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)^{\mathcal{A}_0}$ *is asymptotically zero-mean Gaussian. On the other hand, if the tail condition* ($\mathcal{T}_3$) *holds with tail index* $\gamma \in (1, 2)$*, then an application of the Stable Central Limit Theorem to* $\boldsymbol{W}^{\mathcal{A}_0}$ *implies that for each* $j \in \mathcal{A}_0$*, the* j*-th component of* $n^{1-1/\gamma}(\hat{\boldsymbol{\beta}}_{go} - \boldsymbol{\beta}_0)$ *converges weakly to a linear combination of random variables distributed under a stable law.*

Remark S.9. *Loh and Wainwright (2017) show, under the tail condition* ($\mathcal{T}_1$)*, a sparse Riesz condition on* $\mathcal{C}$ *and a betamin condition* $B_0 = 0$*, that* $\hat{\boldsymbol{\beta}}_{go}$ *is the unique local, hence global, minimum. Assuming a weaker bound* $|\mathcal{A}_0| \preceq n/\lambda$ *than ours, they establish a conservative convergence rate of order* n/λ *for* $\hat{\boldsymbol{\beta}}_{go}$*, which is slower than the rate* $n/\sqrt{T}$ *shown in Corollary 1.*

As in Section 2.4 of the main text, the above established theory can be illustrated

using a schematic diagram similar to Figure 1, with the order $n^{-1/2}$ replaced by $\sqrt{T}/n$.

S.1.2 Post-selection OLS estimator

We first generalise the subgradient conditions (3.1, 3.2) to accommodate the heavy tail condition ($\mathcal{T}_3$):

$$\begin{cases} (2n/\lambda)\big|n^{-1}T^{1/2}W_j - \hat{C}_{\{j\}\hat{\mathcal{A}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\hat{\mathcal{A}}} + \hat{C}_{\{j\}\hat{\mathcal{A}}^c}\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}\big| \leq 1, \ j \in \hat{\mathcal{A}}^c & \text{(S1.2)} \\ \hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big)^{\hat{\mathcal{A}}} = n^{-1}T^{1/2}\boldsymbol{W}^{\hat{\mathcal{A}}} + \mathscr{B}_1 + \mathscr{B}_2, & \text{(S1.3)} \end{cases}$$

where $\mathscr{B}_1 = -(2n)^{-1}\lambda\big[\mathrm{sgn}(\hat{\beta}_j)q'(|\hat{\beta}_j|/\kappa) : j \in \hat{\mathcal{A}}\big]$ and $\mathscr{B}_2 = \hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}^c}\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}$. Recall that $\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) = n^{-1}\hat{C}^{-1}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}\sum_{i=1}^n Y_i\boldsymbol{X}_i^{\hat{\mathcal{A}}}$ is the sparse OLS estimator restricted to the submodel containing only variables in $\hat{\mathcal{A}}$. That $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ satisfies

$$\hat{C}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}\big\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\big\}^{\hat{\mathcal{A}}} = n^{-1}T^{1/2}\boldsymbol{W}^{\hat{\mathcal{A}}} + \mathscr{B}_2 \tag{S1.4}$$

suggests that the bias term $\mathscr{B}_1$ is removed from (S1.3) by post-selection OLS. We now detail the convergence properties of $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ as an estimator of $\boldsymbol{\beta}_0$, under mild conditions on design and signal strength and either one of the tail conditions ($\mathcal{T}_1$), ($\mathcal{T}_2$) or ($\mathcal{T}_3$).

Theorem S.2. *Suppose that* $|\mathcal{A}_0| \preceq 1$ *and the generalised conditions of Proposition 1 hold. If* $\psi \prec \lambda/n$, *then there exists a consistent sparse post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *which is supported on* $\mathcal{A}_0$ *with probability converging to one and converges at a*

rate $B_0^{-1} \wedge (n/\sqrt{T})$.

If, in addition, $B_U \succ \lambda/n$, *then any consistent sparse post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *is supported on* $\hat{\mathcal{A}} \supseteq \mathcal{A}_0$ *with probability converging to one and converges at a rate within the range* $\left[B_0^{-1} \wedge (n/\sqrt{T}),\ n/\sqrt{T}\right]$.

Remark S.10. *As has been discussed in Remark S.3, under the condition* $\psi \prec \lambda/n \prec B_U$, *a consistent sparse local minimum* $\hat{\boldsymbol{\beta}}$ *converges either at a fast rate* $(n/\sqrt{T}) \wedge \psi^{-1}$ *with selected set* $\mathcal{A}_0$ *or at a slow rate* n/λ *with selected set* $\supsetneq \mathcal{A}_0$. *Under the same signal pattern, any post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *has a convergence rate* $\succeq B_0^{-1} \wedge (n/\sqrt{T})$, *which is at least as fast as that of any fast-converging* $\hat{\boldsymbol{\beta}}$ *and strictly faster than the rate of any slowly-converging* $\hat{\boldsymbol{\beta}}$. *If, in addition,* $B_0 \vee (\sqrt{T}/n) \prec (\lambda/n)\{1 - B_U/(\alpha\kappa)\}_+$, *then any post-selection OLS estimator converges strictly faster than any local minimum* $\hat{\boldsymbol{\beta}}$, *provided they are consistent and sparse.*

Remark S.11. *In general, if* $|\mathcal{A}_0| \preceq 1$, $B_U \succ \lambda/n$ *and* $B_0 \prec \sqrt{T}/n$, *we have, for any* $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ *and* $\mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$ *with* $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$, *that* $\mathcal{A} \supseteq \mathcal{A}_0$ *and that the corresponding post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *converges at the fastest possible rate* $n/\sqrt{T}$ *and satisfies* $nT^{-1/2}\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\}^{\mathcal{A}} = \mathcal{C}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{W}^{\mathcal{A}} + o_p(1)$.

Corollary S.2. *(Generalised oracle property) Assume the generalised conditions of Proposition 1,* $|\mathcal{A}_0| \preceq 1$, $\{1 - B_U/(\alpha\kappa)\}_+ \prec 1$ *and* $B_0 \prec \sqrt{T}/n$. *Then, a sequence of*

selected sets $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ *exists such that*

$$\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}_0) \to 1 \quad \textit{and} \quad nT^{-1/2}\big\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\beta}_0\big\}^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1).$$

Remark S.12. *As has been shown in Corollary S.1, existence of a generalised oracle* $\hat{\boldsymbol{\beta}}_{go}$ *requires more restrictive conditions on* B_U*, namely* $\{1 - B_U/(\alpha\kappa)\}_+ \prec \sqrt{T}/\lambda$ *if* $\alpha\kappa \succ \lambda/n$ *or* $B_U \geq \alpha\kappa$ *if* $\alpha\kappa \asymp \lambda/n$*, compared to those required by Corollary S.2. If, in addition,* $B_U \succ \lambda/n$*, then all the post-selection OLS estimators converge at the fastest rate* $n/\sqrt{T}$*, while the corresponding local minima* $\hat{\boldsymbol{\beta}}$ *except* $\hat{\boldsymbol{\beta}}_{go}$ *all converge at the slowest rate* n/λ*.*

By removing the bias term $\mathscr{B}_1$, the post-selection OLS estimators $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ acquire convergence properties more desirable than the local minima $\hat{\boldsymbol{\beta}}$ and, in the case of multiple solutions to the nonconvex optimisation program (2.2), ratewise more robust against the choice of strong signal sets $\hat{\mathcal{A}}$.

S.1.3 Estimation of adjusted effects

From a predictive perspective, it may be of interest to draw inference about the effects of strong signals after adjusting for the omission of weak signals under a weakly sparse model. More specifically, define an "oracle" target to be

$$\boldsymbol{\theta}_0 = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p}\big\{\mathbb{E}(Y - \boldsymbol{X}^\top\boldsymbol{\theta})^2 : \boldsymbol{\theta}^{\mathcal{A}_0^c} = \boldsymbol{0}\big\},$$

so that $\boldsymbol{\theta}_0^{\mathcal{A}_0^c} = \mathbf{0}$ and $\boldsymbol{\theta}_0^{\mathcal{A}_0} = \boldsymbol{\beta}_0^{\mathcal{A}_0} + \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}$, which can be interpreted as the effects of strong signals in $\mathcal{A}_0$ adjusted for the omission of weak signals in $\mathcal{A}_0^c$. In a similar vein, Bühlmann and Van De Geer (2011, sections 6.2.3–4) define an "oracle" active set S_0 that depends on the design matrix and λ, and set as their target for estimation the coefficients which provide the best linear fit using only variables in S_0. An empirical version of the above framework under a fixed design is also considered by Van de Geer et al. (2011). Compared to S_0, which may trade off some strong signals against a desired design condition, our choice of oracle active set $\mathcal{A}_0$ consists of all, and only, strong signals $\succeq \lambda/n$ and appears more natural. In particular, if the strong signals are sufficiently sparse such that $\left|\{j : |\beta_0^{(j)}| \succ \sqrt{T}/n\}\right| \preceq 1$, then setting λ close to $\sqrt{T}$ ensures that the sparse oracle target $\boldsymbol{\theta}_0$ approximates $\boldsymbol{\beta}_0$ well and provides reliable assessments of the effects of all strong signals which are not confounded with sampling noise.

Noting that the bias term $\mathscr{B}_2$ is asymptotically equivalent to $\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}$ if $\hat{\mathcal{A}} = \mathcal{A}_0$, it follows from (S1.3) that switching $\boldsymbol{\beta}_0$ to the oracle target $\boldsymbol{\theta}_0$ may help offset the bias due to $\mathscr{B}_2$. The following two theorems state the generalised oracle properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ as estimators of $\boldsymbol{\theta}_0$, respectively.

Theorem S.3. *Assume the generalised conditions of Proposition 1,* $|\mathcal{A}_0| \preceq 1$, $B_0 \prec \lambda/n$, *and that either* $\sqrt{T}/(n\alpha\kappa) \vee \{1 - B_U/(\alpha\kappa)\}_+ \prec \sqrt{T}/\lambda$ *or* $B_U \geq \alpha\kappa$. *Then,*

there exists a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ *satisfying* $\mathbb{P}\big(\hat{\boldsymbol{\beta}}^{\mathcal{A}_0^c} = \mathbf{0}\big) \to 1$ *and*

$$nT^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1).$$

Theorem S.4. *Assume the generalised conditions of Proposition 1,* $|\mathcal{A}_0| \preceq 1$ *and* $\psi \prec \lambda/n$*. Then, there exists a sequence of selected sets* $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ *with* $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}_0) \to 1$ *such that*

$$nT^{-1/2}\big\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\theta}_0\big\}^{\mathcal{A}_0} = \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1).$$

Remark S.13. *If the local minimum* $\hat{\boldsymbol{\beta}}$ *or the post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *is taken to estimate* $\boldsymbol{\theta}_0$ *rather than* $\boldsymbol{\beta}_0$*, then the condition on* B_0 *for the generalised oracle properties stated in Corollaries S.1 and S.2 can be weakened from* $B_0 \prec \sqrt{T}/n$ *to* $B_0 \prec \lambda/n$*.*

Remark S.14. *If the signals in the inactive set* $\mathcal{A}_0^c$ *are sufficiently weak such that* $B_0 \prec \sqrt{T}/n$*, then inference drawn about* $\boldsymbol{\theta}_0$*, based on either* $\hat{\boldsymbol{\beta}}$ *or* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$*, is equivalent to inference drawn about* $\boldsymbol{\beta}_0$ *to the first order.*

Remark S.15. *If we assume further that* $B_U \succ \lambda/n$*, then we can deduce from Theorem S.1 and* (S1.4) *that any consistent sparse local minimum* $\hat{\boldsymbol{\beta}}$ *with* $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ *gives rise to a post-selection OLS estimator* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *which, with probability converging to one, has support containing* $\mathcal{A}_0$ *and satisfies* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})^{\mathcal{A}^c} = \boldsymbol{\theta}_0^{\mathcal{A}^c} = \mathbf{0}$ *and*

$$\hat{\boldsymbol{b}}(\hat{\mathcal{A}})^{\mathcal{A}} = \boldsymbol{\theta}_0^{\mathcal{A}} + n^{-1}T^{1/2}\hat{C}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{W}^{\mathcal{A}} + \hat{C}_{\mathcal{A}\mathcal{A}}^{-1}\hat{C}_{\mathcal{A}\mathcal{A}^c}\boldsymbol{\beta}_0^{\mathcal{A}^c} - \begin{bmatrix} \mathcal{C}_{\mathcal{A}_0\mathcal{A}_0}^{-1}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} \\ -\boldsymbol{\beta}_0^{\mathcal{A}\setminus\mathcal{A}_0} \end{bmatrix}.$$

We see from the above expansion that the estimation error $\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\theta}_0\}^{\mathcal{A}}$ *converges to* $\mathbf{0}$ *at a rate* $(n/\sqrt{T}) \wedge \left(\mathbf{1}\{\mathcal{A}_0 \neq \mathcal{A}\}B_0\right)^{-1}$ *in general. It follows that the fastest rate* $n/\sqrt{T}$ *is achieved by the generalised oracle* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ *supported on* $\mathcal{A}_0$ *or, if* $B_0 \prec \sqrt{T}/n$, *by any* $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$.

S.1.4 Simulation study

Setting

We select SCAD, a popular nonconvex penalisation method, for investigation in our simulation studies, under both Gaussian and heavy-tailed settings. When SCAD performs well in both selection and estimation, as would have been expected under pattern 5 in Figure 1, the bias would be small and the post-SCAD OLS method may not make a significant improvement. To better illustrate the effects of post-selection OLS, we focus in our studies on cases where SCAD estimators are likely to succumb to large or moderate biases, which are by no means uncommon.

Since our assumptions allow all regression coefficients to be nonzero, hence $B_0 > 0$, and the oracle active set $\mathcal{A}_0$ is defined in terms of asymptotic orders, there does not exist a definitive demarcation between $\mathcal{A}_0$ and $\mathcal{A}_0^c$ based on numerical values of the $\beta_0^{(j)}$'s. As a finite-sample benchmark for evaluating empirical performance of different estimators, we fix in our simulations $\mathcal{A}_0 = \text{supp}(\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ is a solution for the penalised parameter satisfying (2.3), obtained by the optim function using

the "BFGS" method in the R package stats. The number of simulated replications in each example is denoted by m. The focus of this study is on not so much the choice of optimal tuning parameters as the changes in performances of SCAD and post-SCAD OLS across different signal patterns. We fix the tuning parameters at values consistent with the conditions set in our theoretical investigation. In particular, we set in the R package ncvreg $\alpha = 3.7$, the default value advised by Fan and Li (2001), and the SCAD penalty weight $\lambda_{\mathsf{R}} = \lambda/(2n)$ to be a fixed value by reference to cross-validation outcomes of a few trial runs. The tail conditions $(\mathcal{T}_1)$–$(\mathcal{T}_3)$ on $(\boldsymbol{X}, Y)$ cover both the regular sub-Gaussian case and the heavy-tailed power law. The conditions of Lemma 1 imply $\log p \prec n$, $p \prec n$ and $p \prec n^{\gamma-1}$ $(1 < \gamma < 2)$ under $(\mathcal{T}_1)$, $(\mathcal{T}_2)$ and $(\mathcal{T}_3)$, respectively, which inform our settings of (n, p) in the simulation study.

Measures of estimation accuracy

As a benchmark for SCAD and post-SCAD OLS, an oracle OLS estimator is calculated to be $\hat{\boldsymbol{b}}(\mathcal{A}_0)$, assuming prior knowledge of the correct model. It is denoted as Oracle in the tables of results.

Denote by $\tilde{\boldsymbol{\beta}}$ a generic estimator of $\boldsymbol{\beta}_0$. We now describe the indicators used for evaluating the performance of $\tilde{\boldsymbol{\beta}}$. The estimation error $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ for the entire signal vector is summarised into the total bias, Bias $= \left\| m^{-1} \sum_{i=1}^{m} \tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0 \right\|_2$, and the total mean squared error, MSE $= m^{-1} \sum_{i=1}^{m} \|\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0\|_2^2$, where $\tilde{\boldsymbol{\beta}}^{*i}$ denotes the replicate

of $\tilde{\boldsymbol{\beta}}$ obtained in the i^{th} simulation. Under the heavy-tailed settings, we report in lieu of MSE the total mean absolute deviation, MAD $= m^{-1}\sum_{i=1}^{m}\|\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0\|_1$, which is more robust against outliers. Replacing $\boldsymbol{\beta}_0$ by the "oracle" target $\boldsymbol{\theta}_0$ (Section S.1.3), we also calculate the adjusted total bias PBias $= \left\|m^{-1}\sum_{i=1}^{m}\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\theta}_0\right\|_2$ for SCAD and post-SCAD OLS. To save space we do not report PBias for Oracle, which has zero bias against $\boldsymbol{\theta}_0$ by definition. Similarly, we do not report the MSE figures with respect to $\boldsymbol{\theta}_0$, for the change of target from $\boldsymbol{\beta}_0$ to $\boldsymbol{\theta}_0$ does not affect the variance of $\tilde{\boldsymbol{\beta}}$.

