BIOMETRICS 000, 000–000 DOI: 000
000 0000

Supplementary Web Materials For
"Fused Lasso with the Adaptation of Parameter Ordering in Combining Multiple Studies with Repeated Measurements"

Fei Wang, Lu Wang, and Peter X.-K. Song

This paper has been submitted for consideration for publication in *Biometrics*

1. Regularity Conditions

ASSUMPTION 1: The parameter $\boldsymbol{\phi}^0$ lies in the interior of a compact set $\mathcal{B}$ and $\mathbf{g}_i(\boldsymbol{\phi})$ is unbiased, namely $E_0(\mathbf{g}_i(\boldsymbol{\phi})) = 0$, and continuously differentiable in $\boldsymbol{\phi}$.

ASSUMPTION 2: The meta extended score vector $\bar{\mathbf{g}}(\boldsymbol{\phi})$ converges to $E\mathbf{g}_i(\boldsymbol{\phi})$ in probability uniformly over $\mathcal{B}$ and $E\mathbf{g}_i(\boldsymbol{\phi})$ is continuous in $\boldsymbol{\phi}$.

ASSUMPTION 3: The sensitivity matrix $\partial\bar{\mathbf{g}}(\boldsymbol{\phi})$ converges to $E\{\partial\mathbf{g}_i(\boldsymbol{\phi})\}$ in probability uniformly in a neighbourhood $\mathcal{N}$ of $\boldsymbol{\phi}^0$. $E\{\partial\mathbf{g}_i(\boldsymbol{\phi})\}$ is continuous in $\boldsymbol{\phi}$ and $E\{\partial\mathbf{g}_i(\boldsymbol{\phi}^0)\} = \mathbf{G}$. Moreover, under the partition of $\mathbf{G} = (\mathbf{G}_{\mathcal{A}_0^c}, \mathbf{G}_{\mathcal{A}_0})^T$, $\mathbf{G}_{\mathcal{A}_0^c}$ corresponds to the nonzero component $\boldsymbol{\phi}_{\mathcal{A}_0^c}$ of $\boldsymbol{\phi}$.

ASSUMPTION 4: The weight matrix $\mathbf{C}(\boldsymbol{\phi})$ in the QIF $Q(\cdot)$ is positive definite and is continuous in $\boldsymbol{\phi} \in \mathcal{B}$ and $\mathbf{C}(\boldsymbol{\phi})$ converges to $\mathrm{cov}\{\mathbf{g}_i(\boldsymbol{\phi})\}$ in probability uniformly over $\mathcal{B}$. At $\boldsymbol{\phi} = \boldsymbol{\phi}^0$, let $\boldsymbol{\Sigma} = \mathrm{cov}\{\mathbf{g}_i(\boldsymbol{\phi}^0)\}$.

ASSUMPTION 5: For any $k \neq k'$ and $\phi_k^0 = \phi_{k'}^0 = 0$, the initial root-$n$ consistent estimators ϕ_k^* and $\phi_{k'}^*$ satisfy $\phi_k^* - \phi_{k'}^* = o_p(1)$.

Assumptions 1 - 4 are the standard regularity conditions required in the theory of Generalized Method of Moment estimation (Hansen, 1982). Assumption 5 is required for the estimator $\widehat{\boldsymbol{\phi}}_F$ to evaluate the ratio of two weights, which is, however, not required for the two proposed FLAPO estimators $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ and $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$. To establish finite-sample L_1-norm error bounds, we need two extra Assumptions 6 and 7 for λ and $\partial\mathbf{Q}(\boldsymbol{\phi})$.

ASSUMPTION 6: The tuning parameter λ satisfies $\lambda > J\|\partial\mathbf{Q}(\boldsymbol{\phi}^0)\|_\infty$ for a constant $J > 1$.

ASSUMPTION 7: For a given weight matrix $\mathbf{F}$, there exists a constant $\kappa > 0$ so that $Q(\boldsymbol{\phi}^0 + u) - Q(\boldsymbol{\phi}^0) + \partial\mathbf{Q}(\boldsymbol{\phi}^0)^T u \geqslant \kappa\|u\|_2^2$ for $u \in \mathbf{C}(\mathbf{F})$ defined below in (1).

Let $\widehat{u} = \widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0$, where $\widehat{\boldsymbol{\phi}}$ is the minimizer of the proposed penalized objective function when a weight matrix $\mathbf{F}$ is used in the penalty function. By Assumptions 1 and 4, we know $Q(\boldsymbol{\phi})$ is convex around $\boldsymbol{\phi}^0$, and therefore $Q(\widehat{u} + \boldsymbol{\phi}^0) \geqslant \partial \mathbf{Q}(\boldsymbol{\phi}^0)^T \widehat{u}$. Thus, Assumption 6 leads to

$$\lambda(\|\mathbf{F}\widehat{u} + \mathbf{F}\boldsymbol{\phi}^0\|_1 - \|\mathbf{F}\boldsymbol{\phi}^0\|_1) \leqslant -Q(\widehat{u} + \boldsymbol{\phi}^0) \leqslant \|\partial \mathbf{Q}(\boldsymbol{\phi}^0)\|_\infty \|\widehat{u}\|_1 \leqslant J^{-1}\lambda \|\widehat{u}\|_1.$$

It follows that $\|\mathbf{F}_{\mathcal{A}_0}\widehat{u}_{\mathcal{A}_0}\|_1 - \|\mathbf{F}_{\mathcal{A}_0^c}\widehat{u}_{\mathcal{A}_0^c}\|_1 \leqslant J^{-1}\|\widehat{u}\|_1$. Using this fact, we can define a set

$$\mathbf{C}(\mathbf{F}) = \{u \in \mathbf{R}^{Kp} : \|\mathbf{F}_{\mathcal{A}_0} u_{\mathcal{A}_0}\|_1 - \|\mathbf{F}_{\mathcal{A}_0^c} u_{\mathcal{A}_0^c}\|_1 \leqslant J^{-1}\|u\|_1\}. \tag{1}$$

When the Hessian matrix of $Q(\boldsymbol{\phi})$ exists, assumption 7 essentially requires the smallest eigenvalue of the Hessian matrix bounded away from 0 in a neighborhood of $\boldsymbol{\phi}^0$ defined by $\mathbf{C}(\mathbf{F})$.

2. Algorithm for optimization

Since the proposed penalties use adjacent contrasts, their optimization procedures appear more challenging than that of the popular lasso method. Our idea is to convert the optimization problem into a computationally more manageable setup in order to facilitate numerical calculation. In the following presentation, for convenience we focus on penalty $\widetilde{P}_e(\boldsymbol{\beta})$ in the algorithm; but the entire procedure can also be applicable to the other two penalties $\widetilde{P}(\boldsymbol{\beta})$ and $P(\boldsymbol{\beta})$ as well. As discussed in Section 3 of the paper, we begin by approximating QIF $Q(\boldsymbol{\beta})$ by a second-order Taylor expansion at an initial consistent estimate $\boldsymbol{\beta}^*$. These initial estimates may be obtained by performing routine generalized estimating equation analysis with one study at a time, where the estimation consistency holds when individual study mean models are correctly specified. Specifically, the second-order approximation to the objective function $\Phi(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + \lambda \widetilde{P}_e(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$ is given by

$$\Phi(\boldsymbol{\beta}) \approx Q_* + \left(\partial \mathbf{Q}_*^T\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \left(\partial^2 \mathbf{Q}_*\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \lambda \|\widetilde{\mathbf{D}}_e \boldsymbol{\beta}\|_1, \tag{2}$$

where Q_*, $\partial \mathbf{Q}_*$ and $\partial^2 \mathbf{Q}_*$ denote $Q(\boldsymbol{\beta}^*)$, the first-order and second-order derivatives of $Q(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^*$, respectively. Following the argument of Kim et al. (2009), let $\mathbf{z} = \widetilde{\mathbf{D}}_e \boldsymbol{\beta} \in R^{Kp}$

and rewrite the local quadratic minimization of (2) as follows:

$$\begin{cases} \min_{\boldsymbol{\beta}\in R^{Kp},\mathbf{z}\in R^{Kp}} \left\{Q_* + \left(\partial \mathbf{Q}_*^T\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left(\partial^2\mathbf{Q}_*\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \lambda\|\mathbf{z}\|_1\right\} \\ \text{subject to } \widetilde{\mathbf{D}}_e\boldsymbol{\beta} = \mathbf{z}. \end{cases}$$

It follows that the Lagrangian formulation takes the form:

$$L(\boldsymbol{\beta},\mathbf{z},\boldsymbol{\tau}) = Q_* + \left(\partial \mathbf{Q}_*^T\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^T\left(\partial^2\mathbf{Q}_*\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + \lambda\|\mathbf{z}\|_1 + \boldsymbol{\tau}^T(\widetilde{\mathbf{D}}_e\boldsymbol{\beta}-\mathbf{z}),$$

where $\boldsymbol{\tau} \in R_+^{Kp}$ is the Lagrangian multiplier. Being a function of $\boldsymbol{\beta}$, the above objective function $L(\boldsymbol{\beta},\mathbf{z},\boldsymbol{\tau})$ is actually minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}^* - (\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\boldsymbol{\tau})$, provided the existence of $(\partial^2\mathbf{Q}_*)^{-1}$. Moreover, the corresponding minimum is given by, up to a constant,

$$\min_{\boldsymbol{\beta}\in R^{Kp}} L(\boldsymbol{\beta},\mathbf{z},\boldsymbol{\tau}) = \boldsymbol{\tau}^T\widetilde{\mathbf{D}}_e\boldsymbol{\beta}^* - \frac{1}{2}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\boldsymbol{\tau})^T(\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\boldsymbol{\tau}).$$

