Biometrics 000, 000–000 DOI: 000
000 0000

Supporting Information for "Histopathological Imaging-based Cancer Heterogeneity Analysis via Penalized Fusion with Model Averaging" by Baihua He, Tingyan Zhong, Jian Huang, Yanyan Liu, Qingzhao Zhang, and Shuangge Ma

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

1. Additional details on computing Step 2

In the ℓth iteration, first consider the estimation of $\boldsymbol{\eta}_{im(b)}$, while treating the other parameters as fixed. The relevant minimization problem can be rewritten as:

$$\frac{\kappa}{2}\|\boldsymbol{\delta}_{im(b)}^{(\ell)} - \boldsymbol{\eta}_{im(b)}\|^2 + p_2(\|\boldsymbol{\eta}_{im(b)}\|, \lambda_2).$$

The closed-form solution under SCAD with $\gamma > 2$ is:

$$\boldsymbol{\eta}_{im(b)}^{(\ell)} = \begin{cases} S(\boldsymbol{\delta}_{im(b)}^{(\ell)}, \lambda_2/\kappa) & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \leqslant \lambda_2 + \lambda_2/\kappa \\ \frac{S\left(\boldsymbol{\delta}_{im(b)}^{(\ell)}, \gamma\lambda_2/((\gamma-1)\kappa)\right)}{1-1/((\gamma-1)\kappa)} & \text{if } \lambda_2 + \lambda_2/\kappa < \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \leqslant \gamma\lambda_2 \\ \boldsymbol{\delta}_{im(b)}^{(\ell)} & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| > \gamma\lambda_2. \end{cases}$$

We then consider $\boldsymbol{\theta}_{(b)}$, and the minimization has a quadratic form:

$$\boldsymbol{\theta}_{(b)} = \arg\min \left\{ \frac{1}{2}\left\|\boldsymbol{Y} - \mathbf{X}_{(b)}\boldsymbol{\theta}_{(b)}\right\|^2 + \frac{\kappa}{2}\left[\left\|\mathbf{A}\boldsymbol{\theta}_{(b)} - \boldsymbol{\eta}_{(b)}^{(\ell)} + \frac{\mathbf{v}_{(b)}^{(\ell-1)}}{\kappa}\right\|^2 + \left\|\boldsymbol{\theta}_{(b)} - \boldsymbol{\mu}_{(b)}^{(\ell-1)} + \frac{\bar{\mathbf{v}}_{(b)}^{(\ell-1)}}{\kappa}\right\|^2\right] \right\},$$

which leads to the closed-form solution in (9).

For the update of $\boldsymbol{\mu}_{(b)}$, we have the objective function:

$$\frac{\kappa}{2}\|\boldsymbol{\xi}_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{|A_b|} p_1(|\mu_{i(b)}^j|, \lambda_1),$$

which has a closed-form solution as in (10).

2. Proof of Corollary 1

From (11),

$$\begin{aligned} \bar{\mathbf{v}}_{(b)}^{(\ell)} &= \bar{\mathbf{v}}_{(b)}^{(\ell-1)} + \kappa(\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}^{(\ell)}) \\ &= \kappa(\boldsymbol{\xi}_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}^{(\ell)}). \end{aligned}$$

With $\gamma > 2$,

$$\left|\bar{v}_{i(b)}^{j(\ell)}\right| = \begin{cases} \kappa\left|\xi_{i(b)}^{j(\ell)}\right| & \text{if } |\xi_{i(b)}^{j(\ell)}| \leqslant \lambda_1/\kappa \\ \lambda_1 & \text{if } \lambda_1/\kappa < |\xi_{i(b)}^{j(\ell)}| \leqslant \lambda_1 + \lambda_1/\kappa \\ \frac{\kappa\gamma\lambda_1 - \kappa\left|\xi_{i(b)}^{j(\ell)}\right|}{(\gamma-1)\kappa-1} & \text{if } \lambda_1 + \lambda_1/\kappa < |\xi_{i(b)}^{j(\ell)}| \leqslant \gamma\lambda_1 \\ 0 & \text{if } |\xi_{i(b)}^{j(\ell)}| > \gamma\lambda_1. \end{cases}$$