The performance of post-SCAD OLS in estimation is necessarily intertwined with that of SCAD in variable selection, which may vary considerably across different signal patterns and correlation structures of $\boldsymbol{X}$ under the same λ_{R}. For more insights into the effects of post-SCAD OLS, we also calculate a second group of indicators, denoted respectively by CS Bias, CS MSE (or CS MAD under heavy-tailed settings) and CS PBias, which compare the errors between SCAD and post-SCAD OLS estimators calculated only over simulation runs where the method correctly selected the oracle active set $\mathcal{A}_0$. Specifically, we have CS Bias $= \left\||S|^{-1}\sum_{i\in S}\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0\right\|_2$, where $S = \{i : \text{supp}(\tilde{\boldsymbol{\beta}}^{*i}) = \mathcal{A}_0\}$, and CS MSE, CS MAD and CS PBias are similarly defined.

A comparison between the scaled errors (3.2) and (3.3) of the nonzero estimators $\hat{\boldsymbol{\beta}}^{\hat{\mathcal{A}}}$ and $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})^{\hat{\mathcal{A}}}$ shows that post-selection OLS estimation helps eliminate a bias term

$\mathscr{B}_1$ arising from the nonconvex penalty, which mainly affects estimators of strong signals. It would therefore be interesting to investigate the performance of these estimators on the oracle active set $\mathcal{A}_0$ alone, leaving aside the errors attributed to the inactive group $\mathcal{A}_0^c$. To this end we report the total bias restricted to $\mathcal{A}_0$, namely OS Bias $= \left\| m^{-1} \sum_{i=1}^{m} (\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0)^{\mathcal{A}_0} \right\|_2$ and, conditional on $\{\hat{\mathcal{A}} = \mathcal{A}_0\}$, OS-CS Bias $= \left\| |S|^{-1} \sum_{i \in S} (\tilde{\boldsymbol{\beta}}^{*i} - \boldsymbol{\beta}_0)^{\mathcal{A}_0} \right\|_2$. Likewise we report also the other indicators restricted to $\mathcal{A}_0$: OS MSE (or OS MAD), OS-CS MSE (or OS-CS MAD) and OS PBias, noting that OS-CS PBias = OS PBias.

All figures shown in the tables of results are rounded to an appropriate number of decimal places such that any comparison between the figures would not be swamped by Monte Carlo error.

Examples and results

Example S.1. *(Gaussian setting with exact sparsity)* To address the question of how the methods perform in a traditional sparse setting, we consider a model where most coefficients are exactly zero

$$\mathbb{E}[Y|\boldsymbol{X}] = 0.5X^{(1)} + 0.6X^{(2)} + 0.7X^{(3)} + 5X^{(4)} + 10X^{(5)}.$$

There are $p = 500$ potential covariates, but only the first five have non-zero coefficients. For each simulation run, we set the sample size to $n = 139$. The covariates $X^{(1)}, \ldots, X^{(500)}$ and the error term ϵ are generated from a standard normal distri-

bution, $N(0,1)$. The entire simulation process is repeated 1000 times. The SCAD-penalized regression is fitted using the `ncvreg` R package with the tuning parameter $\lambda_{\mathsf{R}} = 0.3$. We obtain multiple sparse solutions by changing the initial guesses, and choose the sparse solution which yields the minimum empirical loss.

We investigate the performance of the various methods under three different correlation structures for the covariates: (1) the independent case ($\rho = 0$), where all covariates are independent; (2) the mildly correlated case ($\rho_{12} = 0.5$), where the correlation between $X^{(1)}$ and $X^{(2)}$ is set to 0.5, while all other pairs remain independent; (3) the highly correlated case ($\rho = 0.5$), where each pairwise correlation is set to 0.5.

The performance is assessed in terms of variable selection accuracy and estimation precision in Table S.1. In the independent and mildly correlated cases, the SCAD-based methods demonstrate superior performance in variable selection. They achieve a high rate of correct selection and select models of size close to the true value 5. In terms of estimation, post-SCAD OLS proves highly effective in reducing bias, yielding the lowest bias and MSE. Standard SCAD exhibits considerable bias, which is effectively corrected by the post-selection refitting step.

When strong multicollinearity exists among all covariates, the performance of all methods deteriorates, as expected. The correct selection rate of SCAD drops to 22.9%. In this challenging setting, post-selection OLS remains beneficial to SCAD,

significantly reducing both bias and MSE.

Table S.1: The average selected number of variables (N), percentages (%) of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), variance (Var) and mean squared error (MSE).

		variable selection					estimation		
ρ	Method	N	CS	FP	FN	FPN	Bias	Var	MSE
0	SCAD	**5.466**	**59.9**	36.5	2.2	1.4	0.8656	**0.0743**	0.8236
	post-SCAD						**0.1311**	0.0938	**0.1110**
	Oracle	5	–	–	–	–	0.0165	0.0453	0.0456
ρ_{12} =0.5	SCAD	**5.332**	**64.1**	29.3	4.5	2.1	0.6161	0.1373	0.5170
	post-SCAD						**0.1161**	0.1069	**0.1204**
	Oracle	5	–	–	–	–	0.0168	0.0508	0.0511
0.5	SCAD	**6.819**	**22.9**	54.3	10.5	12.3	1.47	0.231	2.39
	post-SCAD						**0.787**	0.254	**0.872**
	Oracle	5	–	–	–	–	0.00159	0.0701	0.0703

Figure S.1 displays heat maps of componentwise absolute errors of the SCAD and post-SCAD OLS estimates. They provide compelling visual evidence in support of the stronger theoretical guarantee our theory confers on post-SCAD OLS, in the sense that the second-stage OLS refitting step has the effect of removing the shrinkage bias of strong signals introduced by SCAD.

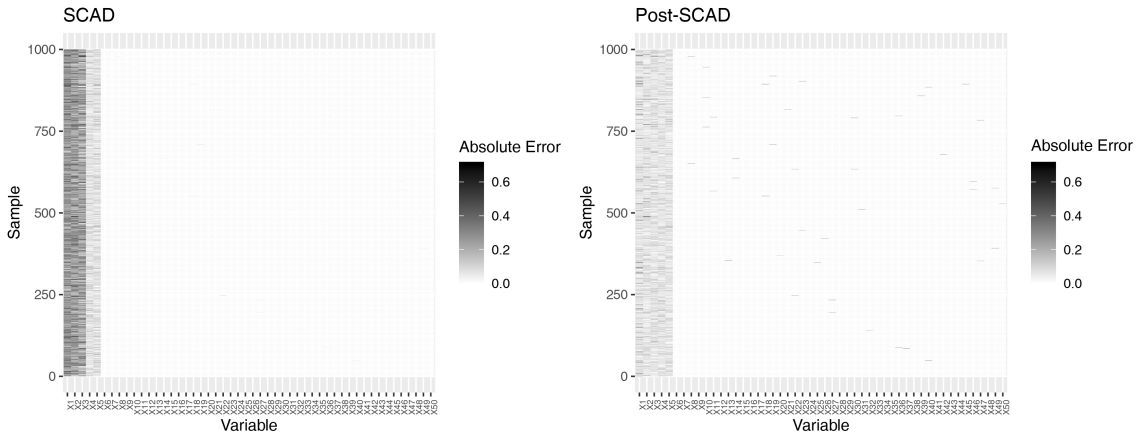

Figure S.1: Example S.1 — heat maps of componentwise absolute errors $|\tilde{\beta}_j - \beta_0^{(j)}|, j \in \{1, \ldots, 50\}$, for SCAD and post-SCAD estimators across 1000 replications under $\rho = 0.5$.

A comparison of the empirical cumulative distribution functions (ecdf) plotted in Figure S.2 consistently shows that the SCAD estimator has a distribution shifted to the left of the oracle. The shifts are pronounced for the first 3 coefficients and become significantly less conspicuous for $\beta_0^{(4)}$ and $\beta_0^{(5)}$. For each of the 5 coefficients, the post-SCAD OLS ecdf is nearly indistinguishable from the oracle, both of which are centred around zero and exhibit nearly identical shapes and dispersions. The results show that SCAD-based variable selection is sufficiently accurate and the subsequent step of OLS refitting effectively eliminates the shrinkage bias of SCAD estimates, leading to a feasible data-driven estimator whose finite-sample distribution almost perfectly matches that of the infeasible oracle.

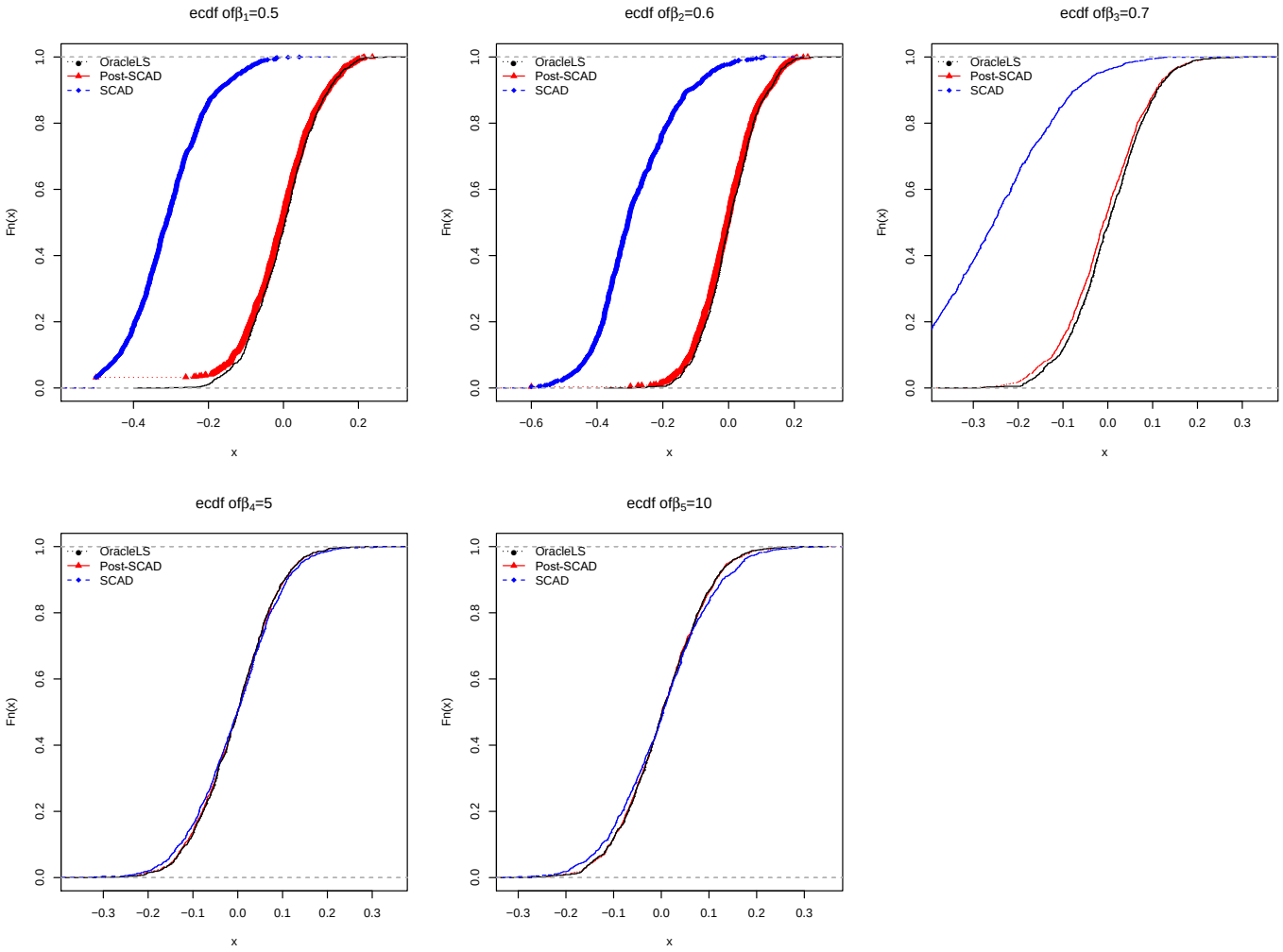

Figure S.2: Example S.1 — empirical cumulative distribution functions of $\tilde{\beta}_j - \beta_0^{(j)}$ $(j = 1, \ldots, 5)$ in the independent case.

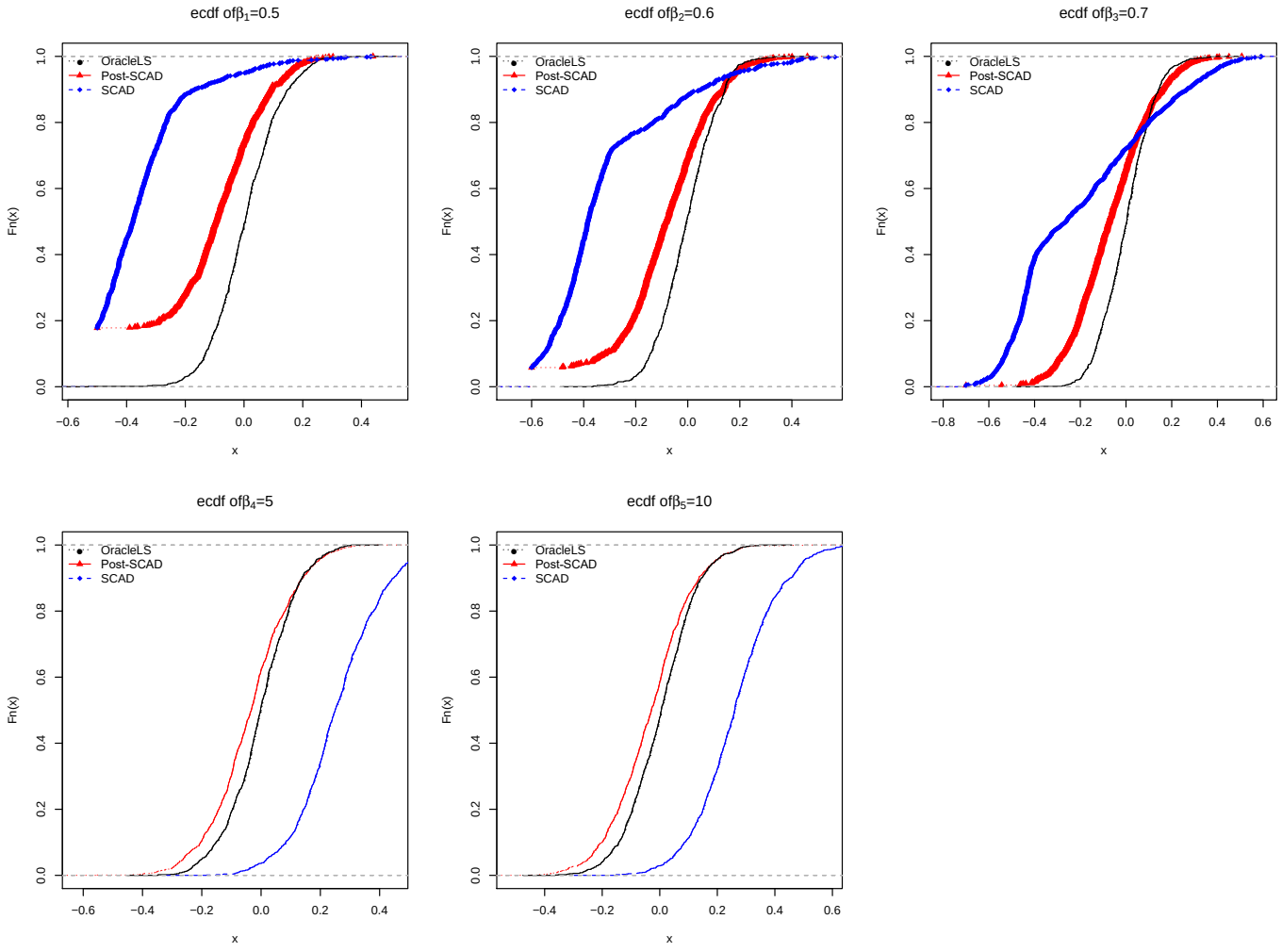

Figure S.3: Example S.1 — empirical cumulative distribution functions of $\tilde{\beta}_j - \beta_0^{(j)}$ $(j = 1, \ldots, 5)$ in the high-correlation case.