Also, minimizing the above objective function $L(\boldsymbol{\beta},\mathbf{z},\boldsymbol{\tau})$ with respect to $\mathbf{z}$ gives the minimum:

$$\min_{\mathbf{z}\in R^{Kp}} L(\boldsymbol{\beta},\mathbf{z},\boldsymbol{\tau}) = \begin{cases} 0, & \text{if } \|\boldsymbol{\tau}\|_\infty < \lambda, \\ -\infty, & \text{otherwise,} \end{cases}$$

where $\|\cdot\|_\infty$ is the sup-norm. Therefore the dual optimization is to minimize the following (3) with respect to $\boldsymbol{\tau}$:

$$\begin{cases} \min_{\boldsymbol{\tau}\in R_+^{(K-1)p}} -\boldsymbol{\tau}^T\widetilde{\mathbf{D}}_e\boldsymbol{\beta}^* + \frac{1}{2}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\boldsymbol{\tau})^T(\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\boldsymbol{\tau}) \\ \text{subject to } \|\boldsymbol{\tau}\|_\infty < \lambda. \end{cases} \tag{3}$$

Having the solution $\hat{\boldsymbol{\tau}}$ of (3), we can update $\boldsymbol{\beta}$ via $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* - (\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}_e^T\hat{\boldsymbol{\tau}})$. In effect, the optimization required in (3) is a quadratic programming problem with boundedness restrictions, $\|\boldsymbol{\tau}\|_\infty < \lambda$, which can be easily solved by applying one of the standard convex optimization algorithms, e.g. the interior-point method (Nocedal and Wright, 2006).

3. Theorem

3.1 *Lemmas*

We first prove Lemma 1 given in Section 4 of the paper that is needed for Theorem 3, and then establish Lemma 2 that is needed for the proof of Theorem A in a later section 3.5.

Proof of Lemma 3.1. The estimated ordering of distinct parameter groups is determined by the estimated ordering of different parameters. Parameters in the same parameter group do not matter at all. This is because when $\beta^0_{k,l} = \beta^0_{k',l}$, for any $\epsilon \in (0, 1)$

$$\text{pr}(\delta\{\beta^*_{k',l} \geqslant \beta^*_{k,l}\} > \epsilon) = \text{pr}(\beta^*_{k',l} \geqslant \beta^*_{k,l}),$$

which depends on the asymptotic distributions of $\beta^*_{k',l}$ and $\beta^*_{k,l}$ and does not necessarily converge to 0. However, the estimated ordering of $\beta^0_{k,l}$ and $\beta^0_{k',l}$ when $\beta^0_{k,l} = \beta^0_{k',l}$ can not change the estimated ordering of the parameter group that $\beta^0_{k,l}$ and $\beta^0_{k',l}$ belong to.

Thus to prove this lemma, it is sufficient to show that when $\beta^0_{k,l} > \beta^0_{k',l}$,

$$\delta\{\beta^*_{k',l} \geqslant \beta^*_{k,l}\} \rightarrow \delta\{\beta^0_{k',l} \geqslant \beta^0_{k,l}\} = 0$$

in probability as $n \to \infty$. For any $\epsilon \in (0, 1)$,

$$\begin{aligned}\text{pr}(\delta\{\beta^*_{k',l} \geqslant \beta^*_{k,l}\} > \epsilon) &= \text{pr}(\beta^*_{k',l} \geqslant \beta^*_{k,l}) \\ &\leqslant \text{pr}\left\{|\beta^*_{k',l} - \beta^*_{k,l} - (\beta^0_{k',l} - \beta^0_{k,l})| \geqslant \beta^0_{k,l} - \beta^0_{k',l}\right\} \to 0,\end{aligned}$$

by the fact that $\beta^*_{k',l}$ and $\beta^*_{k,l}$ are root-n consistent. This completes the proof of Lemma 3.1.

For a general $m \times n$ matrix $\mathbf{A}$, we define $(1, 1)$ operator norm $\|\mathbf{A}\|_1$ for $\mathbf{A}$ by $\sup_{u \in R^n : \|u\|_1 = 1} \|\mathbf{A}u\|_1$.

LEMMA 2: *For any weight matrix* $\mathbf{W}$*,*

$$\|\mathbf{W}(\boldsymbol{\phi}^0 + u)\|_1 - \|\mathbf{W}\boldsymbol{\phi}^0\|_1 \geqslant \rho a_0^{-1/2} \|u_{\mathcal{A}_0}\|_1 - \|\mathbf{W}_{\mathcal{A}_0^c}\|_1 \|u_{\mathcal{A}_0^c}\|_1,$$

where $\rho = \{\rho_{min}(\mathbf{W}^T_{\mathcal{A}_0}\mathbf{W}_{\mathcal{A}_0})\}^{1/2}$*,* $\rho_{min}(\cdot) > 0$ *is the smallest nonzero eigenvalue of any square matrix,* $a_0 = |\mathcal{A}_0|$*, and* $\|\mathbf{W}_{\mathcal{A}_0^c}\|_1$ *is* $(1, 1)$ *operator norm for matrix* $\mathbf{W}_{\mathcal{A}_0^c}$*.*

Proof of Lemma 2.

$$\begin{aligned}\|\mathbf{W}(\boldsymbol{\phi}^0 + u)\|_1 - \|\mathbf{W}\boldsymbol{\phi}^0\|_1 =& \|\mathbf{W}_{\mathcal{A}_0} u_{\mathcal{A}_0}\|_1 + \|\mathbf{W}_{\mathcal{A}_0^c}(\boldsymbol{\phi}^0_{\mathcal{A}_0^c} + u_{\mathcal{A}_0^c})\|_1 - \|\mathbf{W}_{\mathcal{A}_0^c}\boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1 \\ \geqslant& \|\mathbf{W}_{\mathcal{A}_0} u_{\mathcal{A}_0}\|_1 - \|\mathbf{W}_{\mathcal{A}_0^c} u_{\mathcal{A}_0^c}\|_1 \\ \geqslant& \rho a_0^{-1/2} \|u_{\mathcal{A}_0}\|_1 - \|\mathbf{W}_{\mathcal{A}_0^c}\|_1 \|u_{\mathcal{A}_0^c}\|_1,\end{aligned}$$

where $\|\mathbf{W}_{\mathcal{A}_0^c} u_{\mathcal{A}_0^c}\|_1 \leqslant \|\mathbf{W}_{\mathcal{A}_0^c}\|_1 \|u_{\mathcal{A}_0^c}\|_1$ and $\|\mathbf{W}_{\mathcal{A}_0} u_{\mathcal{A}_0}\|_1 \geqslant \|\mathbf{W}_{\mathcal{A}_0} u_{\mathcal{A}_0}\|_2 \geqslant \rho a_0^{-1/2} \|u_{\mathcal{A}_0}\|_1$.

3.2 *Proof of Proposition 1*

Proof of Proposition 1. For notational simplicity, let $\widehat{\phi} = \widehat{\phi}_F$ and $\Phi(\phi) = Q(\phi) + \lambda P(\phi)$.

Part (a): Following the argument of Fan and Li (2001) and Peng and Fan (2004), we show for any $\epsilon > 0$, there exists a large L independent of n such that the following (4) holds for all n sufficiently large depending on ϵ

$$\mathrm{pr}\left\{\inf_{\|u\|_2=L} \Phi(\phi^0 + un^{-1/2}) > \Phi(\phi^0)\right\} > 1 - \epsilon. \tag{4}$$