So $|\bar{v}_{i(b)}^{j(\ell)}| \leqslant \lambda_1$. As for $\mathbf{v}_{im(b)}^{(\ell)}$,

$$\mathbf{v}_{im(b)}^{(\ell)} = \kappa \boldsymbol{\delta}_{im(b)}^{(\ell)} - \kappa \boldsymbol{\eta}_{im(b)}^{(\ell)},$$

$$\mathbf{v}_{im(b)}^{(\ell)} = \begin{cases} \kappa \boldsymbol{\delta}_{im(b)}^{(\ell)} & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \leqslant \lambda_2/\kappa \\ \frac{\lambda_2}{\|\boldsymbol{\delta}_{im(b)}^{(\ell)}\|} \boldsymbol{\delta}_{im(b)}^{(\ell)} & \text{if } \lambda_2/\kappa < \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \leqslant \lambda_2 + \lambda_2/\kappa \\ \frac{\left(\frac{\gamma\kappa\lambda_2}{\|\boldsymbol{\delta}_{im(b)}^{(\ell)}\|} - \kappa\right)\boldsymbol{\delta}_{im(b)}^{(\ell)}}{(\gamma-1)\kappa-1} & \text{if } \lambda_2 + \lambda_2/\kappa < \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| \leqslant \gamma\lambda_2 \\ \mathbf{0} & \text{if } \|\boldsymbol{\delta}_{im(b)}^{(\ell)}\| > \gamma\lambda_2. \end{cases}$$

So $\|\mathbf{v}_{im(b)}^{(\ell)}\| \leqslant \lambda_2$. With the update (11), we can conclude that:

$$\begin{aligned} & T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}) - T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) \\ = \; & \frac{1}{\kappa}\|\mathbf{v}_{(b)}^{(\ell)} - \mathbf{v}_{(b)}^{(\ell-1)}\| + \frac{1}{\kappa}\|\bar{\mathbf{v}}_{(b)}^{(\ell)} - \bar{\mathbf{v}}_{(b)}^{(\ell-1)}\|. \end{aligned} \tag{A.1}$$

With the definitions of $\boldsymbol{\mu}_{(b)}^{(\ell)}$ and $\boldsymbol{\eta}_{(b)}^{(\ell)}$,

$$T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) - T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) \leqslant 0. \tag{A.2}$$

Moreover, since the Hessian matrix $\mathbf{X}_{(b)}^{\mathrm{T}}\mathbf{X}_{(b)} + \kappa\mathbf{A}^{\mathrm{T}}\mathbf{A} + \kappa\mathbf{I}_{n|A_b|}$ is positive definite, and since $T_n(\boldsymbol{\theta}_{(b)}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)})$ is strongly convex with respect to $\boldsymbol{\theta}_{(b)}$, there exists a constant $c > 0$ such that:

$$\begin{aligned} & T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) - T_n(\boldsymbol{\theta}_{(b)}^{(\ell-1)}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) \\ \leqslant \; & -\frac{c}{2}\|\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\theta}_{(b)}^{(\ell-1)}\|. \end{aligned} \tag{A.3}$$

Combining (A.1), (A.2), and (A.3), we have:

$$\begin{aligned} & T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}) - T_n(\boldsymbol{\theta}_{(b)}^{(\ell-1)}, \boldsymbol{\mu}_{(b)}^{(\ell-1)}, \boldsymbol{\eta}_{(b)}^{(\ell-1)}, \mathbf{v}_{(b)}^{(\ell-1)}, \bar{\mathbf{v}}_{(b)}^{(\ell-1)}) \\ \leqslant & -\frac{c}{2}\|\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\theta}_{(b)}^{(\ell-1)}\| + \frac{1}{\kappa}\|\mathbf{v}_{(b)}^{(\ell)} - \mathbf{v}_{(b)}^{(\ell-1)}\| + \frac{1}{\kappa}\|\bar{\mathbf{v}}_{(b)}^{(\ell)} - \bar{\mathbf{v}}_{(b)}^{(\ell-1)}\|. \end{aligned} \tag{A.4}$$

Since $\left\{\boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}\right\}_{\ell=1}^{\infty}$ are bounded, together with that $\left\{\mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\right\}_{\ell=1}^{\infty}$ are bounded, $\boldsymbol{\theta}_{(b)}^{(\ell)}$ is also bounded. So $T^{(\ell)}$ and $\left\{\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}\mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\right\}_{\ell=1}^{\infty}$ are bounded, where $T^{(\ell)} = T_n(\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}, \mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)})$. Further denote $y^{(\ell)} = \frac{c}{2}\|\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\theta}_{(b)}^{(\ell-1)}\|$ and $z^{(\ell)} = \frac{1}{\kappa}\|\mathbf{v}_{(b)}^{(\ell)} - \mathbf{v}_{(b)}^{(\ell-1)}\| + \frac{1}{\kappa}\|\bar{\mathbf{v}}_{(b)}^{(\ell)} - \bar{\mathbf{v}}_{(b)}^{(\ell-1)}\|$.