Figure S.3 presents the most challenging scenario, where all covariates have a uniform pairwise correlation of $\rho = 0.5$. Unlike the independent case, a clear gap now emerges between the post-SCAD and oracle ecdf's. While the two curves no

longer coincide perfectly, the post-SCAD OLS ecdf remains much closer to the oracle than does the SCAD ecdf. Both SCAD and post-SCAD OLS ecdf's are visibly flatter, more spread out and more prone to an atomic lower bound than those displayed in Figure S.2, indicating an increase in the variance of the estimates and a decrease in stability of model selection across the simulation runs.

Broadly speaking, as covariates become more strongly correlated, variances of the SCAD and post-SCAD OLS estimators increase and bias correction becomes less effective. In a sparse setting with inactive covariates associated with exact zero coefficients, the post-selection OLS method provides a viable strategy for improving estimation accuracy, even when reliability of variable selection in the first step has been plagued by multicollinearity.

Example S.2. *(Gaussian setting with weak sparsity)* Consider a regression model

$$\mathbb{E}[Y|\boldsymbol{X}] = 20 + \beta_0^{(2)} X^{(2)} + X^{(3)} + 1.5X^{(4)} + 3X^{(5)} + 10X^{(6)} + \sum_{j=7}^{501} \beta_0^{(j)} X^{(j)}.$$

We set the sample size $n = 140$, dimension $p = 501$ and the number of replications $m = 1000$. We generate $\beta_0^{(7)}, \ldots, \beta_0^{(501)}$ independently from $N(0, 0.01)$, reflecting a presence of many weak signals, and the random error ϵ from $N(0, 1)$. The covariate vector $[X^{(2)}, \ldots, X^{(501)}]^\top$ is generated from a 500-variate normal distribution $N(\mathbf{0}, \Sigma)$. The 500×500 covariance matrix Σ is constructed in line with the assumption (A1), with its (i, j)-th entry given by $\Sigma_{ij} = \boldsymbol{M}_i^\top \boldsymbol{M}_j \|\boldsymbol{M}_i\|_2^{-1} \|\boldsymbol{M}_j\|_2^{-1}$, for $i \neq j$, where $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_{500}$ denote 500 independent random vectors in $[0, 1]^{500}$, each consisting of

500 random components independently generated from the beta $(0.5, 10)$ distribution. In this example, the Σ_{ij} values range from about 0.224 to 0.517, about half of which lie between 0.345 and 0.388. We set $\Sigma_{ii} = \mathrm{Var}(X^{(i)}) \in \{1, 2\}$ and $\beta_0^{(2)} \in \{0.5, 0.73\}$. Note that changing Σ_{ii} from 1 to 2 reduces all correlations by 50% between each pair of covariates. The choices $\beta_0^{(2)} \in \{0.5, 0.73\}$ can be identified with a moderate signal close in magnitude to the sampling noise, with respect to the sample size 140 under a Gaussian setting. The optimisation program (2.3) yields an oracle active set $\mathcal{A}_0 = \{1, \ldots, 6\}$ for all four cases.

The performance of SCAD in variable selection under the four cases is summarised in Table S.2. With λ_{R} fixed at 0.3, the average numbers of selected variables (N) are all around 6. In general, a bigger value of $B_U = \beta_0^{(2)}$ or a weaker correlation between covariates gives rise to a higher percentage of correct selection.

Table S.2: Example S.2 — average number of selected variables (N), percentages (%) of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), over 1000 replications.

$\text{Var}(X^{(i)})$	$\beta_0^{(2)}$	Variable Selection N	CS	FP	FN	FPN
1	0.50	6.106	**64.2**	16.9	15.8	3.1
	0.73	6.208	**84.7**	15.0	0.3	0.0
2	0.50	6.322	**76.3**	23.6	0.1	0.0
	0.73	6.165	**85.8**	14.2	0.0	0.0

Table S.3: Example S.2 — different measures of total bias and total mean squared error.

$\textbf{Var}(X^{(i)})$	$\beta_0^{(2)}$	Method	Bias	MSE	PBias	CS Bias	CS MSE	CS PBias	OS Bias	OS MSE	OS PBias	OS-CS Bias	OS-CS MSE
1	0.5	SCAD	0.471	0.321	0.403	0.45	0.295	0.379	0.412	0.269	0.403	0.388	0.243
		Post-SCAD	**0.248**	**0.196**	**0.086**	**0.231**	**0.109**	**0.032**	**0.099**	**0.116**	**0.086**	**0.035**	**0.057**
		Oracle	0.23	0.113	0	0.23	0.113	0	0.03	0.061	0	0.03	0.061
	0.73	SCAD	0.428	0.315	0.354	0.421	0.301	0.345	0.363	0.263	0.354	0.354	0.249
		Post-SCAD	**0.231**	**0.142**	**0.023**	**0.23**	**0.113**	**0.011**	**0.043**	**0.069**	**0.022**	**0.03**	**0.061**
		Oracle	0.23	0.113	0	0.23	0.113	0	0.03	0.061	0	0.03	0.061
2	0.5	SCAD	0.313	0.137	0.21	0.31	0.135	0.205	0.214	0.085	0.21	0.209	0.083
		Post-SCAD	**0.228**	**0.1**	**0.015**	**0.228**	**0.083**	**0.005**	**0.023**	**0.034**	**0.014**	**0.012**	**0.031**
		Oracle	0.229	0.083	0	0.229	0.083	0	0.015	0.031	0	0.015	0.031
	0.73	SCAD	0.236	0.095	0.058	0.237	0.095	0.058	0.063	0.043	0.058	0.063	0.043
		Post-SCAD	**0.228**	**0.094**	**0.007**	**0.228**	**0.083**	**0.005**	**0.015**	**0.032**	**0.006**	**0.014**	**0.031**
		Oracle	0.229	0.083	0	0.229	0.083	0	0.015	0.031	0	0.015	0.031

Table S.3 reports a variety of measures of estimation errors of SCAD, post-SCAD OLS and Oracle. In all four settings of $\left(\text{Var}(X^{(i)}), \beta_0^{(2)}\right)$, post-SCAD OLS consistently outperforms SCAD in yielding smaller total bias and smaller total mean squared error calculated by any measure. A comparison between Bias and CS Bias shows that the performance of either estimator does not change much if the bias is evaluated conditional on successful selection of $\mathcal{A}_0$. Remarkably, SCAD still experiences a substantial bias even when the correct oracle set $\mathcal{A}_0$ has been selected. The improvement made by post-SCAD OLS over SCAD is most noticeable in OS Bias and OS-CS Bias, in comparison with that made in Bias and CS Bias, implying that a significant reduction in bias is achieved by post-SCAD OLS for the estimation of the active coefficients $\boldsymbol{\beta}_0^{\mathcal{A}_0}$. The results can be attributed to the fact that a significant bias term $\mathscr{B}_1$ in the scaled error (3.2) of the active SCAD estimator is removed by post-SCAD OLS (3.3), thus pushing the latter much closer to the oracle estimator. It can be derived from the bias and MSE figures that the SCAD estimator has a total variance smaller than that of the post-SCAD OLS estimator, due possibly to the shrinking effect of the penalty. After adjusting $\boldsymbol{\beta}_0$ to the oracle target $\boldsymbol{\theta}_0$, the measures PBias, CS PBias and OS PBias show a trend similar to, but are in magnitude smaller than, Bias, CS Bias and OS Bias, respectively, which agrees with the theory given in Section S.1.3. The distinction between Bias and OS Bias is more appreciable than that between PBias and OS PBias.

Under the same $\text{Var}(X^{(i)}) = 1$ or 2, if we increase $B_U = \beta_0^{(2)}$ from 0.5 to 0.73, the biases of SCAD decrease, which conforms with an increase in convergence rate induced by a progression to the generalised oracle phase (Corollary 1) under a sufficiently small ψ. The biases of post-SCAD OLS exhibit a much steadier change as B_U increases and B_0 remains fixed, corroborating its robust convergence rate stipulated in Theorem 3 and Corollary 2. The selection performance of SCAD is in line with the phase changes from patterns 2 to 5 shown in Figure 1.

A comparison between the cases $\text{Var}(X^{(i)}) = 1$ and 2 suggests that the performances of SCAD in both selection and estimation are sensitive to the correlations between covariates, resulting in a bigger bias under the case $\text{Var}(X^{(i)}) = 1$ where covariate correlations are stronger. By contrast, the results on Bias, CS Bias, OS Bias and OS-CS Bias of post-SCAD OLS appear more robust and stable against changes to covariate correlations. Such difference in the trend of bias between the two approaches can be accounted for by the shrinking effect of the SCAD penalty. Consider for illustration a simple scenario where $\mathcal{A}_0 = \{1, 2\} = \hat{\mathcal{A}}$ is correctly selected, Σ_{12} is fixed and $B_0 = 0$. Then the absolute bias of SCAD can be shown to decrease with $\text{Var}(X^{(i)})$, or equivalently, increase with the correlations between covariates, while the post-SCAD OLS estimator remains unbiased.

The three heat maps in Figure S.4 show in more detail the selection results of SCAD (top panel) and componentwise estimation errors of SCAD and post-SCAD

OLS (bottom panel) with respect to the first 50 signals, for each of the 1000 replications under the case $(\mathrm{Var}(X^{(i)}), \beta_0^{(2)}) = (1, 0.5)$, where SCAD shows the lowest rate of correct selection ($\hat{\mathcal{A}} = \mathcal{A}_0$). We see that the variables in the oracle active set $\mathcal{A}_0$ except $X^{(2)}$ had been selected in all 1000 replications, while $X^{(2)}$, associated with the small $\beta_0^{(2)} = 0.5$, had been selected only 811 times. Variables outside $\mathcal{A}_0$ were rarely selected. The heat maps of absolute errors show that post-SCAD OLS outperforms SCAD slightly in estimating $\{\beta_0^{(j)} : j = 1, 3, 4, 5, 6\}$ and is notably more accurate in estimating $\beta_0^{(2)}$. On the other hand, there exist a few cases where SCAD is found to be more accurate than post-SCAD OLS in estimating $\beta_0^{(7)}, \ldots, \beta_0^{(50)}$, when their corresponding, inactive, variables have been wrongly selected by SCAD. Here the SCAD penalty has played a role in shrinking the SCAD estimates towards the inactive $\beta_0^{(j)}$'s, hence a smaller estimation error compared to that given by the non-penalised OLS estimates.

Figure S.5 plots the marginal ecdf's of $\tilde{\beta}_j - \beta_0^{(j)}$, $j = 1, \ldots, 6$, for SCAD, post-SCAD OLS and Oracle under the case $(\mathrm{Var}(X^{(i)}), \beta_0^{(2)}) = (1, 0.5)$. Except for the intercept $\beta_0^{(1)}$, which has a large value 20, distributions of SCAD estimates of $\beta_0^{(2)}, \ldots, \beta_0^{(6)}$ are markedly different from the oracle with either a location shift (cases $j = 2, 4, 5, 6$) or a scale shift (case $j = 3$). Such discrepancy is noted even for the SCAD estimate of a large coefficient like $\beta_0^{(6)} = 10$, attesting to the adverse impact on the oracle property of SCAD exerted by the presence of a moderate signal $X^{(2)}$

which is correlated with the strong ones. The problem is resolved successfully by post-SCAD OLS, resulting in ecdf's almost indistinguishable from the oracle. For the case $j = 2$, a high false negative rate 15.8% gives rise to a conspicuous atom at $-\beta_0^{(2)} = -0.5$ in the ecdf's of both SCAD and post-SCAD OLS.

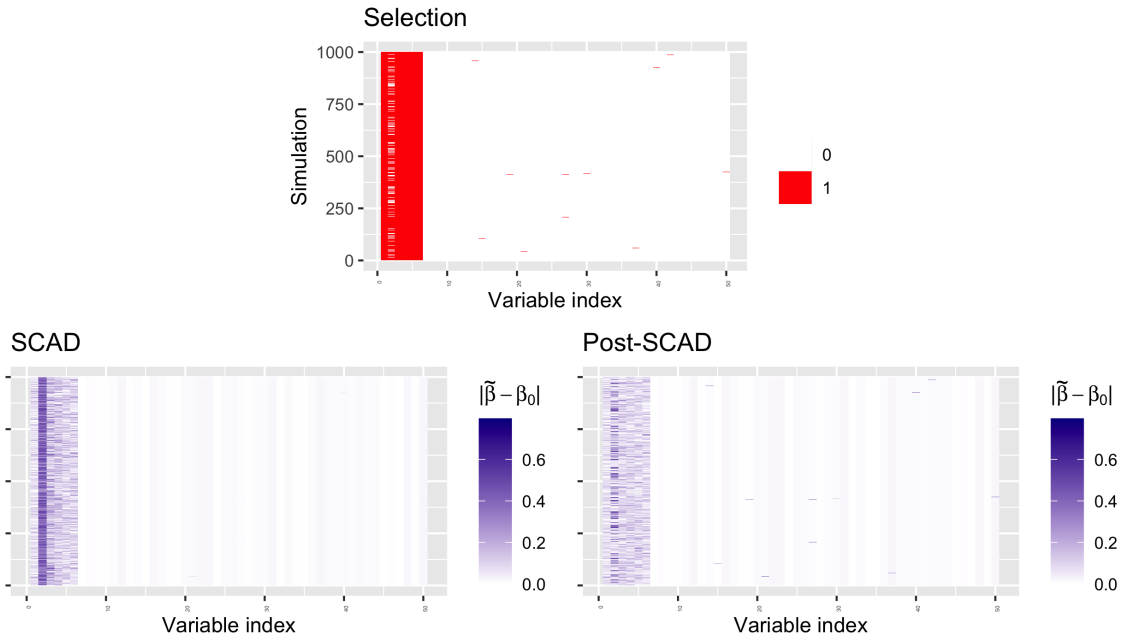

Figure S.4: Example S.2 — heat maps of SCAD selection results and componentwise absolute errors of SCAD and post-SCAD OLS estimates, across 1000 replications under the setting $(\mathrm{Var}(X^{(i)}), \beta_0^{(2)}) = (1, 0.5)$.

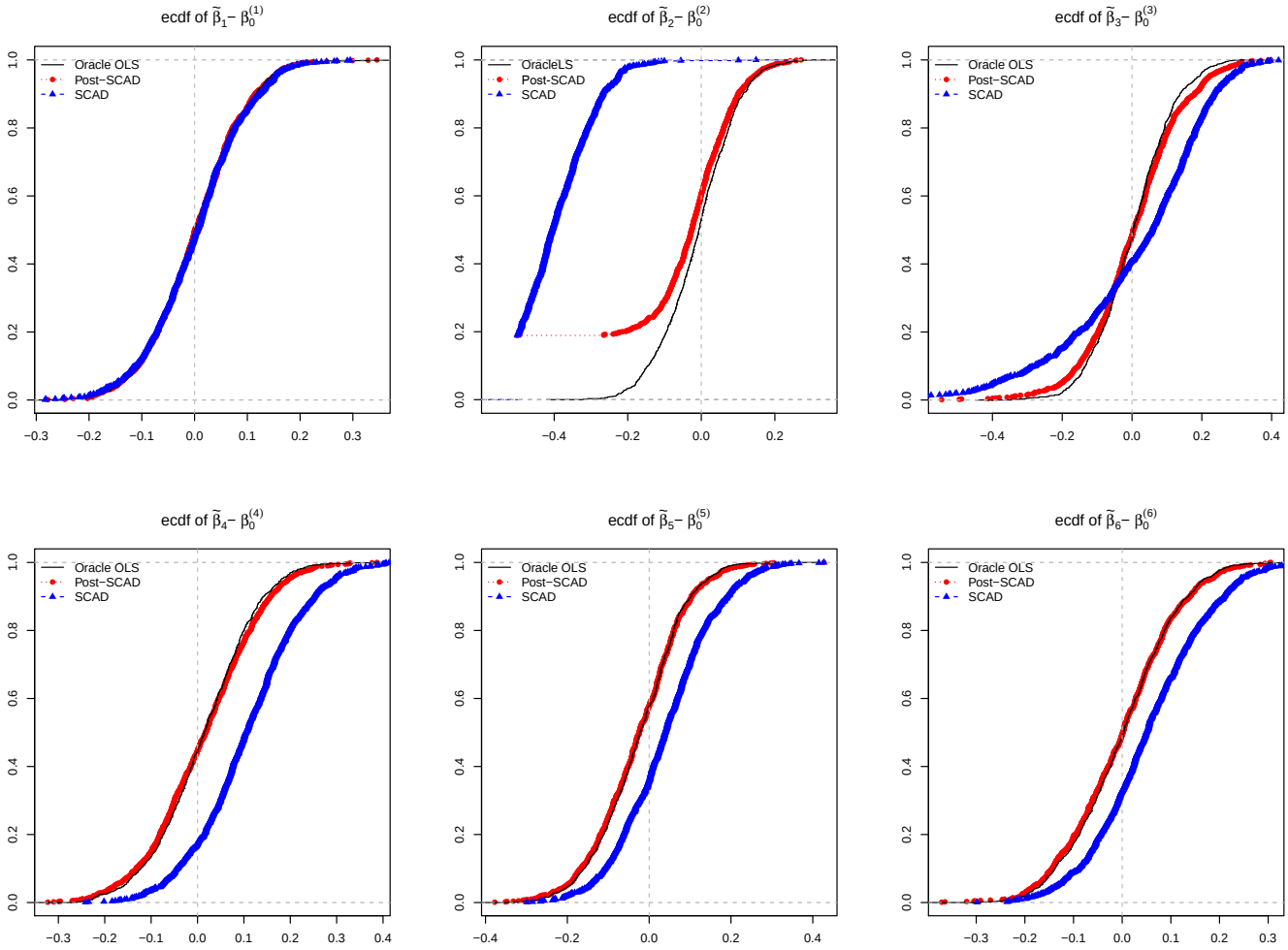

Figure S.5: Example S.2 — empirical cumulative distribution functions of $\tilde{\beta}_j - \beta_0^{(j)}$ $(j = 1, \ldots, 6)$ under the setting $(\mathrm{Var}(X^{(i)}), \beta_0^{(2)}) = (1, 0.5)$.