This implies a local minimizer satisfying that $\widehat{\phi} - \phi^0 = O_p(n^{-1/2})$. Let $\phi = \phi^0 + un^{-1/2}$. By the Taylor expansion of $\bar{\mathbf{g}}(\phi)$ at ϕ^0, $\bar{\mathbf{g}}(\phi) = \bar{\mathbf{g}}(\phi^0) + \partial\bar{\mathbf{g}}(\widetilde{\phi})^T(\phi - \phi^0)$, where $\widetilde{\phi}$ is between ϕ and ϕ^0. Applying Lemma 2 to $\widetilde{\mathbf{F}}$ and $\overline{\mathbf{F}}$ respectively, we obtain

$$\begin{aligned}
&\Phi(\phi) - \Phi(\phi^0)\\
=&Q(\phi) + \lambda\|\widetilde{\mathbf{F}}\phi\|_1 + \lambda\|\overline{\mathbf{F}}\phi\|_1 - Q(\phi^0) - \lambda\|\widetilde{\mathbf{F}}\phi^0\|_1 - \lambda\|\overline{\mathbf{F}}\phi^0\|_1\\
=&n\bar{\mathbf{g}}(\phi^0)^T\mathbf{C}(\phi)^{-1}\bar{\mathbf{g}}(\phi^0) + 2n(\phi - \phi^0)^T\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1}\bar{\mathbf{g}}(\phi^0)\\
&+ n(\phi - \phi^0)^T\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1}\partial\bar{\mathbf{g}}(\widetilde{\phi})^T(\phi - \phi^0) - n\bar{\mathbf{g}}(\phi^0)^T\mathbf{C}(\phi^0)^{-1}\bar{\mathbf{g}}(\phi^0)\\
&+ \lambda\left(\|\widetilde{\mathbf{F}}\phi\|_1 + \|\overline{\mathbf{F}}\phi\|_1 - \|\widetilde{\mathbf{F}}\phi^0\|_1 - \|\overline{\mathbf{F}}\phi^0\|_1\right)\\
\geqslant&2u^T\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1}n^{1/2}\bar{\mathbf{g}}(\phi^0) + u^T\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1}\partial\bar{\mathbf{g}}(\widetilde{\phi})^Tu\\
&+ \lambda n^{-1/2}\left\{(\widetilde{\rho} + \overline{\rho})a_0^{-1/2}\|u_{\mathcal{A}_0}\|_1 - (\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1 + \|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1)\|u_{\mathcal{A}_0^c}\|_1\right\} + o_p(1)\\
=&O_p(\|u\|_2) + u^T\partial\bar{\mathbf{g}}(\phi^0)\mathbf{C}(\phi^0)^{-1}\partial\bar{\mathbf{g}}(\phi^0)^Tu + o_p(1),
\end{aligned}$$

where by Assumption 1 and the central limit theory, the unbiased estimating function $n^{1/2}\bar{\mathbf{g}}(\phi^0)$ is the order of $O_p(1)$, and

$$\begin{aligned}
&n\bar{\mathbf{g}}(\phi^0)^T\mathbf{C}(\phi)^{-1}\bar{\mathbf{g}}(\phi^0) - n\bar{\mathbf{g}}(\phi^0)^T\mathbf{C}(\phi^0)^{-1}\bar{\mathbf{g}}(\phi^0) = o_p(1)\\
&\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1}n^{1/2}\bar{\mathbf{g}}(\phi^0)\\
=&\left[\partial\bar{\mathbf{g}}(\widetilde{\phi})\mathbf{C}(\phi)^{-1} - E\{\partial\bar{\mathbf{g}}(\phi^0)\mathbf{C}(\phi^0)^{-1}\}\right]n^{1/2}\bar{\mathbf{g}}(\phi^0) + \mathbf{G}\mathbf{\Sigma}^{-1}n^{1/2}\bar{\mathbf{g}}(\phi^0)\\
=&o_p(1) + O_p(1),
\end{aligned}$$

$\widetilde{\rho} = \{\rho_{min}(\widetilde{\mathbf{F}}_{\mathcal{A}_0}^T \widetilde{\mathbf{F}}_{\mathcal{A}_0})\}^{1/2}$, $\overline{\rho} = \{\rho_{min}(\overline{\mathbf{F}}_{\mathcal{A}_0}^T \overline{\mathbf{F}}_{\mathcal{A}_0})\}^{1/2}$, and by the condition of this theorem $\lambda n^{-1/2} \to 0$. Thus, when $\|u\|_2 = L$ is large, $\Phi(\boldsymbol{\phi}) - \Phi(\boldsymbol{\phi}^0)$ is dominated by $u^T \partial \bar{\mathbf{g}}(\boldsymbol{\phi}^0)\mathbf{C}(\boldsymbol{\phi}^0)^{-1}\partial \bar{\mathbf{g}}(\boldsymbol{\phi}^0)^T u$, which is bounded below by a spectral lower bound $\|u\|_2^2 \lambda(\partial \bar{\mathbf{g}}(\boldsymbol{\phi}^0)\mathbf{C}(\boldsymbol{\phi}^0)^{-1}\bar{\mathbf{g}}(\boldsymbol{\phi}^0)^T)$, positive definite and independent of n. This implies (4) and the proof is completed.

Part (b): Without loss of generality, we may assume $p = 1$ for simplicity of exposition. The case of $p > 1$ can be proved similarly. Besides $\mathcal{A}_0$, we also define sets $\mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_{K-1}$ where $\mathcal{E}_0 = \{(1), (1,2), (1,2,3), \ldots, (1, \ldots, K)\}$, $\mathcal{E}_1 = \{2, \ldots, K\}$, $\mathcal{E}_2 = \{(2,3), \ldots, (k-1,k), (k,k+1) \ldots, (K-1,K)\}$, $\mathcal{E}_3 = \{(2,3,4), \ldots, (k-2,k-1,k), (k-1,k,k+1), (k,k+1,k+2), \ldots, (K-2, K-1, K)\}$, and $\mathcal{E}_{K-1} = \{(2,3,\ldots,K)\}$. By this definition, $\mathcal{E}_0$ corresponds to K regression coefficients represented by $\phi_1, \phi_1 + \phi_2, \ldots, \phi_1 + \cdots + \phi_K$ and $\mathcal{E}_1$ represents all differences of adjacent parameter pairs, namely $\phi_2, \ldots, \phi_K$.

Let $u = (u_1, \ldots, u_K)^T$ be a K-dimensional vector where $u_k \in [-\epsilon, \epsilon]$, $\epsilon > 0$, for $k = 1, \ldots, K$. Around the true $\boldsymbol{\phi}^0$, the objective function can be rewritten as

$$\begin{aligned}
\Phi(u) &= Q(\boldsymbol{\phi}^0 + un^{-1/2}) + \lambda \sum_{k \in \mathcal{E}_0} w_{1,\ldots,k} |\sum_{j=1}^{k} \phi_j| + \lambda \sum_{k \in \mathcal{E}_1} w_k (\phi_k^0 + u_k n^{-1/2}) \\
&\quad + \lambda \sum_{(k,k+1) \in \mathcal{E}_2} w_{k,k+1} \Big\{ \phi_k^0 + \phi_{k+1}^0 + (u_k + u_{k+1}) n^{-1/2} \Big\} + \cdots \\
&\quad + \lambda \sum_{(2,\ldots,K) \in \mathcal{E}_{K-1}} w_{2,\ldots,K} \Big\{ \phi_2^0 + \phi_3^0 + \cdots + \phi_K^0 + (u_2 + u_2 + \cdots + u_K) n^{-1/2} \Big\},
\end{aligned}$$

where $w_{1,\ldots,k}, w_k, w_{k,k+1}, \ldots, w_{2,\ldots,K}$ are exactly the weights defined in equation (2) in the paper but represented by $\phi_1, \ldots, \phi_K$. For example, $w_k = |\beta_{k+1}^* - \beta_k^*|^{-1} = |\phi_k^*|^{-1}$, $w_{k,k+1} = |\beta_{k+1}^* -$

$\beta^*_{k-1}|^{-1} = |\phi^*_{k+1} + \phi^*_k|^{-1}$, and $w_{2,3,\dots,K} = |\beta^*_K - \beta^*_1|^{-1} = |\phi^*_2 + \cdots + \phi^*_K|^{-1}$. Using $\Phi(u)$, we let

$$\begin{aligned} V(u) =& \Phi(u) - \Phi(\mathbf{0}) \\ =& Q(\boldsymbol{\phi}^0 + un^{-1/2}) - Q(\boldsymbol{\phi}^0) + \lambda \sum_{k\in\mathcal{E}_1} w_k(|\phi^0_k + u_k n^{-1/2}| - |\phi^0_k|) \\ & + \lambda \sum_{(k_1,k_2)\in\mathcal{E}_2} w_{k_1,k_2}\Big[|\phi^0_{k_1} + \phi^0_{k_2} + (u_{k_1} + u_{k_2})n^{-1/2}| - |\phi^0_{k_1} + \phi^0_{k_2}|\Big] + \cdots \\ & + \lambda \sum_{(1,2,\dots,K)\in\mathcal{E}_K} w_{1,2,\dots,K}\Big[|\phi^0_1 + \phi^0_2 + \cdots + \phi^0_K + (u_1 + u_2 \cdots + u_K)n^{-1/2}| \\ & - |\phi^0_1 + \phi^0_2 + \cdots + \phi^0_K|\Big] \\ =& Q(\boldsymbol{\phi}^0 + un^{-1/2}) - Q(\boldsymbol{\phi}^0) + \lambda I_1(u) + \cdots + \lambda I_K(u), \end{aligned} \tag{5}$$

which is minimized at $\widehat{u} = n^{1/2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$. We first consider the limit of $\lambda I_1(u), \dots, \lambda I_K(u)$.

(i) If $\phi^0_k = 0$ and $u_k \neq 0$,

$$\lambda w_k(|\phi^0_k + u_k n^{-1/2}| - |\phi^0_k|) = \lambda n^{-1/2} w_k |u_k| \to \infty$$

by $\lambda \to \infty$ and $(n^{-1/2}w_k)^{-1} = n^{1/2}\phi^*_k = O_p(1)$.