Since $T^{(\ell)}$ is bounded, there exists a subsequence $\{T^{(\ell_j)}\}$ such that:

$$\lim_{\ell_j\to\infty} T^{(\ell_j)} = \liminf_{\ell\to\infty} T^{(\ell)}. \tag{A.5}$$

By (A.4) and $\lim_{\ell\to\infty} z^{(\ell)} \to 0$, we have:

$$\begin{aligned}\liminf_{\ell_j\to\infty} y^{(\ell_j)} &\leqslant \liminf_{\ell_j\to\infty} (T^{(\ell_j-1)} - T^{(\ell_j)} + z^{(\ell_j)}) \\ &= \liminf_{\ell\to\infty} T^{(\ell)} - \liminf_{\ell_j\to\infty} T^{(\ell_j+1)} \leqslant 0.\end{aligned}$$

As $y^{(\ell_j)} \geqslant 0$, $\liminf_{\ell_j\to\infty} y^{(\ell_j)} = 0$, which means that $\|\boldsymbol{\theta}_{(b)}^{(\ell_j)} - \boldsymbol{\theta}_{(b)}^{(\ell_j-1)}\| = 0$. Together with $\liminf_{\ell_j\to\infty} z^{(\ell_j)} = 0$, we have:

$$\liminf_{\ell_j\to\infty} \|\boldsymbol{\mu}_{(b)}^{(\ell_j)} - \boldsymbol{\mu}_{(b)}^{(\ell_j-1)}\| + \|\bar{\boldsymbol{\eta}}_{(b)}^{(\ell_j)} - \bar{\boldsymbol{\eta}}_{(b)}^{(\ell_j-1)}\| = 0.$$

So the sequence $\left\{\boldsymbol{\theta}_{(b)}^{(\ell)}, \boldsymbol{\mu}_{(b)}^{(\ell)}, \boldsymbol{\eta}_{(b)}^{(\ell)}\mathbf{v}_{(b)}^{(\ell)}, \bar{\mathbf{v}}_{(b)}^{(\ell)}\right\}_{\ell=1}^{\infty}$ has a subsequence $\left\{\boldsymbol{\theta}_{(b)}^{(\ell_j)}, \boldsymbol{\mu}_{(b)}^{(\ell_j)}, \boldsymbol{\eta}_{(b)}^{(\ell_j)}\mathbf{v}_{(b)}^{(\ell_j)}, \bar{\mathbf{v}}_{(b)}^{(\ell_j)}\right\}_{\ell_j=1}^{\infty}$, which converges to a point $\left\{\boldsymbol{\theta}_{(b)}^{\#}, \boldsymbol{\mu}_{(b)}^{\#}, \boldsymbol{\eta}_{(b)}^{\#}\mathbf{v}_{(b)}^{\#}, \bar{\mathbf{v}}_{(b)}^{\#}\right\}$. Combining with (11), we have:

$$\begin{cases} \mathbf{A}\boldsymbol{\theta}_{(b)}^{\#} - \boldsymbol{\eta}_{(b)}^{\#} = \mathbf{0} \\ \boldsymbol{\theta}_{(b)}^{\#} - \boldsymbol{\mu}_{(b)}^{\#} = \mathbf{0}. \end{cases} \tag{A.6}$$

Further, the procedure satisfies the following optimality conditions:

$$\begin{cases} \mathbf{X}_{(b)}^{\mathrm{T}}(-\mathbf{y} + \mathbf{X}_{(b)}\boldsymbol{\theta}_{(b)}^{(\ell)}) + \kappa\mathbf{A}^{\mathrm{T}}\{\mathbf{A}\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\eta}_{(b)}^{(\ell)} + \kappa^{-1}\mathbf{v}_{(b)}^{(\ell-1)}\} + \kappa\{\boldsymbol{\theta}_{(b)}^{(\ell)} - \boldsymbol{\mu}_{(b)}^{(\ell-1)} + \kappa^{-1}\bar{\mathbf{v}}_{(b)}^{(\ell-1)}\} = \mathbf{0} \\ 0 \in -\kappa\{\theta_{i(b)}^{j(\ell)} - \mu_{i(b)}^{j(\ell)} + \kappa^{-1}\bar{v}_{i(b)}^{j(\ell-1)}\} + \frac{\partial p_1(|\mu_{i(b)}^{j}|,\lambda_1)}{\partial \mu_{i(b)}^{j}}\big|_{\mu_{i(b)}^{j}=\mu_{i(b)}^{j(\ell)}} \\ \mathbf{0} \in -\kappa\{\mathbf{A}\boldsymbol{\theta}_{(b)}^{(\ell-1)} - \boldsymbol{\eta}_{(b)}^{(\ell)} + \kappa^{-1}\mathbf{v}_{(b)}^{(\ell-1)}\} + \frac{\partial p_2(\|\boldsymbol{\eta}_{im(b)}\|,\lambda_2)}{\partial \boldsymbol{\eta}_{im(b)}}\big|_{\boldsymbol{\eta}_{im(b)}=\boldsymbol{\eta}_{im(b)}^{(\ell)}}. \end{cases} \tag{A.7}$$