Example S.3. *(Heavy-tailed setting)* Consider a regression model

$$\mathbb{E}[Y|\boldsymbol{X}] = 20 + \beta_0^{(2)} X^{(2)} + 4X^{(3)} + 5X^{(4)} + 6X^{(5)} + 12X^{(6)} + \sum_{j=7}^{51} \beta_0^{(j)} X^{(j)}.$$

We set $n = 200$, $p = 51$ and $m = 10^5$. The variables $\{X^{(j)} : j = 2, \ldots, 51\}$ and the coefficients $\{\beta_0^{(j)} : j = 7, \ldots, 51\}$ are generated in the same way as in Example S.2. The random error ϵ has a density function with Pareto tails, given by

$$f_\epsilon(t) = (1/6)\,\mathbf{1}\{|t| < 1\} + (1/2)|t|^{-5/2}\mathbf{1}\{|t| \geq 1\},$$

so that ϵ has zero mean and an infinite variance, and falls within the intervals $(-1, 1)$, $(-\infty, -1]$ and $[1, \infty)$ with equal probabilities. It satisfies the heavy-tailed condition $(\mathcal{T}_3)$ with tail index $\gamma = 3/2$. Lemma 1 holds in this case under a more stringent condition on the dimension, $p \prec n^{1/2}$, with the penalty weight λ satisfying $(np)^{2/3} \prec \lambda \prec n$. Such consideration leads us to set p to be smaller than n in this example, as opposed to the large p attempted in Example S.2. The SCAD penalty weight λ_{R} is fixed at 2.3.

We set $\mathrm{Var}(X^{(i)}) \in \{2, 3\}$ and $\beta_0^{(2)} \in \{3, 4\}$. These choices differ slightly from those considered in Example S.2 and coordinate better with the heavy-tailed setting to reveal the effects of moderate signals. Again, the oracle active set is found to be $\mathcal{A}_0 = \{1, \ldots, 6\}$ in all four cases. The error indicator MSE is changed to MAD for a more robust measure under heavy tails.

Tables S.4 and S.5 report the results on selection and estimation, respectively. They reveal patterns quite similar to those obtained in Example S.2, so that our comments on the latter example carry over here in general. Note that the total magnitude of weak signals, $\|\boldsymbol{\beta}_0^{\mathcal{A}_0^c}\|_2 \approx 0.077$, is much smaller than the magnitude

Table S.4: Example S.3 — average number of selected variables (N), percentages (%) of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), over 1000 replications.

$\mathrm{Var}(X^{(i)})$	$\beta_0^{(2)}$	Variable Selection N	CS	FP	FN	FPN
2	3	5.467	**93.1**	5.1	0.2	1.7
	4	5.465	**93.6**	5.3	0.0	1.2
3	3	5.483	**94.0**	5.1	0.0	0.9
	4	5.479	**94.3**	5.1	0.0	0.6

0.228 found in Example S.2, as a result of a reduction in the dimension p. Thus, different bias measures yield similar results within each method. Despite a larger degree of disturbance caused by the heavy-tailed noise than its Gaussian counterpart in Example S.2, post-SCAD OLS makes a more remarkable improvement over SCAD in both bias and MAD. Compared to the regular order $n^{-1/2}$ in the Gaussian case, the tail condition ($\mathcal{T}_3$) entails a noise level of a higher order $\sqrt{T}/n = n^{-1/3}$. This necessitates a heavier SCAD penalty weight $\succ n^{-1/3}$ for consistent selection of strong signals, which in turn amplifies the penalty-driven bias term $\mathscr{B}_1$. Bias reduction made by post-SCAD OLS, effected mainly through elimination of $\mathscr{B}_1$, is therefore much more remarkable under a heavy-tailed setting than that achieved in Example S.2, as can be seen by comparing Tables S.3 and S.5.

Table S.5: Example S.3 — different measures of total bias and total mean absolute deviation.

$\text{Var}(X^{(i)})$	$\beta_0^{(2)}$	Method	Bias	MAD	PBias	CS Bias	CS MAD	CS PBias	OS Bias	OS MAD	OS PBias	OS-CS Bias	OS-CS MAD
2	3	SCAD	2.153	7.7	2.153	2.167	5.08	2.167	2.151	5.21	2.153	2.165	4.66
		Post-SCAD	**0.098**	**5.0**	**0.061**	**0.077**	**1.62**	**0.003**	**0.047**	**1.86**	**0.048**	**0.007**	**1.20**
		Oracle	0.078	2.2	0	0.078	2.16	0	0.014	1.73	0	0.014	1.73
	4	SCAD	1.816	7.2	1.817	1.819	4.56	1.820	1.814	4.70	1.816	1.818	4.13
		Post-SCAD	**0.093**	**5.0**	**0.055**	**0.077**	**1.63**	**0.003**	**0.038**	**1.85**	**0.040**	**0.007**	**1.21**
		Oracle	0.078	2.2	0	0.078	2.15	0	0.011	1.73	0	0.011	1.73
3	3	SCAD	1.353	5.3	1.352	1.358	3.24	1.356	1.351	3.28	1.352	1.356	2.82
		Post-SCAD	**0.082**	**4.1**	**0.034**	**0.077**	**1.45**	**0.002**	**0.021**	**1.53**	**0.022**	**0.004**	**1.03**
		Oracle	0.077	1.9	0	0.077	1.88	0	0.008	1.46	0	0.008	1.46
	4	SCAD	0.826	4.6	0.824	0.818	2.51	0.816	0.823	2.58	0.824	0.814	2.09
		Post-SCAD	**0.080**	**4.1**	**0.030**	**0.077**	**1.45**	**0.002**	**0.016**	**1.53**	**0.017**	**0.004**	**1.03**
		Oracle	0.077	1.9	0	0.077	1.88	0	0.008	1.46	0	0.008	1.46

Example S.4. *(Gaussian setting, highly-correlated covariates)* Consider a regression model

$$\mathbb{E}[Y|\boldsymbol{X}] = 20 + \beta_0^{(2)} X^{(2)} + 1.8X^{(3)} + 2X^{(4)} + 3X^{(5)} + 10X^{(6)} + \sum_{j=7}^{501} \beta_0^{(j)} X^{(j)}.$$

The parameters $n, p, m, \lambda_{\mathsf{R}}$ and the distributions of $\epsilon, \beta_0^{(7)}, \ldots, \beta_0^{(501)}$ are the same as those set in Example S.2. The covariates $[X^{(2)}, \ldots, X^{(501)}]^\top$ are also generated in the same way as in Example S.2, except that $\text{Var}(X^{(i)})$ is fixed at 1 and the parameters of

the beta distribution are changed from $(0.5, 10)$ to $(2, 5)$. The latter change gives rise to a covariance matrix Σ with Σ_{ij} ranging from 0.704 to 0.815 for $i \neq j \in \{2, \ldots 501\}$, implying strong correlations between the covariates $X^{(2)}, \ldots, X^{(501)}$. We set $\beta_0^{(2)} \in \{1.3, 1.7\}$, under which the oracle active set is found by (2.3) to be $\mathcal{A}_0 = \{1, \ldots, 6\}$, with B_U identified with $\beta_0^{(2)}$.

In general, SCAD shows a good performance in selection, yielding a rate of correct selection 90.4% in the case $\beta_0^{(2)} = 1.3$ and 99.8% in the case $\beta_0^{(2)} = 1.7$. The results on estimation accuracy are qualitatively very similar to those obtained in Example S.2 and are therefore omitted here.

Figures S.6 and S.7 show the ecdf's of $\tilde{\beta}_j - \beta_0^{(j)}$, $j \in \mathcal{A}_0$, under the two cases of $\beta_0^{(2)}$, respectively. Apparently, for the case $\beta_0^{(2)} = 1.7$ (Figure S.7), both SCAD and post-SCAD OLS perform almost as well as the oracle in selection and estimation, echoing an oracle phase exemplified by pattern 5 on the local asymptotic spectrum shown in Figure 1. However, when we switch $B_U = \beta_0^{(2)}$ to a smaller value 1.3 (Figure S.6), all the SCAD estimates except $\hat{\beta}_1$ yield ecdf's markedly different from the oracle, suggesting a move into non-oracle phases exemplified by patterns 1 to 3 in Figure 1, where the SCAD estimates suffer from a slower rate of convergence. The problem is resolved to some extent by post-SCAD OLS, which reduces the bias of SCAD and helps achieve a distribution closer to the oracle. Similar comments also hold for the previous examples, albeit to a lesser extent.

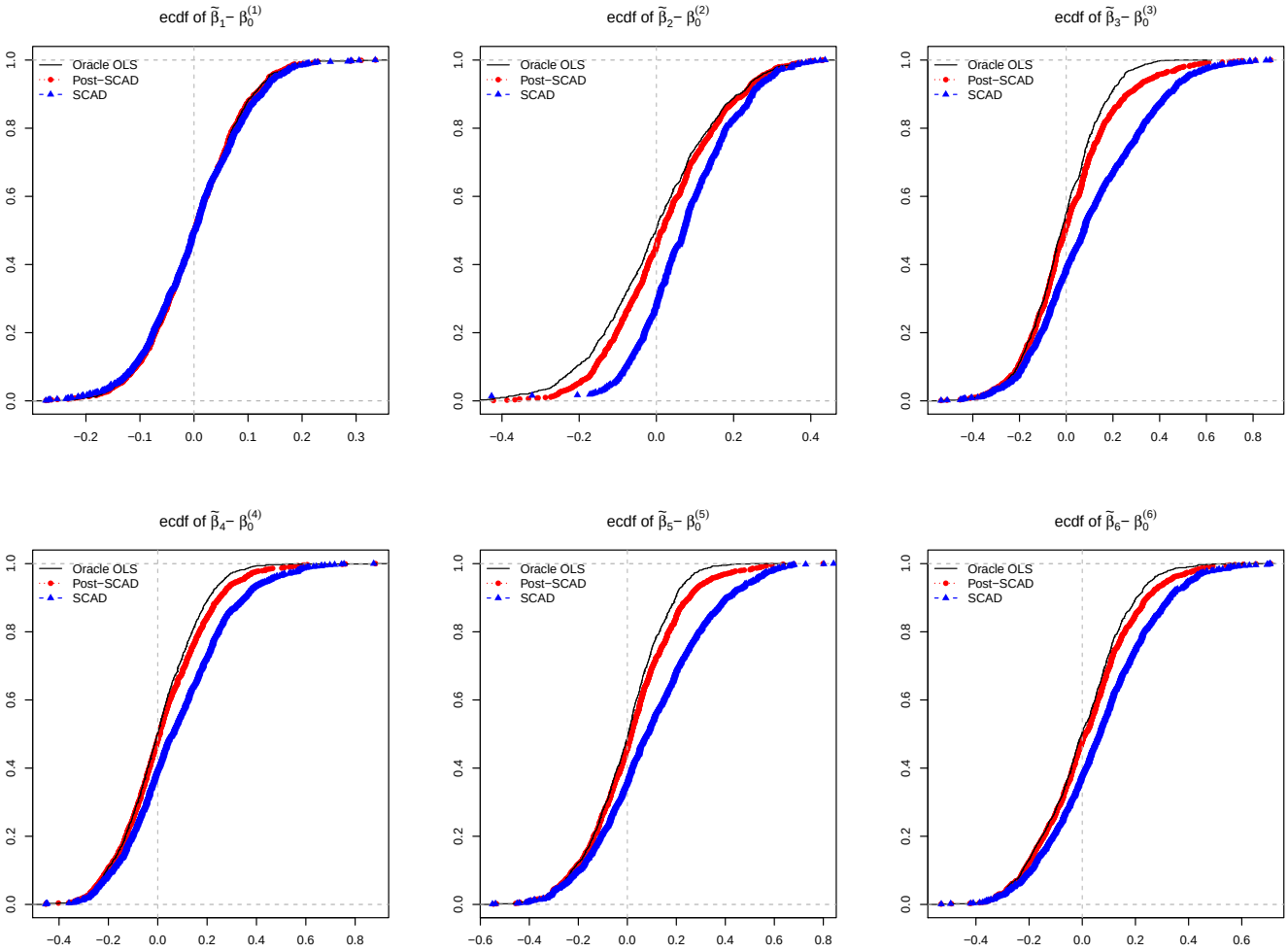

Figure S.6: Example S.4 — empirical cumulative distribution functions of $\tilde{\beta}_j - \beta_0^{(j)}$ $(j = 1, \ldots, 6)$ under the setting $\beta_0^{(2)} = 1.3$.