(ii) If $\phi^0_k \neq 0$ and $u_k \neq 0$,

$$\begin{aligned} & \lambda w_k(|\phi^0_k + u_k n^{-1/2}| - |\phi^0_k|) \\ & = \lambda n^{-1/2} w_k \left[(n^{1/2}\phi^0_k + u_k)\mathrm{sgn}(n^{1/2}\phi^0_k + u_k) - n^{1/2}\phi^0_k \mathrm{sgn}(\phi^0_k)\right] \to 0 \quad \text{in probability} \end{aligned}$$

by $\lambda n^{-1/2} \to 0$, w_k converges to $(\phi^0_k)^{-1}$ in probability and $\mathrm{sgn}(n^{1/2}\phi^0_k + u_k) \to \mathrm{sgn}(\phi^0_k) = 1$ by $\phi^0_k > 0$.

(iii) If $u_k = 0$,

$$\lambda w_k(|\phi^0_k + u_k n^{-1/2}| - |\phi^0_k|) = 0.$$

Let $u_{\mathcal{A}_0}$ is the subvector of u indexed by the set $\mathcal{A}_0$. Thus, we can write (i), (ii) and (iii) as

$$\lambda I_1(u) \to \begin{cases} 0 & \text{if } u_{\mathcal{A}_0} = \mathbf{0} \\ \infty & \text{if } u_{\mathcal{A}_0} \neq \mathbf{0} \end{cases} \quad \text{in probability.}$$

For $\lambda I_2(u)$, we check (i) if $\phi^0_{k_1} + \phi^0_{k_2} \neq 0$ and $u_{k_1} + u_{k_2} \neq 0$, (ii) if $\phi^0_{k_1} + \phi^0_{k_2} = 0$ and $u_{k_1} + u_{k_2} \neq 0$

and (iii) if $u_{k_1} + u_{k_2} = 0$ to show

$$\lambda I_2(u) \to \begin{cases} 0 & \text{if } u_{\mathcal{A}_0} = \mathbf{0} \\ \infty & \text{if } u_{\mathcal{A}_0} \neq \mathbf{0} \end{cases} \quad \text{in probability.}$$

Thus, we have that for $k = 2, \ldots, K$,

$$\lambda I_k(u) \to \begin{cases} 0 & \text{if } u_{\mathcal{A}_0} = \mathbf{0} \\ \infty & \text{if } u_{\mathcal{A}_0} \neq \mathbf{0} \end{cases} \quad \text{in probability.}$$

Now we consider the term $\Phi(\boldsymbol{\phi}^0 + un^{-1/2}) - \Phi(\boldsymbol{\phi}^0)$ in (5). Notice that

$$n^{1/2}\bar{\mathbf{g}}(\boldsymbol{\phi}^0 + un^{-1/2}) = n^{1/2}\bar{\mathbf{g}}(\boldsymbol{\phi}^0) + \partial\bar{\mathbf{g}}(\phi^*)^T u \to \Psi + \mathbf{G}^T u \quad \text{in distribution}$$

$$\mathbf{C}(\boldsymbol{\phi}^0 + un^{-1/2})^{-1} \to \boldsymbol{\Sigma}^{-1} \quad \text{in probability,}$$

where $\boldsymbol{\phi}^*$ is between $\boldsymbol{\phi}^0$ and $\boldsymbol{\phi}^0 + un^{-1/2}$, $n^{1/2}\bar{\mathbf{g}}(\boldsymbol{\phi}^0)$ converges to $\Psi \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ in distribution and $\partial\bar{\mathbf{g}}(\boldsymbol{\phi}^*)$ converges to $\mathbf{G}$ in probability. Thus as $n \to \infty$

$$\begin{aligned} \Phi(\boldsymbol{\phi}^0 + un^{-1/2}) - \Phi(\boldsymbol{\phi}^0) &\to (\Psi + \mathbf{G}^T u)^T \boldsymbol{\Sigma}^{-1} (\Psi + \mathbf{G}^T u) - \Psi^T \boldsymbol{\Sigma}^{-1} \Psi \\ &= 2\Psi^T \boldsymbol{\Sigma}^{-1} \mathbf{G}^T u + u^T \mathbf{G} \boldsymbol{\Sigma}^{-1} \mathbf{G}^T u, \end{aligned}$$

and

$$V(u) \to \begin{cases} 2\Psi^T \boldsymbol{\Sigma}^{-1} \mathbf{G}^T_{\mathcal{A}_0^c} u_{\mathcal{A}_0^c} + u^T_{\mathcal{A}_0^c} \mathbf{G}_{\mathcal{A}_0^c} \boldsymbol{\Sigma}^{-1} \mathbf{G}^T_{\mathcal{A}_0^c} u_{\mathcal{A}_0^c} & \text{if } u_{\mathcal{A}_0} = \mathbf{0} \\ \infty & \text{if } u_{\mathcal{A}_0} \neq \mathbf{0} \end{cases},$$

in distribution, which is minimized at $u_{\mathcal{A}_0^c} = -(\mathbf{G}_{\mathcal{A}_0^c} \boldsymbol{\Sigma}^{-1} \mathbf{G}^T_{\mathcal{A}_0^c})^{-1} \mathbf{G}_{\mathcal{A}_0^c} \boldsymbol{\Sigma}^{-1} \Psi$ and $u_{\mathcal{A}_0 0} = 0$.

By the part (a), we know $\widehat{u} = O_p(1)$. Then, according to the Lemma 5.1 and Proposition 4.2 in Geyer (1994), we have

$$n^{1/2}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) = \widehat{u} = \begin{pmatrix} \widehat{u}_{\mathcal{A}_0^c} \\ \widehat{u}_{\mathcal{A}_0} \end{pmatrix} \to \begin{pmatrix} -(\mathbf{G}_{\mathcal{A}_0^c} \boldsymbol{\Sigma}^{-1} \mathbf{G}^T_{\mathcal{A}_0^c})^{-1} \mathbf{G}_{\mathcal{A}_0^c} \boldsymbol{\Sigma}^{-1} \Psi \\ \mathbf{0} \end{pmatrix}.$$

in distribution as $n \to \infty$.

Part (c): Next we show the sparsity result, $\widehat{\mathcal{A}}_0 \to \mathcal{A}_0$ in probability, which is equivalent to

$$\text{pr}(k \in \widehat{\mathcal{A}}_0 \mid k \in \mathcal{A}_0) \to 1, \quad \text{pr}(k \in \widehat{\mathcal{A}}_0 \mid k \in \mathcal{A}_0^c) \to 0.$$

When $k \in \mathcal{A}_0^c$, the above normality results show $\text{pr}(k \in \widehat{\mathcal{A}}_0 \mid k \in \mathcal{A}_0^c) \to 0$. When $k \in \mathcal{A}_0$ but

$\widehat{\phi}_k \neq 0$, we consider the following KKT optimality condition,

$$\partial_k \Phi(\widehat{\boldsymbol{\phi}}) = \partial_k Q(\widehat{\boldsymbol{\phi}}) + b_k(\widehat{\boldsymbol{\phi}}),$$

and

$$\begin{aligned}
b_1(\widehat{\boldsymbol{\phi}}) =& \lambda w_1 \Big\{ \mathrm{sgn}(\widehat{\phi}_1) + \frac{w_{1,2}}{w_1} \partial_k |\sum_{j=1}^{2} \widehat{\phi}_j| + \cdots + \frac{w_{1,\ldots,K}}{w_1} \partial_k |\sum_{j=1}^{K} \widehat{\phi}_j| \Big\} \quad k = 1, \\
b_k(\widehat{\boldsymbol{\phi}}) =& \lambda w_k \Big\{ 1 + \sum_{k' \in \mathcal{E}_0, k' \geqslant k} \frac{w_{1,\ldots,k'}}{w_k} |\sum_{j=1}^{k'} \phi_j| + \frac{w_{k,k+1}}{w_k} (\widehat{\phi}_k + \widehat{\phi}_{k+1}) + \frac{w_{k-1,k}}{w_k} (\widehat{\phi}_{k-1} + \widehat{\phi}_k) \\
& + \cdots + \frac{w_{2,\ldots,K}}{w_k} (\widehat{\phi}_2 + \cdots + \widehat{\phi}_K) \Big\} \quad 1 < k < K, \\
b_K(\widehat{\boldsymbol{\phi}}) =& \lambda w_K \Big\{ 1 + \frac{w_{1,\ldots,K}}{w_K} |\sum_{j=1}^{K} \phi_j| + \frac{w_{K-1,K}}{w_K} (\widehat{\phi}_{K-1} + \widehat{\phi}_K) \\
& + \cdots + \frac{w_{2,\ldots,K}}{w_K} (\widehat{\phi}_2 + \cdots + \widehat{\phi}_K) \Big\} \quad k = K,
\end{aligned}$$

where $\partial_k Q(\boldsymbol{\phi}) = \partial Q(\boldsymbol{\phi}) / \partial \phi_k$ and ∂_k is the subdifferential operation for $\widehat{\phi}_k$. By the root-n consistency of $\widehat{\boldsymbol{\phi}}$, it can be obtained that

$$\begin{aligned}
\partial_k Q(\widehat{\boldsymbol{\phi}}) &= 2n \bar{\mathbf{g}}(\widehat{\boldsymbol{\phi}})^T \mathbf{C}(\widehat{\boldsymbol{\phi}})^{-1} \partial_k \bar{\mathbf{g}}(\widehat{\boldsymbol{\phi}}) + o_p(1) \\
&= 2n \bar{\mathbf{g}}(\boldsymbol{\phi}^0)^T \mathbf{C}(\widehat{\boldsymbol{\phi}})^{-1} \partial_k \bar{\mathbf{g}}(\widehat{\boldsymbol{\phi}}) + n(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0)^T \partial \mathbf{g}(\widetilde{\boldsymbol{\phi}})^T \mathbf{C}(\widehat{\boldsymbol{\phi}})^{-1} \partial_k \mathbf{g}(\widehat{\boldsymbol{\phi}}) + o_p(1) = O_p(n^{1/2}).
\end{aligned}$$