Hence,

$$\begin{cases} \mathbf{X}_{(b)}^{\mathrm{T}}(-\mathbf{y} + \mathbf{X}_{(b)}\boldsymbol{\theta}_{(b)}^{\#}) + \mathbf{A}^{\mathrm{T}}\mathbf{v}_{(b)}^{\#} + \bar{\mathbf{v}}_{(b)}^{\#} = \mathbf{0} \\ 0 \in -\bar{v}_{i(b)}^{j\#} + \frac{\partial p_1(|\mu_{i(b)}^{j}|,\lambda_1)}{\partial \mu_{i(b)}^{j}}\big|_{\mu_{i(b)}^{j}=\mu_{i(b)}^{j\#}} \\ \mathbf{0} \in -\mathbf{v}_{im(b)}^{\#} + \frac{\partial p_2(\|\boldsymbol{\eta}_{im(b)}\|,\lambda_2)}{\partial \boldsymbol{\eta}_{im(b)}}\big|_{\boldsymbol{\eta}_{im(b)}=\boldsymbol{\eta}_{im(b)}^{\#}}. \end{cases} \tag{A.8}$$

Therefore, $\left\{\boldsymbol{\theta}_{(b)}^{\#}, \boldsymbol{\mu}_{(b)}^{\#}, \boldsymbol{\eta}_{(b)}^{\#}\mathbf{v}_{(b)}^{\#}, \bar{\mathbf{v}}_{(b)}^{\#}\right\}$ is a KKT point of (5).

3. Conditions

To establish the theoretical results described in the main text, we first assume the following conditions.

C1 $C_{nk(b)} = \frac{1}{|\mathcal{G}_k|}\sum_{i\in\mathcal{G}_k}\mathbf{x}_{i(b)}\mathbf{x}_{i(b)}^{\mathrm{T}} \to C_{k(b)}$, for $k = 1,\ldots,K$ and $b = 1,\ldots,B_n$, where $C_{k(b)}$ is a nonnegative definite matrix and $\sup_i \sup_b \|\mathbf{x}_{i(b)}\| \leqslant C\sqrt{|A_1|\log(nB_n)}$.

C2 (i) $p_1^{'}(t)$, the derivative of $p_1(t)$, exists and is bounded. $\liminf_{t\to 0+} p_1^{'}(t) > 0$. There are constants c and C such that, when $t_1, t_2 > c\lambda_1$, $\lambda_1|p_1^{''}(t_1) - p_1^{''}(t_2)| \leqslant C|t_1 - t_2|$, where $p_1^{''}(\cdot)$ is the second order derivative of $p_1(t)$. (ii) $p_2(t)$ is a symmetric, non-decreasing, and concave on $[0,\infty)$. It is constant for $t > a\lambda_2$ for some constant $a > 0$, and $p_2(0) = 0$. $p_2^{'}(t)$ exists and is continuous except for a finite number of values of t and $p_2^{'}(0+) = 1$.

C3 $\boldsymbol{\epsilon}_{(b)} = (\epsilon_{1(b)},\ldots,\epsilon_{n(b)})^{\mathrm{T}}$ is sub-Gaussian, that is, $P(|\mathbf{a}^{\mathrm{T}}\boldsymbol{\epsilon}_{(b)}| > \|\mathbf{a}\|x) \leqslant 2\exp(-cx^2)$ for any vector $\mathbf{a}\in\mathbb{R}^n$ and $x > 0$, where $0 < c < \infty$.

C4 For each k and b, $\mathcal{A}_{k,b}$, the parameter space of $\boldsymbol{\alpha}_{k(b)}^*$, is compact, and $\bigcup_{k=1}^{K}\bigcup_{b=1}^{B_n}\mathcal{A}_{k,b}$ is bounded, as $n\to\infty$.

C5 Consider $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_B)^{\mathrm{T}} \in \boldsymbol{\Omega}_n = \{\boldsymbol{\omega} : 0 \leqslant \omega_b \leqslant 1 \text{ and } \sum_{b=1}^{B}\omega_b = 1\}$. If submodel b^* is fitted and $\sum_{b\in\mathcal{M}}\omega_b \neq 0$, then $\sum_{i=1}^{n}\left(\sum_b \omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^* - \mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^*\right)^2 > 0$.

C6 (i) $\lambda_1|\mathcal{G}_{\max}|^{1/2} = O(1)$; (ii) $b_n^{-1}|\mathcal{G}_{\min}|^{-1/2}|A_1|^{1/2} = o(1)$, and (iii) $|\mathcal{G}_{\min}|^{-1/2}|A_1|^{3/2}\log(nB_n) = o(1)$.