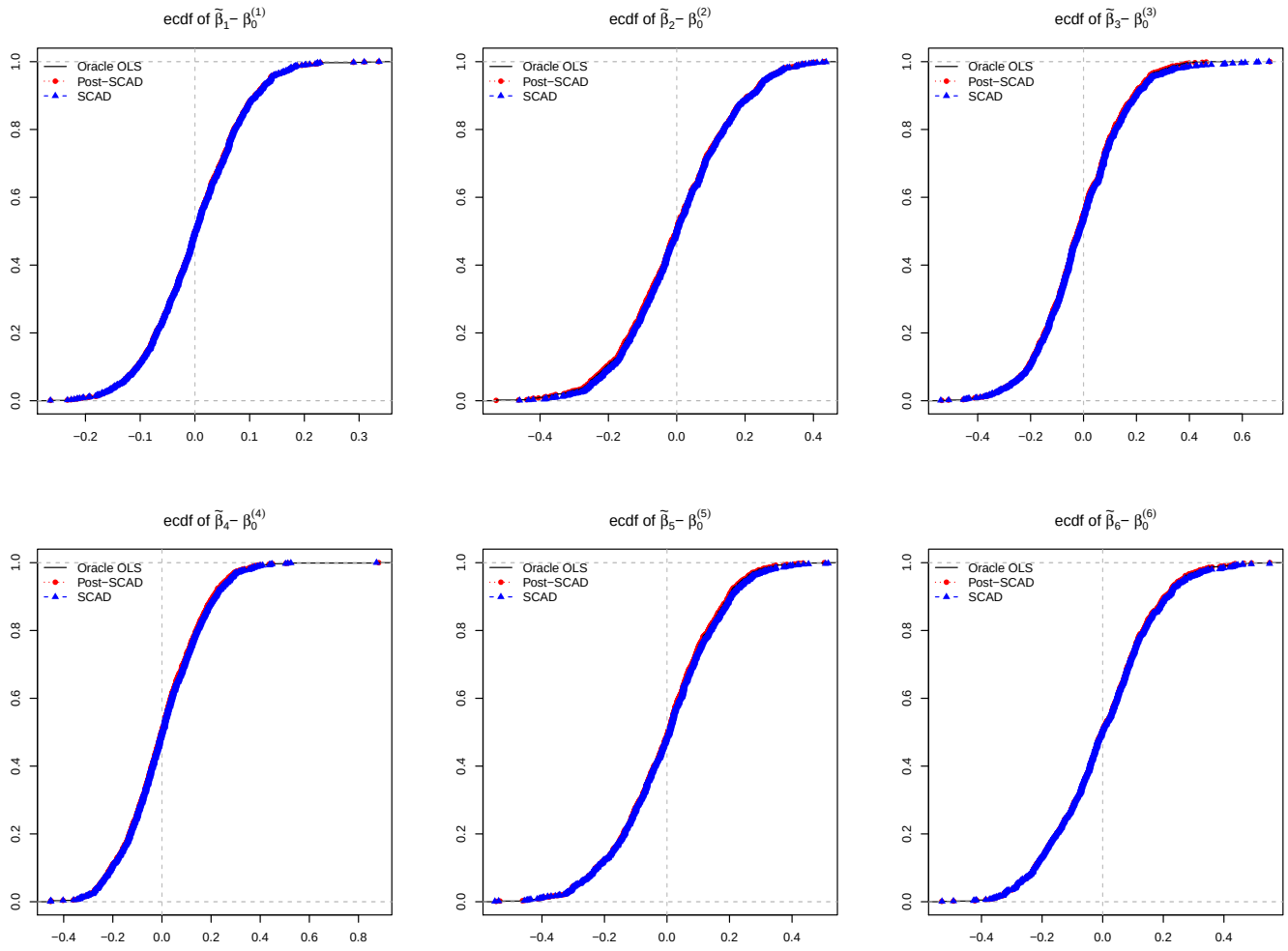

Figure S.7: Example S.4 — empirical cumulative distribution functions of $\tilde{\beta}_j - \beta_0^{(j)}$ $(j = 1, \ldots, 6)$ under the setting $\beta_0^{(2)} = 1.7$.

S.1.5 Concluding remarks

In summary, we show under all three types of tail conditions that phase changes in the asymptotics of $\hat{\boldsymbol{\beta}}$ are determined critically by (B_0, B_U), and provide a necessary and sufficient condition, namely $\psi \equiv B_0 \vee \{(\lambda/n)(1 - B_U/(\alpha\kappa))\}_+ \prec \lambda/n$, for the existence of a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ which selects $\mathcal{A}_0$ consistently and has a fast componentwise convergence rate $(n/\sqrt{T}) \wedge \psi^{-1}$. It follows that when B_U is not large enough or B_0 is not small enough, the generalised oracle property fails to hold for any consistent sparse local minimum $\hat{\boldsymbol{\beta}}$. By removing the bias term $\mathscr{B}_1$, the post-selection OLS estimators $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ acquire convergence properties more desirable than the local minima $\hat{\boldsymbol{\beta}}$ and, in the case of multiple solutions to the nonconvex optimisation program (2.2), ratewise more robust against the choice of strong signal sets $\hat{\mathcal{A}}$. If $B_U \succ \lambda/n$, our Corollary S.2 shows that the fastest rate $n/\sqrt{T}$ is achieved by all choices of $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$, while the corresponding local minima $\hat{\boldsymbol{\beta}}$ except $\hat{\boldsymbol{\beta}}_{go}$ converge only at the slowest rate n/λ.

We have argued that from a predictive perspective, adjusting $\boldsymbol{\beta}_0$ for the omission of weak signals makes for a practically more relevant target $\boldsymbol{\theta}_0$ under a weakly sparse model. With this change of target, we may weaken the condition on B_0 from $B_0 \prec \sqrt{T}/n$ to $B_0 \prec \lambda/n$ for $\hat{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ to satisfy the generalised oracle property.

We have conducted elaborate simulation studies to compare SCAD with post-SCAD OLS by a variety of numerical and graphical measures. The numerical findings

corroborate our theory in general, suggesting that post-SCAD OLS successfully reduces the bias of SCAD and displays a more robust performance. The improvement made by post-SCAD OLS is especially significant under a heavy-tailed setting, which calls for a heavier SCAD penalty weight for consistent selection.

S2 Appendix 2

S.2.1 Proof of Lemma 1

We first show that $n^{-1}T^{1/2}\|\boldsymbol{W}\|_\infty$ and $\max_{1\le j,j'\le p}|\hat{C}_{jj'}-\mathcal{C}_{jj'}|$ are uniformly bounded by a negligibly small sequence in probability, under different tail conditions. Then we establish two sub-lemmas. Lemma S.1 shows that $\mathbb{P}(\mathscr{K}_n = \hat{\mathscr{K}}_n) \to 1$. Under $\lambda \succ \sqrt{T}$, Lemma S.2 establishes asymptotic "noiselessness" of $\hat{\mathcal{A}}_n$, in the sense that $\hat{\mathcal{A}}_n$ is not affected, to first order, by the noise term $\boldsymbol{W}^{\hat{\mathcal{A}}_n}$. The two lemmas are employed to build a one-to-one correspondence between $\hat{\mathcal{A}}_n \in \hat{\mathscr{K}}_n$ and $\mathcal{A}_n \in \mathscr{K}_n$ when n is large enough. Finally, the above results are combined to prove Lemma 1. For brevity we write $\hat{\mathcal{A}} = \hat{\mathcal{A}}_n$ and $\hat{\mathscr{K}} = \hat{\mathscr{K}}_n$ when there is no confusion, and denote by c a positive constant which may assume different values in different expressions.

Let $\{a_n\}$ be a positive sequence depending on n. Define events $E_1 = \{n^{-1}T^{1/2}\|\boldsymbol{W}\|_\infty \le a_n\}$ and $E_2 = \big\{\max_{1\le j,j'\le p}|\hat{C}_{jj'}-\mathcal{C}_{jj'}| \le a_n\big\}$. We shall show that there exists some $\{a_n\}$ such that $\mathbb{P}(E_1 \cap E_2) \to 1$ under tail conditions ($\mathcal{T}_1$)–($\mathcal{T}_3$).

Suppose that ($\mathcal{T}_1$) holds. Denote the survival function of a random variable Z by $\bar{F}_Z(t) = \mathbb{P}(Z > t)$, so that $\bar{F}_{|\epsilon|}(t) = \mathbb{P}(|\epsilon| > t) \le ke^{-ct^2}$ and $\bar{F}_{|X|}(t) = \mathbb{P}(|X^{(j)}| > t)$. Note, for $j = 1, \ldots, p$ and sufficiently large $t > 0$, that

$$\begin{aligned}\mathbb{P}(|\epsilon X^{(j)}| > t) &= \mathbb{E}\bar{F}_{|\epsilon|}(t/|X^{(j)}|)\mathbf{1}_{\{0\le|X^{(j)}|\le\sqrt{t}\}} + \mathbb{E}\bar{F}_{|\epsilon|}(t/|X^{(j)}|)\mathbf{1}_{\{|X^{(j)}|>\sqrt{t}\}} \\ &\le \bar{F}_{|\epsilon|}(\sqrt{t}) + \mathbb{P}\big(|X^{(j)} - \mathbb{E}X^{(j)}| > \sqrt{t} - |\mathbb{E}X^{(j)}|\big) \le ke^{-ct},\end{aligned}$$

for some constants $c, k > 0$. Letting $T = n$ and $\{a_n\}$ satisfy $\sqrt{n^{-1}\log p} \prec a_n \prec \lambda/n$, it follows from Lemmas 2.2.11 (Bernstein Inequality) and 2.2.10 of Van Der Vaart et al. (1996) that, as $n \to \infty$,

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &\geq 1 - \mathbb{P}\big(n^{-1/2}\|\boldsymbol{W}\|_\infty > a_n\big) - \mathbb{P}\big(\max_{1\leq j,j'\leq p} |\hat{C}_{jj'} - \mathcal{C}_{jj'}| > a_n\big) \\ &\geq 1 - \Big\{\exp\Big(\frac{na_n}{k_1\log(1+p) + k_2\sqrt{n\log(1+p)}}\Big) - 1\Big\}^{-1} \\ &\quad - \Big\{\exp\Big(\frac{na_n}{k_3\log(1+p^2) + k_4\sqrt{n\log(1+p^2)}}\Big) - 1\Big\}^{-1} \to 1, \quad \text{(S2.1)}\end{aligned}$$

for some positive constants $k_1, \ldots, k_4$.

Under tail conditions $(\mathcal{T}_2)$ or $(\mathcal{T}_3)$, it follows by Theorem 4 of Sarantsev (2011) that $\mathbb{P}(|\epsilon X^{(j)}| > x) \asymp x^{-\gamma}$ as $x \to \infty$. Denote by $Z_1, Z_2, \ldots$ a generic sequence of i.i.d. random variables with mean zero and satisfying $\mathbb{P}(|Z_i| > x) \leq cx^{-\gamma}$ for all $x > 0$. Consider first the case $(\mathcal{T}_3)$. Let $T = n^{2/\gamma}$ and $\{a_n\}$ satisfy $n^{-1+1/\gamma}p^{1/\gamma} \prec a_n \prec \lambda/n$. Then, by Theorem 3.1.1 of Borovkov (2008), we have, for $x > 0$ and as $v \to 0$,

$$(nc)^{-1}x^\gamma\mathbb{P}\Big(\sum_{i=1}^n Z_i > x\Big) \leq \sup\Big\{(nc)^{-1}x^\gamma\mathbb{P}\Big(\max_{k\leq n}\sum_{i=1}^k Z_i \geq x\Big) : ncx^{-\gamma} \leq v\Big\} \leq 1+o(1).$$

It follows, for sufficiently large n and $x \succ n^{1/\gamma}$, that $\mathbb{P}(|\sum_{i=1}^n Z_i| > x) \leq 4cnx^{-\gamma}$. Substituting $Z_i = \epsilon_i X_i^{(j)}$ and $Z_i = X_i^{(j)}X_i^{(j')} - \mathcal{C}_{jj'}$, respectively, we have, under $(\mathcal{T}_3)$, that

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &\geq 1 - \sum_{j=1}^p \mathbb{P}(n^{-1+1/\gamma}|W_j| > a_n) - \sum_{j,j'=1}^p \mathbb{P}\big(|\hat{C}_{jj'} - \mathcal{C}_{jj'}| > a_n\big) \\ &\geq 1 - 8cn^{1-\gamma}pa_n^{-\gamma} \to 1. \quad \text{(S2.2)}\end{aligned}$$

For the case ($\mathcal{T}_2$), set $T = n$ and $\{a_n\}$ to satisfy $(p/n)^{1/2} \prec a_n \prec \lambda/n$. It follows by Lemma 4 of Sarantsev (2011) that $\mathrm{var}(Z_i) < \infty$, and by Markov inequality that $\mathbb{P}(|\sum_{i=1}^n Z_i| > x) \le nx^{-2}\mathrm{var}(Z_i) \le cnx^{-2}$. Thus, by the same arguments as given above,

$$\mathbb{P}(E_1 \cap E_2) \ge 1 - 2cpn^{-1}a_n^{-2} \to 1. \tag{S2.3}$$

Lemma S.1. *Assume the conditions of Lemma 1. Then we have, under the event $E_1 \cap E_2$, that $\mathscr{K}_n = \hat{\mathscr{K}}_n$ for sufficiently large n.*

Proof of Lemma S.1 . We first prove $\mathscr{K}_n \subset \hat{\mathscr{K}}_n$. Fix a population penalised parameter $\boldsymbol{\beta}^*$ with support $\mathcal{A}_n \in \mathscr{K}_n$. Define, for $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top \in \mathbb{R}^p$,

$$\begin{aligned}\tau(\boldsymbol{\beta}) = \frac{\lambda}{2n}\Bigg[&\mathrm{sgn}(\beta_j)\Big(\mathbf{1}\{0 < |\beta_j| \le \kappa\} + \frac{\alpha}{\alpha-1}\mathbf{1}\{\kappa < |\beta_j| < \alpha\kappa\}\Big) \\ &- \frac{\beta_j}{(\alpha-1)\kappa}\mathbf{1}\{\kappa < |\beta_j| < \alpha\kappa\} : j = 1, \ldots, p\Bigg], \\ D(\boldsymbol{\beta}) = \mathrm{diag}\big(&\mathbf{1}\{\kappa < |\beta_j| < \alpha\kappa\} : j \in \mathrm{supp}(\boldsymbol{\beta})\big).\end{aligned}$$

Writing $\mathcal{B} = \mathrm{supp}(\boldsymbol{\beta})$, let $\hat{C}'(\boldsymbol{\beta}) = \hat{C}_{\mathcal{B}\mathcal{B}} - \{2n(\alpha-1)\kappa\}^{-1}\lambda D(\boldsymbol{\beta})$ and $\mathcal{C}'(\boldsymbol{\beta}) = \mathbb{E}\hat{C}'(\boldsymbol{\beta})$. The latter is invertible for sufficiently large n under (A1). Let $\{\mu_i\}$, $\{\nu_i\}$ and $\{\rho_i\}$ be decreasing sequences of eigenvalues of the matrices $\hat{C}'(\boldsymbol{\beta})$, $\mathcal{C}'(\boldsymbol{\beta})$ and $\hat{C}'(\boldsymbol{\beta}) - \mathcal{C}'(\boldsymbol{\beta}) = \hat{C}_{\mathcal{B}\mathcal{B}} - \mathcal{C}_{\mathcal{B}\mathcal{B}}$, respectively. We have by Weyl's and Jensen's inequalities that $\nu_i + \rho_n \le \mu_i \le \nu_i + \rho_1$ and $|\rho_i| \le |\mathcal{B}| \max_{i,j\in\mathcal{B}} |\hat{C}_{ij} - \mathcal{C}_{ij}|$, respectively. It follows, under the event E_2 and the condition $|\mathcal{B}| = O(1)$ that $\hat{C}'(\boldsymbol{\beta})$ is invertible for sufficiently

large n. This enables us to define, with $\mathcal{B} = \text{supp}(\boldsymbol{\beta})$,

$$g^*(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathcal{B}} + \mathcal{C}'(\boldsymbol{\beta})^{-1}\{\tau(\boldsymbol{\beta})^{\mathcal{B}} - \mathcal{C}_{\mathcal{B}\mathcal{B}^c}\boldsymbol{\beta}_0^{\mathcal{B}^c}\},$$
$$\hat{g}(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathcal{B}} + \hat{C}'(\boldsymbol{\beta})^{-1}\{\tau(\boldsymbol{\beta})^{\mathcal{B}} - \hat{C}_{\mathcal{B}\mathcal{B}^c}\boldsymbol{\beta}_0^{\mathcal{B}^c} - n^{-1}T^{1/2}\boldsymbol{W}^{\mathcal{B}}\}.$$

Note that the second equality of (2.3) is equivalent to $g^*(\boldsymbol{\beta}^*) = \mathbf{0}$. To prove $\mathscr{K}_n \subset \hat{\mathscr{K}}_n$, it remains to show that there exists a $\hat{\boldsymbol{\beta}}$ satisfying (S1.2) and $\hat{g}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ with $\hat{\mathcal{A}}_n = \mathcal{A}_n$.

Define, for b_n satisfying $a_n \prec b_n \prec \lambda/n$, $G^* = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty < b_n, \text{supp}(\boldsymbol{\beta}) = \mathcal{A}_n\}$. Suppose that $0 < |\beta_j^*| \preceq b_n$ for $j \in \mathcal{A}_n$. With boundary cases excluded by (A1), we have $\beta_j^* = \beta_0^{(j)} - \big[\mathcal{C}'(\boldsymbol{\beta}^*)^{-1}\{\tau(\boldsymbol{\beta}^*)^{\mathcal{A}_n} - \mathcal{C}_{\mathcal{A}_n\mathcal{A}_n^c}\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\}\big]^{\{j\}} \asymp \lambda/n \succ b_n \succeq |\beta_j^*|$, a contradiction. It follows that for $j \in \mathcal{A}_n$, $|\beta_j^*| \succ b_n$ and hence $\text{sgn}(\beta_j^*) = \text{sgn}(\beta_j)$ for $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top \in G^*$. If $\kappa \prec \lambda/n$, setting $b_n \succeq \kappa$ ensures that $\beta_j^* \succ \kappa$, hence $|\beta_j^*| \wedge |\beta_j| > \kappa$ for all $j \in \mathcal{A}_n$ and $\boldsymbol{\beta} \in G^*$. If $\kappa \succeq \lambda/n$, then we have $\mathbf{1}\{0 < |\beta_j^*| \le \kappa\} = \mathbf{1}\{0 < |\beta_j| \le \kappa\}$ and $\mathbf{1}\{\kappa < |\beta_j^*| < \alpha\kappa\} = \mathbf{1}\{\kappa < |\beta_j| < \alpha\kappa\}$ for n sufficiently large, using the fact that $|\beta_j - \beta_j^*| < b_n \prec \lambda/n \preceq \kappa$ for $\boldsymbol{\beta} \in G^*$ and $|\beta_j^*| \neq \kappa$ or $\alpha\kappa$ by (A1). The above results together imply that $D(\boldsymbol{\beta}) = D(\boldsymbol{\beta}^*)$ and $\tau(\boldsymbol{\beta}) = \tau(\boldsymbol{\beta}^*)$ for $\boldsymbol{\beta} \in G^*$ and n sufficiently large.