Next we consider the following two cases:

(i) Any weight w^*, whose calculation involves, besides ϕ_k^* for $k \in \mathcal{A}_0$, an initial estimate $\phi_{k'}^*$ for $k' \in \mathcal{A}_0^c$ and $k' \neq k$, is $o_p(w_k)$. For example if $k_2 \in \mathcal{A}_0^c$ and $k_2 \neq k$

$$\frac{w_{k,k_2}}{w_k} = \frac{n^{1/2} \phi_k^*}{n^{1/2} \phi_k^* + n^{1/2} (\phi_{k_2}^* - \phi_{k_2}^0) + n^{1/2} \phi_{k_2}^0} = O_p(n^{-1/2}).$$

(ii) Any weight w^*, whose calculation only involves, besides ϕ_k^* for $k \in \mathcal{A}_0$, an initial zero estimate $\phi_{k'}^*$ for $k' \in \mathcal{A}_0$ and $k' \neq k$, satisfies that

$$\frac{w^*}{w_k} = \zeta^* + o_p(1),$$

where ζ^* is a fraction with $0 < \zeta^* < 1$. For example, for $k_2 \in \mathcal{A}_0$ and $k_2 \neq k$, according to assumption 5,

$$\frac{w_{k,k_2}}{w_k} = \frac{\phi_k^*}{\phi_k^* + \phi_{k_2}^*} = \frac{1}{2} + o_p(1).$$

When $1 \in \mathcal{A}_0$ and $\widehat{\phi}_1 \neq 0$, there always exists k', $1 \leqslant k' \leqslant K$, such that $\phi_1^0 = \cdots = \phi_{k'}^0 = 0$ and we can obtain

$$\begin{aligned} b_1(\widehat{\boldsymbol{\phi}}) =& \lambda w_1 \left\{ \text{sgn}(\widehat{\phi}_1) + \frac{w_{1,2}}{w_1} \partial_1 \Big| \sum_{j=1}^{2} \widehat{\phi}_j \Big| + \cdots + \frac{w_{1,\ldots,K}}{w_1} \partial_1 \Big| \sum_{j=1}^{K} \widehat{\phi}_j \Big| \right\} \\ =& \lambda w_1 \left\{ \text{sgn}(\widehat{\phi}_1) + \frac{1}{2} \partial_1 \Big| \sum_{j=1}^{2} \widehat{\phi}_j \Big| + \cdots + \frac{1}{k'} \partial_1 \Big| \sum_{j=1}^{k'} \widehat{\phi}_j \Big| + O_p(n^{-1/2}) \right\}, \end{aligned}$$

where $\frac{1}{2}\partial_1 | \sum_{j=1}^{2} \widehat{\phi}_j | + \cdots + \frac{1}{k'}\partial_1 | \sum_{j=1}^{k'} \widehat{\phi}_j |$ cannot be an integer by the properties of partial sum of harmonic series. Thus

$$b_1(\widehat{\boldsymbol{\phi}}) n^{-1/2} = \frac{\lambda}{n^{1/2} w_1^{-1}} = \frac{\lambda}{O_p(1)} \to \infty.$$

When $k \in \mathcal{A}_0$, $1 < k \leqslant K$, and $\widehat{\phi}_k \neq 0$, we have

$$b_k(\widehat{\boldsymbol{\phi}}) = \lambda w_k \left\{ 1 + \sum_{k' \in \mathcal{E}_0, k' \geqslant k} \frac{w_{1,\ldots,k'}}{w_k} \partial_k \Big| \sum_{j=1}^{k'} \phi_j \Big| + \frac{w_{k,k+1}}{w_k} + \frac{w_{k-1,k}}{w_k} + \cdots + \frac{w_{2,\ldots,K}}{w_k} \right\},$$

where $\sum_{k' \in \mathcal{E}_0, k' \geqslant k} \frac{w_{1,\ldots,k'}}{w_k} \partial_k | \sum_{j=1}^{k'} \phi_j | + \frac{w_{k,k+1}}{w_k} + \frac{w_{k-1,k}}{w_k} + \cdots + \frac{w_{2,\ldots,K}}{w_k}$ converges in probability to a partial sum of a harmonic series, which can not be an integer. Thus,

$$b_k(\widehat{\boldsymbol{\phi}}) n^{-1/2} = \frac{\lambda}{n^{1/2} w_k^{-1}} = \frac{\lambda}{O_p(1)} \to \infty.$$

When $K \in \mathcal{A}_0$ and $\widehat{\phi}_K \neq 0$, similarly we can obtain

$$b_K(\widehat{\boldsymbol{\phi}}) n^{-1/2} = \frac{\lambda}{n^{1/2} w_K^{-1}} = \frac{\lambda}{O_p(1)} \to \infty.$$

Thus for $k \in \mathcal{A}_0$ and $\widehat{\phi}_k \neq 0$,

$$\partial_k \Phi(\widehat{\boldsymbol{\phi}}) n^{-1/2} = \partial_k Q(\widehat{\boldsymbol{\phi}}) n^{-1/2} + b_k(\widehat{\boldsymbol{\phi}}) n^{-1/2} = O_p(1) + \frac{\lambda}{n^{1/2} w_k^{-1}},$$

which is nonzero and converges to ∞ in probability. This implies

$$\text{pr}(k \in \widehat{A}_0^c \mid k \in \mathcal{A}_0) \to 0.$$

3.3 *Proof of Theorem 2*

Proof of Theorem 2. Part (a) and (b) can be proved in the exact same way as Proposition 1. For part (c), we only need to modify $b_k(\boldsymbol{\phi})$ for $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$, which is

$$b_1(\widehat{\boldsymbol{\phi}}) = \lambda w_1 \mathrm{sgn}(\widehat{\phi}_1)$$

$$b_k(\widehat{\boldsymbol{\phi}}) = \lambda w_k, \quad \text{for } k = 2, \ldots, K.$$

The result is proved by following the steps in Proposition 1.

3.4 *Proof of Theorem 3*

Proof of Theorem 3. Let us define a set $\mathcal{T} = \bigcap_{l=1}^{p} \{T_l^* = T_l\}$ and let $\widehat{\boldsymbol{\phi}}_{\mathcal{T}}$ be $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ when $\mathcal{T}$ occurs; otherwise the estimator is denoted by $\widehat{\boldsymbol{\phi}}_{\mathcal{T}^c}$. Thus $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ may be represented as $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} = \widehat{\boldsymbol{\phi}}_{\mathcal{T}} \delta\{\mathcal{T}\} + \widehat{\boldsymbol{\phi}}_{\mathcal{T}^c} \delta\{\mathcal{T}^c\}$, where $\delta\{\mathcal{T}\}$ is an indicator for whether $\mathcal{T}$ occurs. Then

$$n^{1/2}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} - \boldsymbol{\phi}^0) = n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}} - \boldsymbol{\phi}^0)\delta\{\mathcal{T}\} + n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}^c} - \boldsymbol{\phi}^0)\delta\{\mathcal{T}^c\}.$$

Note the following facts: (i) $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}} - \boldsymbol{\phi}^0) = O_p(1)$ by Theorem 2; (ii) $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}^c} - \boldsymbol{\phi}^0) = O_p(1)$ by part (a) of Proposition 1; and (iii) $\delta\{\mathcal{T}^c\} = o_p(1)$ and $\delta\{\mathcal{T}\} - 1 = o_p(1)$ by Lemma 1. It follows that $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} - \boldsymbol{\phi}^0)$ is $O_p(1)$. Furthermore, given $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}^c} - \boldsymbol{\phi}^0)\delta\{\mathcal{T}^c\} = o_p(1)$, Slutsky's Theorem ensures that $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} - \boldsymbol{\phi}^0)$ and $n^{1/2}(\widehat{\boldsymbol{\phi}}_{\mathcal{T}} - \boldsymbol{\phi}^0)$ converge weakly to the same distribution as n goes to infinity. Let $\widehat{\mathcal{A}}_{0e}$ denote the estimated configuration of parameter homogeneity given by equation (10) in the paper based on the estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$, then we have

$$\mathrm{pr}(\widehat{\mathcal{A}}_{0e} = \mathcal{A}_0) = \mathrm{pr}\left(\widehat{\mathcal{A}}_{0e} = \mathcal{A}_0 \mid \mathcal{T}\right) \mathrm{pr}\,(\mathcal{T}) + \mathrm{pr}\left(\widehat{\mathcal{A}}_{0e} = \mathcal{A}_0 \mid \mathcal{T}^c\right) \mathrm{pr}\,(\mathcal{T}^c) \to 1 \text{ as } n \to \infty.$$

Summarizing the above outline of arguments, Theorem 3 is proved.