Condition C1 bounds $C_{nk(b)}$ and $\|\mathbf{x}_{i(b)}\|$. SCAD and several other penalties satisfy Condition C2. The sub-Gaussian assumption in Condition C3 is common in high-dimensional regression. Conditions C1, C2(i), C3, and C6(i) ensure consistency of the oracle estimators defined below. Denote $\mathcal{B} = \left(\mathbf{X}_{(1)}\mathbf{G}_1\boldsymbol{\alpha}_{(1)}^*,\ldots,\mathbf{X}_{(B_n)}\mathbf{G}_{B_n}\boldsymbol{\alpha}_{(B_n)}^*\right)$. If $\mathrm{rank}(\mathcal{B}) = B_n - s_n + 1$, Condition C5 is satisfied, where s_n is the number of fitted candidate submodels (in the whole candidate submodel set). Further, Condition C6 also gives the divergence rate that the sample size, candidate submodel size, and maximum covariates' dimension of each candidate submodel

need to satisfy. In particular, if we consider a fixed K, since $|\mathcal{G}_{\min}| \leqslant n/K$, Condition C6 is satisfied by choosing $B_n = O(n)$ and $|A_1| = o(n^{1/3}(\log n)^{-2/3})$. As such, $|A_1|$ and B_n are allowed to diverge with n. Or if we consider a fixed $|A_1|$, Condition C6 is satisfied by choosing $K^{1/2}\log(nB_n) = o(n^{1/2})$. It is also noted that, as in a large number of high-dimensional regression studies, we consider non-random covariates. Otherwise, additional conditions on covariate distributions and significantly different theoretical developments would be needed.

4. Proof of Theorem 1

Some published studies such as Ma and Huang (2017) and a few others consider the simpler scenario with a low dimensionality and no sparsity. As such, when our analysis can be reduced to such a scenario, some of the existing theoretical developments can be borrowed. In what follows, we provide references to Ma and Huang (2017) and others when taking advantage of the existing results. However, our theoretical development is considerably more challenging and demands new investigation with the higher dimensionality (and hence the demand for, for example, analyzing the relationship between sample size and number of covariates), presence of sparsity (and hence the additional sparsity penalty), and model averaging (and hence the demand for analyzing each submodel separately and then assembling multiple submodels).

Under Conditions C1 and C2(i), as $\lambda_1 = O(|\mathcal{G}_k|^{-1/2})$, it can be proved following Fan and Peng (2004) that there exists a local minimizer $\widehat{\boldsymbol{\alpha}}^{\text{or}}_{k(b)}$ of the objective function (13) satisfying

$$\|\widehat{\boldsymbol{\alpha}}^{\text{or}}_{k(b)} - \boldsymbol{\alpha}^{*}_{k(b)}\| = O_p(|\mathcal{G}_k|^{-1/2}|A_b|^{1/2}). \tag{A.9}$$

It should be noted that the "true" subgrouping structure may vary across candidates submodels, depending on the covariates in each submodel. So the "true" subgroup size of each submodel K_b may also differ. Similar to in Ma and Huang (2017), define

$$F_n(\boldsymbol{\theta}_{(b)}, \lambda_1) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}_{(b)}\boldsymbol{\theta}_{(b)}\|^2 + \lambda_1 \sum_{i=1}^{n}\sum_{j=1}^{|A_b|} p_1(|\theta^j_{i(b)}|),\ P_n(\boldsymbol{\theta}_{(b)}, \lambda_2) = \lambda_2 \sum_{i<m} p_2(\|\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}_{m(b)}\|).$$

Then $Q_n(\boldsymbol{\theta}_{(b)}, \lambda_1, \lambda_2) = F_n(\boldsymbol{\theta}_{(b)}, \lambda_1) + P_n(\boldsymbol{\theta}_{(b)}, \lambda_2)$. Define $\mathcal{T} : \mathcal{C}^b_{\mathcal{G}} \to R^{K_b|A_b|}$ as a mapping,

under which $\mathcal{T}(\boldsymbol{\theta}_{(b)})$ is the $K_b|A_b|$-vector consisting of K_b vectors with dimension $|A_b|$, and its kth vector component equals the common value of $\boldsymbol{\theta}_{i(b)}$ for $i \in \mathcal{G}_k$, and $\mathcal{T}^\star : R^{n|A_b|} \to R^{K_b|A_b|}$ as another mapping, under which $\mathcal{T}^\star(\boldsymbol{\theta}_{(b)}) = \{|\mathcal{G}_k|^{-1} \sum_{i\in\mathcal{G}_k} \boldsymbol{\theta}_{i(b)}^{\mathrm{T}}, k = 1, \ldots, K_b\}^{\mathrm{T}}$.