Consider next $\hat{g}(\boldsymbol{\beta}) = \hat{g}(\boldsymbol{\beta}) - g^*(\boldsymbol{\beta}^*) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathcal{A}_n} + R(\hat{g})$, where

$$R(\hat{g}) = \mathcal{C}'(\boldsymbol{\beta}^*)^{-1}(\mathcal{C}_{\mathcal{A}_n\mathcal{A}_n^c} - \hat{C}_{\mathcal{A}_n\mathcal{A}_n^c})\boldsymbol{\beta}_0^{\mathcal{A}_n^c} - n^{-1}T^{1/2}\hat{C}'(\boldsymbol{\beta})^{-1}\boldsymbol{W}^{\mathcal{A}_n}$$
$$+ \{\hat{C}'(\boldsymbol{\beta})^{-1} - \mathcal{C}'(\boldsymbol{\beta}^*)^{-1}\}\{\tau(\boldsymbol{\beta})^{\mathcal{A}_n} - \hat{C}_{\mathcal{A}_n\mathcal{A}_n^c}\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\}.$$

Noting that $\|\boldsymbol{\beta}_0\|_1 = O(1)$ under (A3), we may bound each term in $R(\hat{g})$ by a

constant multiple of $\|\boldsymbol{W}\|_\infty + \max_{1\le j,j'\le p} |\hat{C}_{jj'} - \mathcal{C}_{jj'}|$, which implies $\|R(\hat{g})\|_\infty \preceq a_n \prec b_n$. Following Vrahatis (1989), we have, for $j \in \mathcal{A}_n$ and n sufficiently large, $\hat{g}(\hat{\boldsymbol{\beta}})^{\{j\}} \ge b_n - \|R(\hat{g})\|_\infty > 0$ or $\hat{g}(\hat{\boldsymbol{\beta}})^{\{j\}} \le -b_n + \|R(\hat{g})\|_\infty < 0$ according as $\beta_j - \beta_j^* = b_n$ or $-b_n$, respectively. It follows by applying Miranda's existence theorem to the continuous vector-valued function $\hat{g}$ that $\hat{g}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ for some $\hat{\boldsymbol{\beta}} \in G^*$.

To show that (S1.2) holds for $\hat{\boldsymbol{\beta}}$, note first that the first inequality in (2.3) holds strictly by (A1), for sufficiently large n. There thus exist constants $k_6 > 0$ and $k_7 \in (0,1)$ such that, for n sufficiently large and $j \in \mathcal{A}_n^c$,

$$\begin{aligned}
&\left|n^{-1}T^{1/2}W_j - \hat{C}_{\{j\}\mathcal{A}_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{A}_n} + \hat{C}_{\{j\}\mathcal{A}_n^c}\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\right| \\
&\le\ n^{-1}T^{1/2}|W_j| + \left|\hat{C}_{\{j\}\mathcal{A}_n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathcal{A}_n}\right| + \left|(\hat{C}_{\{j\}\mathcal{A}_n} - \mathcal{C}_{\{j\}\mathcal{A}_n})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathcal{A}_n}\right| \\
&\qquad + \left|(\hat{C}_{\{j\}\mathcal{A}_n^c} - \mathcal{C}_{\{j\}\mathcal{A}_n^c})\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\right| + \left|\mathcal{C}_{\{j\}\mathcal{A}_n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathcal{A}_n} - \mathcal{C}_{\{j\}\mathcal{A}_n^c}\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\right| \\
&\le\ k_6 a_n + \left|\mathcal{C}_{\{j\}\mathcal{A}_n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathcal{A}_n} - \mathcal{C}_{\{j\}\mathcal{A}_n^c}\boldsymbol{\beta}_0^{\mathcal{A}_n^c}\right| < k_6 a_n + k_7\lambda/(2n) < \lambda/(2n) \qquad \text{(S2.4)}
\end{aligned}$$

It follows that under the event $E_1 \cap E_2$ and for n sufficiently large, $\hat{\boldsymbol{\beta}}$ is a local minimum of (2.1) with $\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathcal{A}_\mathrm{n}$, hence $\mathscr{K}_n \subseteq \hat{\mathscr{K}_n}$.

To complete our proof, we shall show $\hat{\mathscr{K}_n} \subseteq \mathscr{K}_n$ under the event $E_1 \cap E_2$ for n sufficiently large. Recalling the definition of $f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)$ introduced in Section 2.3, we have

$$\left|f(\mathcal{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}} - f(\hat{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}\right| \preceq \max_{1\le i,i'\le p} |\hat{C}_{ii'} - \mathcal{C}_{ii'}| \preceq a_n, \quad j \in \mathcal{B},$$

on $E_1 \cap E_2$. It then follows by (A1) that

$$\min_{j\in\mathcal{B}} \left\{ \frac{\left|1 - f(\hat{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}/\beta_0^{(j)}\right|}{\mathbf{1}\{\beta_0^{(j)} \asymp \lambda/n\}}, \frac{\left|1 - |\beta_0^{(j)} - f(\hat{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}|/\kappa\right|}{\mathbf{1}\{\kappa \succeq \lambda/n\}}, \right.$$
$$\left. \left|1 - |\beta_0^{(j)} - f(\hat{C}, \mathcal{B}, \boldsymbol{\nu}, \boldsymbol{\beta}_0, k_0)^{\{j\}}|/(\alpha\kappa)\right| \right\} > c \tag{S2.5}$$

on $E_1 \cap E_2$, for sufficiently large n and some sufficiently small constant $c > 0$. Similarly, noting that

$$\begin{aligned}
&(2n/\lambda)\left|n^{-1}T^{1/2}W_j - (\hat{C}_{\{j\}\mathcal{B}} - \mathcal{C}_{\{j\}\mathcal{B}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{B}} + (\hat{C}_{\{j\}\mathcal{B}^c} - \mathcal{C}_{\{j\}\mathcal{B}^c})\boldsymbol{\beta}_0^{\mathcal{B}^c}\right| \\
\preceq\ & (n/\lambda) \max_{1\le i,i'\le p} |\hat{C}_{ii'} - \mathcal{C}_{ii'}| \preceq n a_n/\lambda \prec 1, \quad j \in \mathcal{B}^c,
\end{aligned}$$

and the last inequality under (A1), we have, for each $j \in \mathcal{B}^c$,

$$\left|(2n/\lambda)\left|n^{-1}T^{1/2}W_j - \hat{C}_{\{j\}\mathcal{B}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathcal{B}} + \hat{C}_{\{j\}\mathcal{B}^c}\boldsymbol{\beta}_0^{\mathcal{B}^c}\right| - 1\right| > c \tag{S2.6}$$

on $E_1 \cap E_2$, for sufficiently large n and some small constant $c > 0$. The conditions (S2.5) and (S2.6) rule out boundary cases corresponding to the sample versions of those excluded by (A1). Noting that $n^{-1}T^{1/2}W_j \prec \lambda/n$ for $j \in \hat{\mathcal{A}}_n$ and following similar arguments to those proving $\mathscr{K}_n \subseteq \hat{\mathscr{K}}_n$, we obtain that for sufficiently large n and any local minimum $\hat{\boldsymbol{\beta}}$ with $\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \hat{\mathcal{A}}_{\mathrm{n}}$, a solution $\boldsymbol{\beta}^*$ exists inside the hypercube $\hat{G} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_\infty < b_n, \mathrm{supp}(\boldsymbol{\beta}) = \hat{\mathcal{A}}_{\mathrm{n}}\}$, so that $\hat{\mathscr{K}}_n \subseteq \mathscr{K}_n$. □

Lemma S.2. *Under the conditions (A1) and (A2), each $\hat{\mathcal{A}}_n \in \hat{\mathscr{K}}_n$ is asymptotically noiseless, in the sense that $\hat{\mathcal{A}}_n$ does not depend, to first order, on $\boldsymbol{W}$.*

Proof of Lemma S.2. Define, for $j = 1, \ldots, p$,

$$\check{r}_j = \Big(\sum_{k \neq j} r_k^{-1} \Big)^{-1}, \quad b_j = \min\big\{ n/\sqrt{T}, \check{r}_j \big\}, \quad \tilde{W}_j = -2b_j \Big\{ n^{-1} T^{1/2} W_j - \sum_{k \neq j} r_k^{-1} \hat{C}_{jk} U_k \Big\}.$$

Note that $\tilde{W}_j \asymp 1$ and reduces to $-2W_j + o_p(1)$ or $2\check{r}_j \sum_{k \neq j} r_k^{-1} \hat{C}_{jk} U_k + o_p(1)$ according as $\check{r}_j \succ n/\sqrt{T}$ or $\prec n/\sqrt{T}$, respectively. Using (A2) and the fact that $v = U_j$ locally minimises the univariate function

$$n r_j^{-2} \hat{C}_{jj} v^2 + n(b_j r_j)^{-1} \tilde{W}_j v + \lambda\kappa\, q\left(|\beta_0^{(j)} + r_j^{-1} v| / \kappa \right),$$

we have that the function

$$g_j(v) \triangleq v + (2n\hat{C}_{jj})^{-1} r_j \lambda q'\big(|\beta_0^{(j)} + r_j^{-1} v| / \kappa \big) \mathrm{sgn}(\beta_0^{(j)} + r_j^{-1} v)$$

strictly increases in v and the equation $g_j(v) + (2\hat{C}_{jj} b_j)^{-1} r_j \tilde{W}_j = 0$ admits a unique solution at $v = U_j$, specified as follows for sufficiently large n.

(a) If $\big|\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j\big| \leq (2n\hat{C}_{jj})^{-1}\lambda$, then $\hat{\beta}_j = 0$, hence $U_j = -r_j \beta_0^{(j)}$.

(b) If $(2n\hat{C}_{jj})^{-1}\lambda < \big|\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j\big| \leq (2n\hat{C}_{jj})^{-1}\lambda + \kappa$, then

$$U_j = -(2\hat{C}_{jj})^{-1} r_j \Big\{ b_j^{-1} \tilde{W}_j + (\lambda/n)\, \mathrm{sgn}\big(\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j \big) \Big\}.$$

(c) If $(2n\hat{C}_{jj})^{-1}\lambda + \kappa < \big|\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j\big| < \alpha\kappa$, then

$$\begin{aligned} U_j = -(2\hat{C}_{jj})^{-1} r_j & \Big\{ 1 - \frac{\lambda}{2n\hat{C}_{jj}(\alpha - 1)\kappa} \Big\}^{-1} \\ & \times \Big\{ b_j^{-1} \tilde{W}_j - \frac{\lambda \beta_0^{(j)}}{n(\alpha - 1)\kappa} + \frac{\lambda\alpha}{n(\alpha - 1)} \mathrm{sgn}\big(\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j \big) \Big\}. \end{aligned}$$

(d) If $\left|\beta_0^{(j)} - (2\hat{C}_{jj}b_j)^{-1}\tilde{W}_j\right| \geq \alpha\kappa$, then $U_j = -(2\hat{C}_{jj}b_j)^{-1}r_j\tilde{W}_j$.

Case (a) corresponds to the subgradient condition (S1.2), while cases (b)–(d) correspond to (S1.3), thus leading to $\hat{\mathcal{A}}_n = \{j : |\beta_0^{(j)} - (2\hat{C}_{jj}b_j)^{-1}\tilde{W}_j| > (2n\hat{C}_{jj})^{-1}\lambda\}$. Recall that under either tail condition $(\mathcal{T}_i)$, W_j converges weakly to a nondegenerate distribution, for each $j = 1, \ldots, p$.

Suppose there exists a j' such that the event $\{j' \in \hat{\mathcal{A}}_n\}$ depends to first order on $\boldsymbol{W}$, so that $|\beta_0^{(j')} - (2\hat{C}_{j'j'}b_{j'})^{-1}\tilde{W}_{j'}| - (2n\hat{C}_{j'j'})^{-1}\lambda$ has an asymptotic leading term depending on $\boldsymbol{W}$. Then necessarily $\beta_0^{(j')} \preceq b_{j'}^{-1} \asymp \lambda/n$, which implies, for $j' \in \hat{\mathcal{A}}_n$, that $r_{j'} \asymp b_{j'} \asymp n/\lambda$ and $\tilde{W}_{j'} = 2b_{j'}\sum_{k\neq j'} r_k^{-1}\hat{C}_{j'k}U_k + o_p(1)$. Thus, there must exist $q \in \hat{\mathcal{A}}_n \setminus \{j'\}$ such that U_q has a leading term depending on $\boldsymbol{W}$, which implies $\check{r}_q \preceq r_{j'} \asymp n/\lambda \prec n/\sqrt{T}$, so that $\tilde{W}_q = 2b_q\sum_{k\neq q} r_k^{-1}\hat{C}_{qk}U_k + o_p(1)$. Repeating the above argument iteratively shows that the leading term of U_j $(j \in \hat{\mathcal{A}}_n)$ does not depend on $\boldsymbol{W}$, contradicting the assumption on the leading term of $|\beta_0^{(j')} - (2\hat{C}_{j'j'}b_{j'})^{-1}\tilde{W}_{j'}| - (2n\hat{C}_{j'j'})^{-1}\lambda$. This proves that $\hat{\mathcal{A}}_n$ is asymptotically noiseless. □

Lemma 1 then follows from (S2.3), Lemma S.1 and Lemma S.2.

S.2.2 Proof of Proposition 1

The proofs given in this and subsequent sections are all conducted under the event $\hat{\mathcal{A}} = \mathcal{A}$, which occurs with probability converging to 1.

Define $\beta_{max} = \max_j |\beta_0^{(j)}|$, $B_S = \sum_{j=1}^p r_j^{-1}$, and denote by r_{j^*} the slowest rate.

Then necessarily $\check{r}_k \asymp B_S^{-1}$ and $b_k \asymp b_0 \equiv \min\{n/\sqrt{T}, B_S^{-1}\} \preceq b_{j^*}$ for any $k \neq j^*$. It follows that $b_j \succeq b_0$ for all j.

For any $k \in \mathcal{A} \setminus \{j^*\}$, we have $B_S \succeq r_k^{-1} \succeq b_k^{-1} \asymp b_0^{-1} \succeq B_S \succeq r_{j^*}^{-1} \succeq r_k^{-1}$. In this case, we may assume w.l.o.g. that $j^* \in \mathcal{A}$ and $r_j \asymp B_S^{-1}$ for all $j \in \mathcal{A}$. It also follows that $|\mathcal{A}| \asymp 1$.

If $|\mathcal{A} \setminus \{j^*\}| \prec 1$, then $r_j^{-1} \asymp |\beta_0^{(j)}| \preceq \lambda/n$ for all $j \neq j^*$. Thus, $b_{j^*}^{-1} \asymp (\sqrt{T}/n) \vee \check{r}_{j^*}^{-1} \asymp (\sqrt{T}/n) \vee \sum_{j \neq j^*} |\beta_0^{(j)}|$. If $j^* \in \mathcal{A}$, then $B_S \succeq r_{j^*}^{-1} \succeq b_{j^*}^{-1} \asymp (\sqrt{T}/n) \vee (B_S - r_{j^*}^{-1})$, so that $r_{j^*}^{-1} \asymp B_S$ necessarily.

Summarising the above results, we conclude that

$$\begin{cases} r_j^{-1} \asymp |\beta_0^{(j)}| \preceq \lambda/n, & j \in \mathcal{A}^c, \\ r_0^{-1} \equiv r_j^{-1} = r_{j^*}^{-1} \asymp B_S \succeq (\sqrt{T}/n) \vee \sum_{k \in \mathcal{A}^c} |\beta_0^{(k)}|, & j \in \mathcal{A}, \end{cases}$$

if $|\mathcal{A}| \asymp 1$, and that $r_j^{-1} \asymp |\beta_0^{(j)}| \leq \beta_{max} \asymp r_{j^*}^{-1} \preceq \lambda/n$, for all j, if $|\mathcal{A}| \prec 1$. This proves part (i).