3.5 *Finite-sample error bounds*

Here we establish finite-sample error bounds for the proposed three estimators. Our derivation is made along the lines similar to that given by Negahban et al. (2009) for general M-estimators under the lasso penalty. Although the three estimators $\widehat{\boldsymbol{\phi}}_F$, $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ and $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ share the same asymptotic properties, $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ turns out to have smaller finite-sample error bounds than $\widehat{\boldsymbol{\phi}}_F$, when weights are properly assigned to zero parameters in $\boldsymbol{\phi}^0_{\mathcal{A}_0}$. Let $\rho_{min}(\mathbf{A}) > 0$ generically denote the smallest nonzero eigenvalue of a square matrix $\mathbf{A}$, and let $\widetilde{\rho} = \{\rho_{min}(\widetilde{\mathbf{F}}^T_{\mathcal{A}_0}\widetilde{\mathbf{F}}_{\mathcal{A}_0})\}^{1/2}$ and $\overline{\rho} = \{\rho_{min}(\overline{\mathbf{F}}^T_{\mathcal{A}_0}\overline{\mathbf{F}}_{\mathcal{A}_0})\}^{1/2}$. For a general $m \times n$ matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_1$ to denote $(1,1)$ operator norm defined by $\sup_{u \in R^n : \|u\|_1 = 1} \|\mathbf{A}u\|_1$.

The following Theorem A establishes error bounds for estimators of nonzero and zero parameters $\boldsymbol{\phi}^0_{\mathcal{A}_0^c}$ and $\boldsymbol{\phi}^0_{\mathcal{A}_0}$, respectively, when the parameter ordering is known.

THEOREM A: *Let $r = a_0/b_0$, with $a_0 = card(\mathcal{A}_0)$ and $b_0 = card(\mathcal{A}_0^c)$. Under assumptions 1-5 and 6-7 given in the Supplementary Materials, we have the following finite-sample L_1-norm error bounds.*

(a) The estimator $\widehat{\boldsymbol{\phi}}_F$ satisfies:

$$
\begin{aligned}
\|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0}\|_1 &\leqslant \lambda \frac{-\omega_{\mathcal{A}_0} + \left(\omega^2_{\mathcal{A}_0} + \varphi_{\mathcal{A}_0}\varphi^{-1}_{\mathcal{A}_0^c}\omega^2_{\mathcal{A}_0^c}\right)^{1/2}}{2\varphi_{\mathcal{A}_0}} \\
&\leqslant \lambda \frac{-2\omega_{\mathcal{A}_0}\delta\{\omega_{\mathcal{A}_0} < 0\} + r^{1/2}\omega_{\mathcal{A}_0^c}}{2\varphi_{\mathcal{A}_0}} \\
\|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1 &\leqslant \lambda \frac{\omega_{\mathcal{A}_0^c} + \left(\omega^2_{\mathcal{A}_0^c} + \varphi_{\mathcal{A}_0^c}\varphi^{-1}_{\mathcal{A}_0}\omega^2_{\mathcal{A}_0}\delta\{\omega_{\mathcal{A}_0} < 0\}\right)^{1/2}}{2\varphi_{\mathcal{A}_0^c}} \\
&\leqslant \lambda \frac{2\omega_{\mathcal{A}_0^c} - r^{-1/2}\omega_{\mathcal{A}_0}\delta\{\omega_{\mathcal{A}_0} < 0\}}{2\varphi_{\mathcal{A}_0^c}},
\end{aligned}
\tag{6}
$$

where $\omega_{\mathcal{A}_0} = (\widetilde{\rho} + \overline{\rho})a_0^{-1/2} - J^{-1}$, $\omega_{\mathcal{A}_0^c} = \|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1 + \|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1 + J^{-1} > 0$, $\varphi_{\mathcal{A}_0} = \kappa/a_0$, and $\varphi_{\mathcal{A}_0^c} = \kappa/b_0$. Two constants $J > 1$ and $\kappa > 0$ are given in assumptions 6 and 7, respectively, in the Supplementary Materials.

(b) For the estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$, the error bounds of $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}\mathcal{A}_0}\|_1$ and $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1$ satisfy inequalities in (6) *with $\omega_{\mathcal{A}_0} = \widetilde{\rho}a_0^{-1/2} - J^{-1}$ and $\omega_{\mathcal{A}_0^c} = \|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1 + J^{-1} > 0$.*

(c) When $a_0^{1/2} < \widetilde{\rho}J$, we have $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1 < \|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1$, and $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}\mathcal{A}_0}\|_1 < \|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0}\|_1$.

We have a few remarks on the results of finite sample error bounds given in Theorem A.

Remark 1: For large n, with a probability close to 1 we can estimate the parameter ordering, so the estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ also satisfies $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1 < \|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1$, and $\|\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e\mathcal{A}_0}\|_1 < \|\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0}\|_1$. This theoretical result is confirmed numerically in our simulation study.

Remark 2: $(\widetilde{\rho} + \overline{\rho})a_0^{-1/2}$ may be regarded as the "average smallest weight" for zero coefficients in $\boldsymbol{\phi}_{\mathcal{A}_0}$, while $\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1 + \|\bar{\mathbf{F}}_{\mathcal{A}_0^c}\|_1$ may be regarded as "the largest weight" for nonzero coefficients in $\boldsymbol{\phi}_{\mathcal{A}_0^c}$. If $\omega_{\mathcal{A}_0}$ is nonnegative, namely $(\widetilde{\rho} + \overline{\rho})a_0^{-1/2} \geqslant J^{-1}$, then $\delta\{\omega_{\mathcal{A}_0} < 0\} = 0$, implying that neither the error bounds for $\|\widehat{\boldsymbol{\phi}}_{\mathcal{A}_0}\|$ nor those for $\|\widehat{\boldsymbol{\phi}}_{\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1$ depends on $\omega_{\mathcal{A}_0}$. However, if $\omega_{\mathcal{A}_0}$ is negative or $(\widetilde{\rho} + \overline{\rho})a_0^{-1/2} < J^{-1}$, then $\delta\{\omega_{\mathcal{A}_0} < 0\} = 1$, and therefore the two error bounds are inversely proportional to $\omega_{\mathcal{A}_0}$.

It is interesting to note that the former case ($\omega_{\mathcal{A}_0} \geqslant 0$) may occur for large n. This is because in this case the weights for zero parameters in $\boldsymbol{\phi}_{\mathcal{A}_0}$ diverge, so both $\overline{\rho}$ and $\widetilde{\rho}$ tend to ∞, leading to positive $\omega_{\mathcal{A}_0}$. Thus, for large n, the two error bounds are smaller for $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ than $\widehat{\boldsymbol{\phi}}_F$. The latter case may occur when $b_0 = \text{card}(\mathcal{A}_0^c)$ is large, i.e. the pattern of unequal parameters becomes more complex.

Remark 3: the two error bounds in (6) are both proportional to the weight $\omega_{\mathcal{A}_0^c}$. This implies that large weights used in the penalty for nonzero parameters in $\mathcal{A}_0^c$ may weaken the finite sample performance of the proposed FLAPO method.

Theorem A and the above remarks provide a theoretical basis to apply the penalty $\tilde{P}_e(\cdot)$ in practice, as $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ enjoys smaller error bounds than $\widehat{\boldsymbol{\phi}}_F$ for large n. These properties are further examined and confirmed by the simulation studies in Section 5 of the paper.

Proof of Theorem A.