Consider the neighborhood of $\boldsymbol{\theta}^*_{(b)}$:

$$\Theta = \{\boldsymbol{\theta}_{(b)} \in R^{n|A_b|} : \sup_i \|\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| \leqslant \phi_n\},$$

where $\phi_n = C\sqrt{\frac{|A_b|}{|\mathcal{G}_{\min}|}}$. By (A.9), for any $\epsilon > 0$, there exists an event E_1 under which $\sup_i \|\widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| \leqslant \phi_n$ and $P(E_1^c) \leqslant \epsilon$. Hence, $\widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{(b)} \in \Theta$ under E_1. For any $\boldsymbol{\theta}_{(b)} \in R^{n|A_b|}$, let $\boldsymbol{\theta}^\star_{(b)} = \left\{\left(\theta^{1\star}_{i(b)}, \ldots, \theta^{|A_b|\star}_{i(b)}\right), i = 1, \ldots, n\right\}^{\mathrm{T}} = \mathcal{T}^{-1}(\mathcal{T}^\star(\boldsymbol{\theta}_{(b)}))$.

We will show that $\widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{(b)}$ is a strictly local minimizer of objective function (4) with probability approaching 1 through the following steps: (i) Under event E_1, $Q_n(\boldsymbol{\theta}^\star_{(b)}, \lambda_1, \lambda_2) > Q_n(\widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{(b)}, \lambda_1, \lambda_2)$ for any $\boldsymbol{\theta}_{(b)} \in \Theta$ and $\boldsymbol{\theta}^\star_{(b)} \neq \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{(b)}$. (ii) There is an event E_2 such that $P(E_2^c) \leqslant \epsilon$. Under $E_1 \cap E_2$, there is a neighborhood of $\widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{(b)}$, denoted as Θ_n, such that $Q_n(\boldsymbol{\theta}_{(b)}, \lambda_1, \lambda_2) > Q_n(\boldsymbol{\theta}^\star_{(b)}, \lambda_1, \lambda_2)$ for any $\boldsymbol{\theta}^\star_{(b)} \in \Theta_n \cap \Theta$ for sufficiently small ϵ.

With the assumptions $b_n > a\lambda_2$ and $b_n \gg \phi_n$ and Condition C2, we can establish (i) similarly as in Ma and Huang (2017). Next, we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\theta}_{i(b)} : \sup_i \|\boldsymbol{\theta}_{i(b)} - \widehat{\boldsymbol{\theta}}^{\mathrm{or}}_{i(b)}\| \leqslant t_n\}$. For $\boldsymbol{\theta}_{(b)} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

$$Q_n(\boldsymbol{\theta}_{(b)}, \lambda_1, \lambda_2) - Q_n(\boldsymbol{\theta}^\star_{(b)}, \lambda_1, \lambda_2) = \Gamma_1 + \Gamma_2 + \Gamma_3,$$

where $\Gamma_1 = -(\boldsymbol{Y} - \mathbf{X}_{(b)}\boldsymbol{\theta}^\circ_{(b)})^{\mathrm{T}}\mathbf{X}_{(b)}(\boldsymbol{\theta}_{(b)} - \boldsymbol{\theta}^\star_{(b)})$, $\Gamma_2 = \lambda_1 \sum_{i=1}^n \sum_{j=1}^{|A_b|}\{p_1'(|\theta^{j\circ}_{i(b)}|)\mathrm{sign}(\theta^{j\circ}_{i(b)})\}(\theta^j_{i(b)} - \theta^{j\star}_{i(b)})$, $\Gamma_3 = \sum_{i=1}^n \frac{\partial P_n(\boldsymbol{\theta}^\circ_{(b)})}{\partial \boldsymbol{\theta}^{\mathrm{T}}_{i(b)}}(\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}^\star_{i(b)})$, $\boldsymbol{\theta}^\circ_{(b)} = \left\{\left(\boldsymbol{\theta}^\circ_{1(b)}\right)^{\mathrm{T}}, \ldots, \left(\boldsymbol{\theta}^\circ_{n(b)}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$, $\boldsymbol{\theta}^\circ_{i(b)} = \left(\theta^{1\circ}_{n(b)}, \ldots, \theta^{|A_b|\circ}_{n(b)}\right)^{\mathrm{T}}$ and $\theta^{j\circ}_{i(b)} = \alpha\theta^j_{i(b)} + (1-\alpha)\theta^{j\star}_{i(b)}$ for some constant $\alpha \in (0, 1)$. Following similar steps as in Ma and Huang (2017), we conclude that:

$$\begin{aligned}\Gamma_1 \;\geqslant\; & \{-2C|\mathcal{G}_{\min}|^{-1}\sqrt{2c^{-1}|A_b|\log n \log(nB_n)} - 2C'|\mathcal{G}_{\min}|^{-3/2}\log(nB_n)|A_b|^{3/2}\} \\ & \times \sum_{k=1}^{K_b} \sum_{i,j\in\mathcal{G}_k, i<m} \|\boldsymbol{\theta}_{m(b)} - \boldsymbol{\theta}_{i(b)}\|, \end{aligned} \tag{A.10}$$

and

$$\Gamma_3 \geqslant \sum_{k=1}^{K_b} \sum_{i,j\in\mathcal{G}_k, i<m} \lambda_2 p_2'(4t_n)\|\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}_{m(b)}\|, \tag{A.11}$$

under Conditions C1–C3. Further,

$$\begin{aligned}
|\Gamma_2| &= \left|\lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i\in\mathcal{G}_k}\{p_1'(|\theta_{i(b)}^{l\circ}|)\mathrm{sign}(\theta_{i(b)}^{l\circ})\}(\theta_{i(b)}^{l} - \theta_{i(b)}^{l\star})\right| \\
&= \left|\lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i,m\in\mathcal{G}_k}\{p_1'(|\theta_{i(b)}^{l\circ}|)\mathrm{sign}(\theta_{i(b)}^{l\circ})\}\frac{(\theta_{i(b)}^{l} - \theta_{m(b)}^{l})}{|\mathcal{G}_k|}\right| \\
&= \left|\lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i,m\in\mathcal{G}_k}\{p_1'(|\theta_{i(b)}^{l\circ}|)\mathrm{sign}(\theta_{i(b)}^{l\circ})\}\frac{(\theta_{i(b)}^{l} - \theta_{m(b)}^{l})}{2|\mathcal{G}_k|}\right. \\
&\quad \left. + \lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i,m\in\mathcal{G}_k}\{p_1'(|\theta_{m(b)}^{l\circ}|)\mathrm{sign}(\theta_{m(b)}^{l\circ})\}\frac{(\theta_{m(b)}^{l} - \theta_{i(b)}^{l})}{2|\mathcal{G}_k|}\right| \\
&= \left|\lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i,m\in\mathcal{G}_k}\{p_1'(|\theta_{i(b)}^{l\circ}|)\mathrm{sign}(\theta_{i(b)}^{l\circ}) - p_1'(|\theta_{m(b)}^{l\circ}|)\mathrm{sign}(\theta_{m(b)}^{l\circ})\}\frac{(\theta_{i(b)}^{l} - \theta_{m(b)}^{l})}{2|\mathcal{G}_k|}\right| \\
&= \left|\lambda_1 \sum_{k=1}^{K_b}\sum_{l=1}^{|A_b|}\sum_{i,m\in\mathcal{G}_k, i<m}\{p_1'(|\theta_{i(b)}^{l\circ}|)\mathrm{sign}(\theta_{i(b)}^{l\circ}) - p_1'(|\theta_{m(b)}^{l\circ}|)\mathrm{sign}(\theta_{m(b)}^{l\circ})\}\frac{(\theta_{i(b)}^{l} - \theta_{m(b)}^{l})}{|\mathcal{G}_k|}\right| \\
&\leqslant 2\lambda_1\sqrt{|A_b|}\sup_i\sup_l|p_1'(|\theta_{i(b)}^{l\circ}|)|\sum_{k=1}^{K_b}\sum_{i,m\in\mathcal{G}_k,i<m}\frac{\|\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}_{m(b)}\|}{|\mathcal{G}_{\min}|} \\
&\leqslant \frac{2C\lambda_1\sqrt{|A_b|}}{|\mathcal{G}_{\min}|}\sum_{k=1}^{K_b}\sum_{i,m\in\mathcal{G}_k,i<m}\|\boldsymbol{\theta}_{i(b)} - \boldsymbol{\theta}_{m(b)}\|.
\end{aligned} \tag{A.12}$$

Therefore, with (A.10), (A.11), and (A.12), we have:

$$\begin{aligned}
& Q_n(\boldsymbol{\theta}_{(b)}, \lambda_1, \lambda_2) - Q_n(\boldsymbol{\theta}_{(b)}^{\star}, \lambda_1, \lambda_2) \\
= \;& \Gamma_1 + \Gamma_2 + \Gamma_3 \\
\geqslant \;& \Gamma_1 + \Gamma_3 - |\Gamma_2| \\
\geqslant \;& \{\lambda_2 p_2'(4t_n) - 2C|\mathcal{G}_{\min}|^{-1}\sqrt{2c^{-1}|A_b|\log n\log(nB_n)} - 2C'|\mathcal{G}_{\min}|^{-3/2}\log(nB_n)|A_b|^{3/2} \\
& -2C\lambda_1|\mathcal{G}_{\min}|^{-1}\sqrt{|A_b|}\}\sum_{k=1}^{K_b}\sum_{i,m\in\mathcal{G}_k,i<m}\|\boldsymbol{\theta}_{m(b)} - \boldsymbol{\theta}_{i(b)}\| \geqslant 0.
\end{aligned}$$

So we can conclude:

$$P\left(\widehat{\boldsymbol{\theta}}_{(b)} = \widehat{\boldsymbol{\theta}}_{(b)}^{\text{or}}\right) \to 1, \tag{A.13}$$

under Condition C6. (A.9) and (A.13) lead to $\sup_i \|\widehat{\boldsymbol{\theta}}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| = O_p(\sqrt{|A_b|/|\mathcal{G}_{\min}|})$.