To prove part (ii), suppose on the contrary that $B_0 \succ \lambda/n$. It follows by noting $B_S \asymp \sum_{j \in \mathcal{A}^c} |\beta_0^{(j)}| + |\mathcal{A}| B_S$ that $|\mathcal{A}| \preceq 1$, so that $\sum_{j \in \mathcal{A}_0^c \cap \mathcal{A}} |\beta_0^{(j)}| \prec \lambda/n$. For any $j \in \mathcal{A} \setminus \{j^*\} \neq \emptyset$, we have

$$\lambda/n \succeq b_j^{-1} \asymp b_0^{-1} \succeq B_S \succeq \sum_{k \in \mathcal{A}^c} |\beta_0^{(k)}| = B_0 + \sum_{k \in \mathcal{A}^c \setminus \mathcal{A}_0^c} |\beta_0^{(k)}| - \sum_{k \in \mathcal{A}_0^c \cap \mathcal{A}} |\beta_0^{(k)}| \succeq B_0 \succ \lambda/n,$$

a contradiction. Thus $\hat{\boldsymbol{\beta}}$ is inconsistent for $\boldsymbol{\beta}_0$, which proves (ii).

If $|\beta_0^{(j)}| \succ \lambda/n$, then $|\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j| \succ \lambda/n$, which violates the condition

of case (a) of the equation $g_j(v) + (2\hat{C}_{jj}b_j)^{-1}r_j\tilde{W}_j = 0$ specified in the proof of Lemma S.2. It follows that $j \in \hat{\mathcal{A}}$, which proves part (iii).

S.2.3 Proof of Theorem S.1

We follow hereafter the notations used in the proof of Proposition 1. Referring to cases (a)–(d), specified in the proof of Lemma S.2, which characterise solutions to the equation $g_j(v)+(2\hat{C}_{jj}b_j)^{-1}r_j\tilde{W}_j = 0$, define $\mathscr{K}_b$, $\mathscr{K}_c$ and $\mathscr{K}_d$ to be subsets of $\{1,\ldots,p\}$ to which cases (b), (c) and (d) apply respectively, so that $\hat{\mathcal{A}} = \mathscr{K}_b \cup \mathscr{K}_c \cup \mathscr{K}_d = \mathcal{A} \in \limsup_{n\to\infty} \mathscr{K}_n$. Define $D = \operatorname{diag}\left(\mathbf{1}\{j \in \mathscr{K}_c\} : j = 1,\ldots,p\right)$ and

$$\tau = \frac{\lambda}{2n}\Bigg[\operatorname{sgn}\Big(\beta_0^{(j)} + (\sqrt{T}/n)\hat{C}_{jj}^{-1}W_j - \hat{C}_{jj}^{-1}\sum_{k\neq j} r_k^{-1}\hat{C}_{jk}U_k\Big)\Big(\mathbf{1}\{j \in \mathscr{K}_b\} + \frac{\alpha}{\alpha-1}\mathbf{1}\{j \in \mathscr{K}_c\}\Big) - \frac{\beta_0^{(j)}}{(\alpha-1)\kappa}\mathbf{1}\{j \in \mathscr{K}_c\} : j = 1,\ldots,p\Bigg].$$

Write $\boldsymbol{U} = [U_1,\ldots,U_p]^\top$. It follows from cases (a)–(d) and the proof of Proposition 1 that

$$\begin{cases} U_j = -\,r_j\beta_0^{(j)}, \quad j \in \mathcal{A}^c, \\ \boldsymbol{U}^{\mathcal{A}} = \Big\{\hat{C}_{\mathcal{A}\mathcal{A}} - \dfrac{\lambda}{2n(\alpha-1)\kappa}D_{\mathcal{A}\mathcal{A}}\Big\}^{-1}\big(r_0 n^{-1}T^{1/2}\boldsymbol{W}^{\mathcal{A}} + r_0\hat{C}_{\mathcal{A}\mathcal{A}^c}\boldsymbol{\beta}_0^{\mathcal{A}^c} - r_0\tau^{\mathcal{A}}\big). \end{cases} \tag{S2.7}$$

Note, by Proposition 1(ii), that $B_0 \preceq \lambda/n$, so that $B_U \succeq \lambda/n \succeq \psi \succeq B_0$.

We first consider the order of $B_S = \sum_{j=1}^p r_j^{-1}$. Suppose that $B_S \prec B_0$. For all $j \neq j^*$, we have $b_j^{-1} \asymp b_0^{-1} = \max\{\sqrt{T}/n, B_S\} \prec \lambda/n$. It follows that $\mathcal{A}_0^c \setminus \{j^*\} \subset \mathcal{A}^c$.

If $j^* \notin \mathcal{A}_0^c$, then $B_S \succeq \sum_{j\in\mathcal{A}_0^c} r_j^{-1} \asymp B_0$, a contradiction. If $j^* \in \mathcal{A}_0^c$, then the fact $b_{j^*}^{-1} \preceq b_0^{-1} \prec \lambda/n$ implies that $j^* \in \mathcal{A}^c$, leading again to $B_S \succeq B_0$, a contradiction. Thus we must have $B_S \succeq B_0$. Suppose that $B_S \prec (\lambda/n)\{1 - B_U/(\alpha\kappa)\}_+$, so that there exists $|\beta_0^{(j_U)}| = B_U < \alpha\kappa$. Since $B_U \succeq \lambda/n \succ B_S \succeq r_{j_U}^{-1}$, we have $j_U \in \mathcal{A}$. Note that $\check{r}_{j_U}^{-1} \preceq B_S \prec \lambda/n$, so that $b_{j_U}^{-1} \prec \lambda/n$. That $B_U < \alpha\kappa$ suggests that $j_U \in \mathscr{K}_b \cup \mathscr{K}_c$. If $j_U \in \mathscr{K}_b$, we have $B_S \succeq r_{j_U}^{-1} \succeq \lambda/n$, a contradiction. If $j_U \in \mathscr{K}_c$, we have

$$B_S \succeq r_{j_U}^{-1} \succeq -\frac{\lambda\beta_0^{(j_U)}}{n(\alpha-1)\kappa} + \frac{\lambda\alpha}{n(\alpha-1)}\operatorname{sgn}\big(\beta_0^{(j_U)}\big) = \frac{\lambda\alpha}{n(\alpha-1)}\Big(1 - \frac{B_U}{\alpha\kappa}\Big)\operatorname{sgn}\big(\beta_0^{(j_U)}\big) \succ B_S,$$

a contradiction. Thus we must have $B_S \succeq (\lambda/n)\{1 - B_U/(\alpha\kappa)\}_+$. Suppose that $B_S \succ \lambda/n$. For $j \neq j^*$, we have $b_j^{-1} \asymp b_0^{-1} \succeq B_S \succ \lambda/n$, so that $j \in \mathcal{A}$, which contradicts sparsity of $\hat{\boldsymbol{\beta}}$. The above results together imply that $(\sqrt{T}/n) \vee \psi \preceq B_S \preceq \lambda/n$. It also follows that $b_j^{-1} \preceq b_0^{-1} \asymp B_S \preceq \lambda/n$ for all j.

Let j_U be such that $|\beta_0^{(j_U)}| = B_U$. If $j_U \in \mathcal{A}^c$, then $r_{j_U}^{-1} \asymp B_U \asymp \lambda/n \succeq B_S \succeq r_{j^*}^{-1} \succeq r_{j_U}^{-1}$. If $j_U \in \mathcal{A}$, then $r_{j_U} = r_{j^*} = r_0$. It follows that $r_{j_U} \asymp r_{j^*}$ in all cases and we may assume without loss of generality that $j^* = j_U \in \mathcal{A}_0$.

For $j \in \mathcal{A}_0 \cap \mathcal{A}^c$, we have $\lambda/n \preceq |\beta_0^{(j)}| \asymp r_j^{-1} \preceq B_S \preceq \lambda/n$, which implies $|\beta_0^{(j)}| \asymp B_S \asymp \lambda/n$. For $j \in \mathcal{A}_0^c \cap \mathcal{A}$, we have $B_S \preceq \lambda/n \preceq b_j^{-1} \preceq r_j^{-1} \asymp r_0^{-1} \asymp B_S$, which implies $B_S \asymp \lambda/n$. Thus, if $B_S \prec \lambda/n$, or equivalently, $r_0 \succ n/\lambda$, then $\mathcal{A} = \mathcal{A}_0$.

That $r_0 \succ n/\lambda$ implies $\psi \prec \lambda/n$ follows immediately from the relation $\psi \preceq B_S \asymp r_0^{-1} \prec \lambda/n$.

Conversely, suppose $\psi \prec \lambda/n$, so that $B_0 \prec \lambda/n$ and $\lim_{n\to\infty} (\alpha\kappa)^{-1} B_U \geq 1$ neces-

sarily. To show that (S2.7) admits a solution with $r_0 \succ n/\lambda$ and $\mathcal{A} = \mathcal{A}_0$, it suffices to show, with U_j set to $-\|\beta_0^{(j)}\|_2^{-1}\beta_0^{(j)}$ for $j \in \mathcal{A}_0^c$, that $\big\|\tau^{\mathcal{A}_0}\big\|_1 \prec \lambda/n$ for $r_0 \succ n/\lambda$ and some partition $\mathcal{A}_0 = \mathscr{K}_b \cup \mathscr{K}_c \cup \mathscr{K}_d$.

Note, under (A2), that $\rho \triangleq \lim_{n\to\infty} \lambda\{2\underline{\mathcal{C}}n(\alpha-1)\kappa\}^{-1} < 1$, so that, for n sufficiently large,

$$(\alpha\kappa)^{-1}\big\{\lambda/(2n\hat{C}_{jj}) + \kappa\big\} \le \rho(\alpha-1)/\alpha + 1/\alpha + o(1) < (3+\rho)/4.$$

For all $j \in \mathcal{A}_0$, we have $b_j^{-1} \preceq b_0^{-1} \asymp (\sqrt{T}/n) \vee r_0^{-1} \prec \lambda/n$, and either $|\beta_0^{(j)}| \ge \alpha\kappa$ or

$$|\beta_0^{(j)}| < \alpha\kappa \quad \text{and} \quad 1 - |\beta_0^{(j)}|/(\alpha\kappa) \le 1 - B_U/(\alpha\kappa) \preceq n\psi/\lambda \prec 1.$$

It follows that for sufficiently large n,

$$|\beta_0^{(j)}| \ge \alpha\kappa\big\{1 - o(1)\big\} > \alpha\kappa(3+\rho)/4 > \lambda/(2n\hat{C}_{jj}) + \kappa,$$

so that $\mathcal{A}_0 \subset \mathscr{K}_c \cup \mathscr{K}_d$ and

$$\|\tau^{\mathcal{A}_0}\|_1 = \frac{\lambda\alpha}{2n(\alpha-1)} \sum_{j\in\mathcal{A}_0\cap\mathscr{K}_c} \Big|1 - \frac{|\beta_0^{(j)}|}{\alpha\kappa}\Big| \preceq \frac{\lambda}{n}\Big(1 - \frac{B_U}{\alpha\kappa}\Big)_+ \preceq \psi \preceq B_S \asymp r_0^{-1} \prec \lambda/n.$$

Moreover, it follows by non-triviality of $\boldsymbol{U}^{\mathcal{A}_0}$ and (S2.7) that $r_0^{-1} \preceq (\sqrt{T}/n) \vee \psi \preceq B_S \asymp r_0^{-1}$, which implies $r_0 \asymp (n/\sqrt{T}) \wedge \psi^{-1}$. The solution (S1.1) then follows by setting $\mathcal{A} = \mathcal{A}_0$ and $r_0 \asymp (n/\sqrt{T}) \wedge \psi^{-1}$ in (S2.7) under the condition $\psi \prec \lambda/n$.

On the other hand, setting $r_0 \asymp n/\lambda$ in (S2.7) reduces $\boldsymbol{U}^{\mathcal{A}}$ to

$$\Big\{\hat{C}_{\mathcal{A}\mathcal{A}} - \frac{\lambda}{2n(\alpha-1)\kappa} D_{\mathcal{A}\mathcal{A}}\Big\}^{-1}\big\{r_0\hat{C}_{\mathcal{A}\mathcal{A}^c}\boldsymbol{\beta}_0^{\mathcal{A}^c} - \tau^{\mathcal{A}}\big\} + o_p(1),$$

which has a non-random leading term. If the above leading term $\asymp 1$ and is therefore non-vanishing, it gives rise to a consistent sparse local minimum having support $\mathcal{A} \supset \{k : |\beta_0^{(k)}| \succ \lambda/n\}$ and the slowest possible componentwise convergence rates $r_j = r_0 \asymp n/\lambda$, for $j \in \mathcal{A}$.

Suppose, in addition, that $B_U \succ \lambda/n$, so that $\psi \prec \lambda/n \prec B_U$. Then we have $\mathcal{A}_0 \subset \mathcal{A}$, $r_0 \hat{C}_{\mathcal{A}\mathcal{A}^c} \boldsymbol{\beta}_0^{\mathcal{A}^c} \preceq (n/\lambda) B_0 \preceq (n/\lambda)\psi \prec 1$, and

$$\tau^{\{j\}} \asymp \begin{cases} \mathbf{1}\{j \in \mathscr{K}_b\} + \dfrac{\alpha}{\alpha - 1}\mathbf{1}\{j \in \mathscr{K}_c\}, & j \in \mathcal{A} \cap \mathcal{A}_0^c, \\ \mathbf{1}\{j \in \mathscr{K}_b\} + \mathbf{1}\{j \in \mathscr{K}_c\}\dfrac{\alpha}{\alpha - 1}\Big(1 - \dfrac{|\beta_0^{(j)}|}{\alpha\kappa}\Big) \prec 1, & j \in \mathcal{A} \cap \mathcal{A}_0. \end{cases}$$

It follows that if $\mathcal{A} = \mathcal{A}_0$, then $\tau^{\mathcal{A}} = o_p(1)$, contradicting non-triviality of $\boldsymbol{U}^{\mathcal{A}}$, and therefore $r_0 \succ n/\lambda$ necessarily.