We here provide the proof of part (a), and proofs of part (b) and part (c) are trivial consequences from part (a), so the detail is omitted. Let $\widehat{\boldsymbol{\phi}} = \widehat{u} + \boldsymbol{\phi}^0$. First, according to the definition of the first estimator $\widehat{\boldsymbol{\phi}}$ using all possible pairwise differences in the penalty, $\Phi(\boldsymbol{\phi}^0 + \widehat{u}) - \Phi(\boldsymbol{\phi}^0) \leqslant 0$. Next we derive the upper bound of $\widehat{u}$ from $\Phi(\boldsymbol{\phi}^0 + \widehat{u}) - \Phi(\boldsymbol{\phi}^0)$. By Assumptions 6 and 7, for $\widehat{u} \in \mathbf{C}(\mathbf{F})$, we have

$$Q(\boldsymbol{\phi}^0+\widehat{u})-Q(\boldsymbol{\phi}^0)\geqslant -\frac{\lambda}{J}\|\widehat{u}\|_1+\kappa\|\widehat{u}\|_2^2, \tag{7}$$

We also can bound $\|\widetilde{\mathbf{F}}(\boldsymbol{\phi}^0+\widehat{u})\|_1-\|\widetilde{\mathbf{F}}\boldsymbol{\phi}^0\|_1$ from below as follows,

$$\begin{aligned}\|\widetilde{\mathbf{F}}(\boldsymbol{\phi}^0+\widehat{u})\|_1-\|\widetilde{\mathbf{F}}\boldsymbol{\phi}^0\|_1 =&\|\widetilde{\mathbf{F}}_{\mathcal{A}_0}\widehat{u}_{\mathcal{A}_0}\|_1+\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}(\boldsymbol{\phi}^0_{\mathcal{A}_0^c}+\widehat{u}_{\mathcal{A}_0^c})\|_1-\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\boldsymbol{\phi}^0_{\mathcal{A}_0^c}\|_1\\ \geqslant&\|\widetilde{\mathbf{F}}_{\mathcal{A}_0}\widehat{u}_{\mathcal{A}_0}\|_1-\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\widehat{u}_{\mathcal{A}_0^c}\|_1\\ \geqslant&\widetilde{\rho}a_0^{-1/2}\|\widehat{u}_{\mathcal{A}_0}\|_1-\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1\|\widehat{u}_{\mathcal{A}_0^c}\|_1,\end{aligned}$$

where $\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\widehat{u}_{\mathcal{A}_0^c}\|_1\leqslant\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1\|\widehat{u}_{\mathcal{A}_0^c}\|_1$, $\|\widetilde{\mathbf{F}}_{\mathcal{A}_0}\widehat{u}_{\mathcal{A}_0}\|_1\geqslant\|\widetilde{\mathbf{F}}_{\mathcal{A}_0}\widehat{u}_{\mathcal{A}_0}\|_2\geqslant\widetilde{\rho}a_0^{-1/2}\|\widehat{u}_{\mathcal{A}_0}\|_1$. Similarly we obtain

$$\|\overline{\mathbf{F}}(\boldsymbol{\phi}^0+\widehat{u})\|_1-\|\overline{\mathbf{F}}\boldsymbol{\phi}^0\|_1\geqslant\overline{\rho}a_0^{-1/2}\|\widehat{u}_{\mathcal{A}_0}\|_1-\|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1\|\widehat{u}_{\mathcal{A}_0^c}\|_1.$$

Therefore

$$\begin{aligned}&\|\widetilde{\mathbf{F}}(\boldsymbol{\phi}^0+\widehat{u})\|_1-\|\widetilde{\mathbf{F}}\boldsymbol{\phi}^0\|_1+\|\overline{\mathbf{F}}(\boldsymbol{\phi}^0+\widehat{u})\|_1-\|\overline{\mathbf{F}}\boldsymbol{\phi}^0\|_1\\ \geqslant&(\widetilde{\rho}+\overline{\rho})a_0^{-1/2}\|\widehat{u}_{\mathcal{A}_0}\|_1-(\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1+\|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1)\|\widehat{u}_{\mathcal{A}_0^c}\|_1.\end{aligned} \tag{8}$$

(7) and (8) imply

$$\begin{aligned}-\frac{\lambda}{J}\|\widehat{u}\|_1+\kappa\|\widehat{u}\|_2^2+\lambda(\widetilde{\rho}+\overline{\rho})a_0^{-1/2}\|\widehat{u}_{\mathcal{A}_0}\|_1-\lambda(\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1+\|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1)\|\widehat{u}_{\mathcal{A}_0^c}\|_1\leqslant 0\\ \kappa\|\widehat{u}_{\mathcal{A}_0}\|_2^2+\kappa\|\widehat{u}_{\mathcal{A}_0^c}\|_2^2+\lambda\omega_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1-\lambda\omega_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1\leqslant 0\\ \varphi_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1^2+\varphi_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1^2+\lambda\omega_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1-\lambda\omega_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1\leqslant 0,\end{aligned} \tag{9}$$

where $\varphi_{\mathcal{A}_0}=\kappa a_0^{-1}$, $\varphi_{\mathcal{A}_0^c}=\kappa b_0^{-1}$, $\omega_{\mathcal{A}_0}=(\widetilde{\rho}+\overline{\rho})a_0^{-1/2}-J^{-1}$, $\omega_{\mathcal{A}_0^c}=\|\widetilde{\mathbf{F}}_{\mathcal{A}_0^c}\|_1+\|\overline{\mathbf{F}}_{\mathcal{A}_0^c}\|_1+J^{-1}>0$ and $\omega_{\mathcal{A}_0}+\omega_{\mathcal{A}_0^c}>0$. Next we solve (9) for two cases:

(i) $\omega_{\mathcal{A}_0}\geqslant 0$.

(i.a) We solve (9) for $\|\widehat{u}_{\mathcal{A}_0}\|_1$ when $\|\widehat{u}_{\mathcal{A}_0^c}\|_1$ is fixed. It is easy to derive two inequalities, namely,

$$b^2-4c\geqslant 0,\quad -b+(b^2-4c)^{1/2}\geqslant 0,$$

where $b = \frac{\lambda\omega_{\mathcal{A}_0}}{\varphi_{\mathcal{A}_0}}$ and $c = \frac{1}{\varphi_{\mathcal{A}_0}}(\varphi_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1^2 - \lambda\omega_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1)$. The two inequalities imply that

$$\|\widehat{u}_{\mathcal{A}_0^c}\|_1 \leqslant \min\left\{\lambda\frac{\omega_{\mathcal{A}_0^c} + (\omega_{\mathcal{A}_0^c}^2 + \varphi_{\mathcal{A}_0^c}\varphi_{\mathcal{A}_0}^{-1}\omega_{\mathcal{A}_0}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}, \lambda\frac{\omega_{\mathcal{A}_0^c}}{\varphi_{\mathcal{A}_0^c}}\right\} = \lambda\frac{\omega_{\mathcal{A}_0^c}}{\varphi_{\mathcal{A}_0^c}}$$
$$\|\widehat{u}_{\mathcal{A}_0^c}\|_1 \geqslant \max\left\{\lambda\frac{\omega_{\mathcal{A}_0^c} - (\omega_{\mathcal{A}_0^c}^2 + \varphi_{\mathcal{A}_0^c}\varphi_{\mathcal{A}_0}^{-1}\omega_{\mathcal{A}_0}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}, 0\right\} = 0.$$

Thus we have

$$0 \leqslant \|\widehat{u}_{\mathcal{A}_0^c}\|_1 \leqslant \lambda\frac{\omega_{\mathcal{A}_0^c}}{\varphi_{\mathcal{A}_0^c}}.$$

(i.b) Similarly by solving (9) for $\|\widehat{u}_{\mathcal{A}_0^c}\|_1$ when $\|\widehat{u}_{\mathcal{A}_0}\|_1$ is fixed, we obtain

$$b^2 - 4c \geqslant 0, \quad -b + (b^2 - 4c)^{1/2} \geqslant 0,$$

where $b = -\frac{\lambda\omega_{\mathcal{A}_0^c}}{\varphi_{\mathcal{A}_0^c}}$ and $c = \frac{1}{\varphi_{\mathcal{A}_0^c}}(\varphi_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1^2 + \lambda\omega_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1)$. Thus, we obtain

$$\|\widehat{u}_{\mathcal{A}_0}\|_1 \leqslant \lambda\frac{-\omega_{\mathcal{A}_0} + (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}$$
$$\|\widehat{u}_{\mathcal{A}_0}\|_1 \geqslant \max\left\{\lambda\frac{-\omega_{\mathcal{A}_0} - (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}, 0\right\} = 0.$$

Thus, $\|\widehat{u}_{\mathcal{A}_0}\|_1$ satisfies

$$0 \leqslant \|\widehat{u}_{\mathcal{A}_0}\|_1 \leqslant \lambda\frac{-\omega_{\mathcal{A}_0} + (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}.$$

(ii) $\omega_{\mathcal{A}_0} < 0$.

(ii.a) We repeat the same procedure to solve (9) for $\|\widehat{u}_{\mathcal{A}_0}\|_1$ when $\|\widehat{u}_{\mathcal{A}_0^c}\|_1$ is fixed as we did in part (i.1). Let $a = \varphi_{\mathcal{A}_0}$, $b = \lambda\omega_{\mathcal{A}_0}$ and $c = \varphi_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1^2 - \lambda\omega_{\mathcal{A}_0^c}\|\widehat{u}_{\mathcal{A}_0^c}\|_1$. Thus, we obtain two inequalities

$$b^2 - 4ac > 0, \quad -b + (b^2 - 4ac)^{1/2} > 0.$$

Solving the two inequalities, we have

$$\|\widehat{u}_{\mathcal{A}_0^c}\|_1 \leqslant \lambda\frac{\omega_{\mathcal{A}_0^c} + (\omega_{\mathcal{A}_0^c}^2 + \varphi_{\mathcal{A}_0^c}\varphi_{\mathcal{A}_0}^{-1}\omega_{\mathcal{A}_0}^2)^{1/2}}{2\varphi_{\mathcal{A}_0^c}},$$
$$\|\widehat{u}_{\mathcal{A}_0^c}\|_1 \geqslant \max\left\{\lambda\frac{\omega_{\mathcal{A}_0^c} - (\omega_{\mathcal{A}_0^c}^2 + \varphi_{\mathcal{A}_0^c}\varphi_{\mathcal{A}_0}^{-1}\omega_{\mathcal{A}_0}^2)^{1/2}}{2\varphi_{\mathcal{A}_0^c}}, 0\right\} = 0,$$

and thus

$$0 \leqslant \|\widehat{u}_{\mathcal{A}_0^c}\|_1 \leqslant \lambda\frac{\omega_{\mathcal{A}_0^c} + (\omega_{\mathcal{A}_0^c}^2 + \varphi_{\mathcal{A}_0^c}\varphi_{\mathcal{A}_0}^{-1}\omega_{\mathcal{A}_0}^2)^{1/2}}{2\varphi_{\mathcal{A}_0^c}}.$$

(ii.b) Similarly we fix $\|\widehat{u}_{\mathcal{A}_0}\|_1$ and obtain two inequalities

$$b^2 - 4ac > 0, \quad -b + (b^2 - 4ac)^{1/2} > 0,$$

where $a = \varphi_{\mathcal{A}_0^c}$, $b = -\lambda\omega_{\mathcal{A}_0^c}$ and $c = \varphi_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1^2 + \lambda\omega_{\mathcal{A}_0}\|\widehat{u}_{\mathcal{A}_0}\|_1$. We solve the two inequalities and obtain

$$\|\widehat{u}_{\mathcal{A}_0}\|_1 \leqslant \lambda \frac{-\omega_{\mathcal{A}_0} + (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}},$$
$$\|\widehat{u}_{\mathcal{A}_0}\|_1 \geqslant \max\left\{\lambda \frac{-\omega_{\mathcal{A}_0} - (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}, 0\right\} = 0,$$

which implies

$$0 \leqslant \|\widehat{u}_{\mathcal{A}_0}\|_1 \leqslant \lambda \frac{-\omega_{\mathcal{A}_0} + (\omega_{\mathcal{A}_0}^2 + \varphi_{\mathcal{A}_0}\varphi_{\mathcal{A}_0^c}^{-1}\omega_{\mathcal{A}_0^c}^2)^{1/2}}{2\varphi_{\mathcal{A}_0}}.$$

Thus we complete the proof of Theorem A.

4. Simulation study II

Here we include the detail of the second simulation experiment for longitudinal binary outcomes. Refer to Section 5 for the aims of the simulation study. To examine our method's performance in a different setting, we design the another simulation study which generates 6 datasets with binary longitudinal outcomes from the following logistic models:

$$\text{logit}\{E(Y_{k,ij} \mid X_{k,i}, Z_{k,ij})\} = \beta_{k,0}^0 + \beta_{k,1}^0 X_{k,i} + \beta_{k,2}^0 Z_{k,ij}$$
$$j = 1, \ldots, 4, k = 1, \ldots, 6, i = 1, \ldots, n_k,$$

where a baseline covariate $X_{k,i}$ is generated from Bernoulli(0.2) and a time-dependent covariate $Z_{k,i} = (Z_{k,i1}, \ldots, Z_{k,i4})^T$ is simulated from 4-variate normal $N(\mathbf{0}, 0.5\mathbf{I}_4)$. We set $\mathbf{R}_k(\cdot)$ for $k = 1, 4, 6$ as AR-1 and for $k = 2, 3, 5$ as CS, respectively, with equal correlation 0.5. The two cases of parameter homogeneity are considered, where Case II is set to have more distinct parameter groups than Case I. The vector of intercepts is set the same for both cases with all elements being -1; in Case I, the vectors of slope parameters for covariates X and Z are respectively set as $\boldsymbol{\beta}_1^0 = (-2, -2, -2, -2, -2, -2)^T$ and $\boldsymbol{\beta}_2^0 = (3, 3, 3, 4, 4, 4)^T$, while in Case

II, $\boldsymbol{\beta}_1^0 = (-2,-2,-1,2,2,2)^T$ and $\boldsymbol{\beta}_2^0 = (3,3,4,4,4,4)^T$. Clearly, the numbers of distinctive parameters in $\boldsymbol{\beta}_{(1)}^0$ and $\boldsymbol{\beta}_{(2)}^0$ are 4 for Case I and 5 for Case II, respectively.

Given the above configuration of slope parameters and correlation structures, the multivariate binary outcome $\mathbf{Y}_{k,i}$ is simulated by an algorithm proposed by Oman (2009). Matrices $\mathbf{D}$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{D}}_e$ are formed in the same way as those given in the first simulation study. The dimensions of resulting penalty matrices $\mathbf{D}$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{D}}_e$ are 30 by 18, 10 by 18 and 10 by 18, respectively.

Based on 200 rounds of simulation, we summarize results in Table 1, from which we can draw similar conclusions to those given in the first simulation study. Again, we see all three methods perform better in Case I than in Case II as Case II has more distinct parameter groups. Using the criteria of sensitivity and specificity, we see that $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$ remains the best performer, and that the overall performance of $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ again outperforms $\widehat{\boldsymbol{\beta}}_D$. The model sizes obtained by the three methods all stay close to the true model sizes.

[Table 1 about here.]

References

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

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics* **22,** 1993–2010.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50,** 1029–1054.

Kim, S.-J., Koh, K., Boyd, S. P., and Gorinevsky, D. M. (2009). l_1 trend filtering. *SIAM Review* **51,** 339–360.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2009). A unified framework for highdimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems 22*, pages 1348–1356.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, 2nd edition.

Oman, S. D. (2009). Easily simulated multivariate binary distributions with given positive and negative correlations. *Computational Statistics & Data Analysis* **53,** 999–1005.

Peng, H. and Fan, J. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32,** 928–961.

Table 1

Sensitivity (se100, se90), specificity (sp100, sp90), model size (size) and standard deviation of model size for the Case I and Case II in the simulation study II using different penalty matrices. Se100 and se90 represent the sensitivities computed based on 100% and 90% correct identification of all equal parameter pairs, respectively. Sp100 and sp90 are defined in the similar way but for unequal parameter pairs.

Case		AR-1			CS		
Penalty	n_k	Se100(Se90)	Sp100(Sp90)	Size(Std)	Se100(Se90)	Sp100(Sp90)	Size(Std)
I, $\widetilde{\mathbf{D}}$	100	0.300(0.400)	1.000(1.000)	5.585(1.978)	0.225(0.305)	0.990(0.990)	5.885(2.050)
	200	0.440(0.545)	1.000(1.000)	4.805(1.064)	0.350(0.470)	1.000(1.000)	5.195(1.533)
	400	0.545(0.630)	1.000(1.000)	4.600(0.737)	0.545(0.650)	1.000(1.000)	4.630(0.910)
I, $\widetilde{\mathbf{D}}_e$	100	0.255(0.385)	0.995(1.000)	5.380(1.999)	0.185(0.290)	1.000(1.000)	5.620(1.906)
	200	0.440(0.550)	1.000(1.000)	4.745(1.075)	0.350(0.460)	1.000(1.000)	5.185(1.491)
	400	0.530(0.620)	1.000(1.000)	4.580(0.711)	0.490(0.600)	1.000(1.000)	4.655(0.754)
I, $\mathbf{D}$	100	0.260(0.480)	0.985(0.985)	5.120(1.904)	0.250(0.400)	0.995(1.000)	5.310(1.950)
	200	0.430(0.650)	1.000(1.000)	4.515(1.121)	0.345(0.555)	1.000(1.000)	4.910(1.443)
	400	0.495(0.740)	1.000(1.000)	4.675(0.838)	0.490(0.740)	1.000(1.000)	4.660(0.817)
II, $\widetilde{\mathbf{D}}$	100	0.290(0.505)	0.610(0.640)	5.935(1.598)	0.200(0.375)	0.650(0.685)	6.470(1.793)
	200	0.495(0.650)	0.790(0.790)	5.535(1.102)	0.405(0.590)	0.775(0.780)	5.805(1.355)
	400	0.600(0.725)	0.960(0.970)	5.515(0.757)	0.540(0.705)	0.965(0.970)	5.580(0.853)
II, $\tilde{\mathbf{D}}_e$	100	0.190(0.370)	0.455(0.535)	5.900(1.742)	0.130(0.300)	0.520(0.610)	6.285(1.800)
	200	0.355(0.585)	0.645(0.645)	5.525(1.129)	0.270(0.490)	0.645(0.660)	5.795(1.265)
	400	0.460(0.640)	0.890(0.925)	5.665(0.828)	0.445(0.635)	0.885(0.920)	5.700(0.857)
II, $\mathbf{D}$	100	0.160(0.410)	0.390(0.465)	5.630(1.693)	0.110(0.310)	0.445(0.540)	6.070(1.664)
	200	0.275(0.615)	0.615(0.615)	5.555(1.306)	0.230(0.550)	0.625(0.635)	5.715(1.372)
	400	0.340(0.630)	0.850(0.885)	5.945(1.085)	0.360(0.650)	0.860(0.890)	5.955(1.131)