To prove the last assertion of Theorem S.1, it suffices to compare two consistent sparse local minima with selected sets $\mathcal{A}_0$ and $\mathcal{A} \supsetneq \mathcal{A}_0$, respectively. Denote by $\boldsymbol{V} = [V_1, \ldots, V_p]^\top$ and $\tilde{\boldsymbol{V}} = [\tilde{V}_1, \ldots, \tilde{V}_p]^\top$ their corresponding solutions for $\boldsymbol{U}$, respectively, and by $\boldsymbol{R} = [R_1, \ldots, R_p]^\top$ and $\tilde{\boldsymbol{R}} = [\tilde{R}_1, \ldots, \tilde{R}_p]^\top$ their corresponding vectors of convergence rates, respectively. Note that

$$\begin{cases} R_j = R_0 \asymp (n/\sqrt{T}) \wedge \psi^{-1}, & j \in \mathcal{A}_0, \\ R_j \beta_0^{(j)} = -V_j, & j \in \mathcal{A}_0^c, \end{cases} \quad \text{and} \quad \begin{cases} \tilde{R}_j = \tilde{R}_0 \asymp n/\lambda, & j \in \mathcal{A}, \\ \tilde{R}_j \beta_0^{(j)} = -\tilde{V}_j, & j \in \mathcal{A}^c. \end{cases}$$

Under the condition $\psi \prec \lambda/n \prec B_U$, a comparison between the objective functions

at the two local minima gives

$$
\begin{aligned}
&\sum_{i=1}^{n}\big\{Y_i-\boldsymbol{X}_i^\top(\beta_0+\mathrm{diag}(\boldsymbol{R})^{-1}\boldsymbol{V})\big\}^2+\lambda\kappa\sum_{j=1}^{p}q(|\beta_0^{(j)}+R_j^{-1}V_j|/\kappa)\\
&\quad-\sum_{i=1}^{n}\big\{Y_i-\boldsymbol{X}_i^\top(\beta_0+\mathrm{diag}(\tilde{\boldsymbol{R}})^{-1}\tilde{\boldsymbol{V}})\big\}^2-\lambda\kappa\sum_{j=1}^{p}q(|\beta_0^{(j)}+\tilde{R}_j^{-1}\tilde{V}_j|/\kappa)\\
&=\ 2T^{1/2}\boldsymbol{W}^{\mathcal{A}_0^c\cap\mathcal{A}\top}\boldsymbol{\beta}_0^{\mathcal{A}_0^c\cap\mathcal{A}}-2T^{1/2}R_0^{-1}\boldsymbol{W}^{\mathcal{A}_0\top}\boldsymbol{V}^{\mathcal{A}_0}+n\boldsymbol{\beta}_0^{\mathcal{A}_0^c\top}\hat{C}_{\mathcal{A}_0^c\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c}+nR_0^{-2}\boldsymbol{V}^{\mathcal{A}_0\top}\hat{C}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{V}^{\mathcal{A}_0}\\
&\quad-2nR_0^{-1}\boldsymbol{\beta}_0^{\mathcal{A}_0^c\top}\hat{C}_{\mathcal{A}_0^c\mathcal{A}_0}\boldsymbol{V}^{\mathcal{A}_0}+\lambda\kappa\sum_{j\in\mathcal{A}_0}\big\{q(|\beta_0^{(j)}+R_0^{-1}V_j|/\kappa)-q(|\beta_0^{(j)}+\tilde{R}_0^{-1}\tilde{V}_j|/\kappa)\big\}\\
&\quad+2T^{1/2}\tilde{R}_0^{-1}\boldsymbol{W}^{\mathcal{A}\top}\tilde{\boldsymbol{V}}^{\mathcal{A}}-n\boldsymbol{\beta}_0^{\mathcal{A}^c\top}\hat{C}_{\mathcal{A}^c\mathcal{A}^c}\boldsymbol{\beta}_0^{\mathcal{A}^c}-n\tilde{R}_0^{-2}\tilde{\boldsymbol{V}}^{\mathcal{A}\top}\hat{C}_{\mathcal{A}\mathcal{A}}\tilde{\boldsymbol{V}}^{\mathcal{A}}\\
&\quad+2n\tilde{R}_0^{-1}\boldsymbol{\beta}_0^{\mathcal{A}^c\top}\hat{C}_{\mathcal{A}^c\mathcal{A}}\tilde{\boldsymbol{V}}^{\mathcal{A}}-\lambda\kappa\sum_{j\in\mathcal{A}\cap\mathcal{A}_0^c}q(\tilde{R}_0^{-1}|\tilde{V}_j|/\kappa)\{1+o_p(1)\}\\
&=\ -n\tilde{R}_0^{-2}\tilde{\boldsymbol{V}}^{\mathcal{A}\top}\hat{C}_{\mathcal{A}\mathcal{A}}\tilde{\boldsymbol{V}}^{\mathcal{A}}-\lambda\kappa\sum_{j\in\mathcal{A}\cap\mathcal{A}_0^c}q(\tilde{R}_0^{-1}|\tilde{V}_j|/\kappa)\{1+o_p(1)\}+O_p(n\psi^2+\sqrt{T}\lambda/n+\lambda B_0)\\
&\quad+\frac{\lambda}{2(\alpha-1)\kappa}\sum_{j\in\mathcal{A}_0}\Big\{\big(\alpha\kappa-|\beta_0^{(j)}+\tilde{R}_0^{-1}\tilde{V}_j|\big)_+^2-\big(\alpha\kappa-|\beta_0^{(j)}+R_0^{-1}V_j|\big)_+^2\Big\}\\
&\le\ -n\tilde{R}_0^{-2}\tilde{\boldsymbol{V}}^{\mathcal{A}\top}\hat{C}_{\mathcal{A}\mathcal{A}}\tilde{\boldsymbol{V}}^{\mathcal{A}}\{1+o_p(1)\}\\
&\quad+\frac{\lambda}{2(\alpha-1)\kappa}\sum_{j\in\mathcal{A}_0}\mathbf{1}\big\{\lim_{n\to\infty}|\beta_0^{(j)}|/(\alpha\kappa)=1\big\}O_p\big\{\lambda^2/n^2+(|\beta_0^{(j)}|-\alpha\kappa)^2\big\}\\
&=\ -n\tilde{R}_0^{-2}\tilde{\boldsymbol{V}}^{\mathcal{A}\top}\hat{C}_{\mathcal{A}\mathcal{A}}\tilde{\boldsymbol{V}}^{\mathcal{A}}\{1+o_p(1)\}+\mathbf{1}\{\lim_{n\to\infty}B_U/(\alpha\kappa)=1\}\big\{o_p(\lambda\alpha\kappa)+O_p(\lambda^3/(n^2\alpha\kappa))\big\}.
\end{aligned}
$$

If $\lim_{n\to\infty}B_U/(\alpha\kappa)>1$, then the second term in the last expression vanishes, so that the above difference between the two objective functions becomes strictly negative asymptotically, which proves the last assertion.

S.2.4 Proof of Theorem 2

Suppose that $\mathcal{A}_0 = \emptyset$, so that $\mathcal{A}_0^c = \{1, \ldots, p\}$, $B_U = \infty$, $\psi = B_0$ and $B_0 \preceq B_S \preceq \lambda/n$. Using the same arguments as in Section S.2.3, if $B_S \prec \lambda/n$, then necessarily $B_0 \prec \lambda/n$ and $\mathcal{A}^c = \mathcal{A}_0^c$, which in turn implies $B_S \asymp B_0 \prec \lambda/n$. Thus we have $B_0 \prec \lambda/n$ and $\mathcal{A}^c = \mathcal{A}_0^c$ if and only if $B_S \prec \lambda/n$.

For $j \in \mathcal{A}^c$, we have, for sufficiently large n, $\hat{\beta}_j - \beta_0^{(j)} = -\beta_0^{(j)} \prec \lambda/n$. If $\mathcal{A} \neq \emptyset$, then we have, for each $j \in \mathcal{A}$, $r_j \asymp n/\lambda$, so that $\hat{\beta}_j - \beta_0^{(j)} \asymp r_j^{-1} \asymp \lambda/n$.

Assume now $B_0 \prec \lambda/n$. To prove the existence of a zero local minimum, it suffices to show that $b_j^{-1} \prec \lambda/n$ for all j if we set $r_j \asymp |\beta_0^{(j)}|^{-1}$, for in this case we have $\left|\beta_0^{(j)} - (2\hat{C}_{jj} b_j)^{-1} \tilde{W}_j\right| \prec \lambda/n$ for all j, yielding a local minimum $\hat{\boldsymbol{\beta}} = \mathbf{0}$. This is accomplished by noting the fact that if $r_j \asymp |\beta_0^{(j)}|^{-1}$ for all j, then $B_S \asymp \|\boldsymbol{\beta}_0\|_1 = B_0$, so that, for all j,

$$b_j^{-1} = (\sqrt{T}/n) \vee \check{r}_j^{-1} \preceq (\sqrt{T}/n) \vee B_S \asymp (\sqrt{T}/n) \vee B_0 \prec \lambda/n.$$

To prove the last assertion, consider a consistent sparse local minimum with selected set $\mathcal{A} \neq \emptyset$. Denote by $\tilde{\boldsymbol{V}} = [\tilde{V}_1, \ldots, \tilde{V}_p]^\top$ and $\tilde{\boldsymbol{R}} = [\tilde{R}_1, \ldots, \tilde{R}_p]^\top$ its corresponding solution for $\boldsymbol{U}$ and the accompanying componentwise convergence rates, respectively. Noting that $\tilde{R}_j = \tilde{R}_0 \asymp n/\lambda$ for $j \in \mathcal{A}$ and $\tilde{R}_j \beta_0^{(j)} = -\tilde{V}_j$ for $j \in \mathcal{A}^c$, a comparison between the objective function values at the zero local minimum and the above

nonzero local minimum gives

$$\begin{aligned}
&\sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} \left\{ Y_i - \boldsymbol{X}_i^\top (\boldsymbol{\beta}_0 + \mathrm{diag}(\tilde{\boldsymbol{R}})^{-1} \tilde{\boldsymbol{V}}) \right\}^2 - \lambda\kappa \sum_{j=1}^{p} q(|\beta_0^{(j)} + \tilde{R}_j^{-1} \tilde{V}_j|/\kappa) \\
&= \; 2T^{1/2} \boldsymbol{W}^{\mathcal{A}_0^c \cap \mathcal{A}\top} \boldsymbol{\beta}_0^{\mathcal{A}_0^c \cap \mathcal{A}} + n \boldsymbol{\beta}_0^{\mathcal{A}_0^c \top} \hat{C}_{\mathcal{A}_0^c \mathcal{A}_0^c} \boldsymbol{\beta}_0^{\mathcal{A}_0^c} + 2T^{1/2} \tilde{R}_0^{-1} \boldsymbol{W}^{\mathcal{A}\top} \tilde{\boldsymbol{V}}^{\mathcal{A}} - n \boldsymbol{\beta}_0^{\mathcal{A}^c \top} \hat{C}_{\mathcal{A}^c \mathcal{A}^c} \boldsymbol{\beta}_0^{\mathcal{A}^c} \\
&\quad -n \tilde{R}_0^{-2} \tilde{\boldsymbol{V}}^{\mathcal{A}\top} \hat{C}_{\mathcal{A}\mathcal{A}} \tilde{\boldsymbol{V}}^{\mathcal{A}} + 2n \tilde{R}_0^{-1} \boldsymbol{\beta}_0^{\mathcal{A}^c \top} \hat{C}_{\mathcal{A}^c \mathcal{A}} \tilde{\boldsymbol{V}}^{\mathcal{A}} - \lambda\kappa \sum_{j \in \mathcal{A}} q(\tilde{R}_0^{-1} |\tilde{V}_j|/\kappa) \{1 + o_p(1)\} \\
&= \; -n \tilde{R}_0^{-2} \tilde{\boldsymbol{V}}^{\mathcal{A}\top} \hat{C}_{\mathcal{A}\mathcal{A}} \tilde{\boldsymbol{V}}^{\mathcal{A}} - \lambda\kappa \sum_{j \in \mathcal{A}} q(\tilde{R}_0^{-1} |\tilde{V}_j|/\kappa) \{1 + o_p(1)\} + O_p(\sqrt{T}\lambda/n + \lambda B_0) \\
&\leq \; -n \tilde{R}_0^{-2} \tilde{\boldsymbol{V}}^{\mathcal{A}\top} \hat{C}_{\mathcal{A}\mathcal{A}} \tilde{\boldsymbol{V}}^{\mathcal{A}} \{1 + o_p(1)\},
\end{aligned}$$

It follows that the zero local minimum has an objective function value strictly smaller than that of any nonzero consistent sparse local minimum, which proves the last assertion.

S.2.5 Proofs of Theorem S.2 and Corollary S.2

Proof of Theorem S.2. For any selected set $\hat{\mathcal{A}}$ given by a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$, its corresponding post-selection OLS estimator $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ satisfies (S1.4), with its nonzero components converging at a rate $\|\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}\|_1^{-1} \wedge (n/\sqrt{T}) \preceq n/\sqrt{T}$. The first assertion then follows by noting that if $\psi \prec \lambda/n$, then, by Theorem S.1, there exists a $\hat{\boldsymbol{b}}(\hat{\mathcal{A}})$ selecting $\hat{\mathcal{A}} = \mathcal{A}_0$ with probability converging to one.

Furthermore, if $\psi \prec \lambda/n \prec B_U$, then, by Theorem S.1 again, the support $\hat{\mathcal{A}}$ of any consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ contains $\mathcal{A}_0$ with probability converging to one, thereby entailing a convergence rate $\|\boldsymbol{\beta}_0^{\hat{\mathcal{A}}^c}\|_1^{-1} \wedge (n/\sqrt{T}) \succeq B_0^{-1} \wedge (n/\sqrt{T})$, which

proves the second assertion. □

Proof of Corollary S.2. Note that the conditions assumed on (B_0, B_U) imply $\psi \prec \lambda/n$. The corollary then follows directly from Theorem S.2 and (S1.4), where the convergence rate is given by $B_0^{-1} \wedge (n/\sqrt{T}) \asymp n/\sqrt{T}$ and the bias term vanishes because $(n/\sqrt{T})\mathscr{B}_2 \asymp (n/\sqrt{T})B_0 \prec 1$. □

S.2.6 Proofs of Theorems S.3 and S.4

Proof of Theorem S.3. If $B_0 \prec \lambda/n$ and either $\{1-B_U/(\alpha\kappa)\}_+ \prec \sqrt{T}/\lambda$ or $B_U \geq \alpha\kappa$, then $\psi \prec \lambda/n$, so that, by Theorem S.1, a consistent sparse local minimum $\hat{\boldsymbol{\beta}}$ exists with $\mathbb{P}(\mathrm{supp}(\hat{\boldsymbol{\beta}}) = \mathcal{A}_0) \to 1$. Note that $|\hat{C}_{ij} - \mathcal{C}_{ij}| \preceq \sqrt{T}/n$ $(i, j = 1, \ldots, p)$ under any tail condition $(\mathcal{T}_j)$.

If $\{1-B_U/(\alpha\kappa)\}_+ \prec \sqrt{T}/\lambda$ and $\alpha\kappa \succ \lambda/n$, then we have $nT^{-1/2}\phi^{\mathcal{A}_0}/r_0 \prec 1$ and, by (S1.1), that

$$\begin{aligned} nT^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)^{\mathcal{A}_0} &= \big\{\mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0} + O_p(n^{-1}T^{1/2}) + o(1)\big\}\big\{\boldsymbol{W}^{\mathcal{A}_0} + nT^{-1/2}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} + O_p(B_0) + o(1)\big\} \\ &\quad - nT^{-1/2}\mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\boldsymbol{\beta}_0^{\mathcal{A}_0^c} \\ &= \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + O_p(n^{-1}T^{1/2} + B_0) + o_p(1) = \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1). \end{aligned}$$

On the other hand, if $B_U \geq \alpha\kappa$, then $\Lambda_{\mathcal{A}_0\mathcal{A}_0} = \mathrm{diag}(\mathbf{0})$ and $\phi^{\mathcal{A}_0} = \mathbf{0}$, so that the above expansion remains valid. □

Proof of Theorem S.4. By Theorem S.1, if $\psi \prec \lambda/n$, selected sets $\hat{\mathcal{A}} \in \hat{\mathscr{K}}$ exist such

that $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}_0) \to 1$. Similar to the proof of Theorem S.3, it follows by (S1.4) and the fact $|\hat{C}_{ij} - \mathcal{C}_{ij}| \preceq \sqrt{T}/n$ $(i, j = 1, \ldots, p)$ that, for sufficiently large n,

$$\begin{aligned} nT^{-1/2}\{\hat{\boldsymbol{b}}(\hat{\mathcal{A}}) - \boldsymbol{\theta}_0\}^{\mathcal{A}_0} &= \hat{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + nT^{-1/2}\big(\hat{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\hat{C}_{\mathcal{A}_0\mathcal{A}_0^c} - \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\mathcal{C}_{\mathcal{A}_0\mathcal{A}_0^c}\big)\boldsymbol{\beta}_0^{\mathcal{A}_0^c} \\ &= \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + O_p(n^{-1}T^{1/2} + B_0) = \mathcal{C}^{-1}_{\mathcal{A}_0\mathcal{A}_0}\boldsymbol{W}^{\mathcal{A}_0} + o_p(1). \end{aligned}$$

□

Bibliography

Borovkov, A. A. (2008). Asymptotic analysis of random walks, Volume 118. Cambridge University Press.

Bühlmann, P. and S. Van De Geer (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

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

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. IEEE Transactions on Information Theory 57(8), 5467–5484.

Hoffmann-Jørgensen, J. (1994). Probability with a View Toward Statistics (1st ed.). Chapman and Hall.

Loh, P.-L. and M. J. Wainwright (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. The Journal of Machine Learning Research 16(1), 559–616.

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. The Annals of Statistics 45(6), 2455–2482.

Sarantsev, A. (2011). Tail asymptotic of sum and product of random variables with applications in the theory of extremes of conditionally gaussian processes. arXiv preprint arXiv:1107.3869.

Van de Geer, S., P. Bühlmann, and S. Zhou (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). Electronic Journal of Statistics 5, 688–749.

Van Der Vaart, A. W., A. van der Vaart, A. W. van der Vaart, and J. Wellner (1996). Weak convergence and empirical processes: with applications to statistics. Springer Science & Business Media.

Vrahatis, M. N. (1989). A short proof and a generalization of miranda's existence theorem. Proceedings of the American Mathematical Society 107(3), 701–703.