5. Proof of Theorem 2

First consider the scenario where there is only one submodel b^* not belonging to $\mathcal{M}$. Note that $\epsilon_{i(b^*)} = y_i - \mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)}$, and $\epsilon_{i(b^*)}$ and $\mathbf{x}_{i(b^*)}$ are independent. Under Conditions C1 and C3–C4, for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}_n$:

$$\begin{aligned}
& P\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)}\right)\left(\mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)} - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)}\right) \geqslant C\sqrt{n^{-1}|A_1|\log n\log(nB_n)}\right) \\
\leqslant\ & P\Bigg(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)}\right)\left(\mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)} - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)}\right) \geqslant \\
& \qquad Cn^{-1}\sqrt{\log n\sum_{i=1}^{n}\left(\mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)} - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)}\right)^2}\Bigg) \\
\leqslant\ & 2\exp(-C\log n) = 2n^{-C} \to 0,
\end{aligned} \tag{A.14}$$

as $n \to \infty$, and C may vary across equations.

Under Condition C4, with the result of Theorem 1 that $\sup_i \|\widehat{\boldsymbol{\theta}}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| = O_p\left(\sqrt{\frac{|A_1|}{|\mathcal{G}_{\min}|}}\right)$ ($|A_1| = \ldots = |A_{B_n}|$) and $\sup_i \sup_b \|\mathbf{x}_{i(b)}\| \leqslant C\sqrt{|A_1|\log(nB_n)}$, we can conclude that:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)} - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\widehat{\boldsymbol{\theta}}_{i(b)}\right)^2 = O_p\left(\frac{|A_1|^2\log(nB_n)}{|\mathcal{G}_{\min}|}\right), \tag{A.15}$$

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)}\right)\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)} - \sum_{b=1}^{B_n}\omega_b\mathbf{x}^{\mathrm{T}}_{i(b)}\widehat{\boldsymbol{\theta}}_{i(b)}\right) \\
& \quad = O_p\left(\sqrt{\frac{|A_1|^2\log n\log(nB_n)}{|\mathcal{G}_{\min}|}}\right),
\end{aligned} \tag{A.16}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\right)\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\right)$$
$$=O_p\left(|A_1|\log(nB_n)\sqrt{\frac{|A_1|}{|\mathcal{G}_{\min}|}}\right), \tag{A.17}$$

for any $\boldsymbol{\omega}\in\boldsymbol{\Omega}_n$. Further, we rewrite:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\omega}) &= \sum_{i=1}^{n}\left(y_i-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\right)^2\\
&= \sum_{i=1}^{n}\Bigg(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}+\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\\
&\quad+\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\Bigg)^2\\
&= \sum_{i=1}^{n}\left(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}\right)^2+\sum_{i=1}^{n}\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\right)^2\\
&\quad+2\sum_{i=1}^{n}\left(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}\right)\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\right)\\
&\quad+\sum_{i=1}^{n}\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\right)^2\\
&\quad+2\sum_{i=1}^{n}\left(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}\right)\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\right)\\
&\quad+2\sum_{i=1}^{n}\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\right)\left(\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b)}\right).
\end{aligned}$$

On the other hand, we also have:

$$\begin{aligned}
\mathcal{L}(\mathbf{e}_{b^*}) &= \sum_{i=1}^{n}\left(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}\right)^2+\sum_{i=1}^{n}\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b^*)}\right)^2\\
&\quad+2\sum_{i=1}^{n}\left(y_i-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}\right)\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\mathbf{x}_{i(b^*)}^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{i(b^*)}\right).
\end{aligned}$$

Hence, for any $\boldsymbol{\omega}\in\boldsymbol{\Omega}_n$ with $\sum_{b\neq b^*}\omega_b>0$, under Conditions C5–C6, (A.14)–(A.17) lead to

$$n^{-1}\mathcal{L}(\boldsymbol{\omega})=n^{-1}\mathcal{L}(\mathbf{e}_{b^*})+n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_{i(b^*)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b^*)}^{*}-\sum_{b=1}^{B_n}\omega_b\mathbf{x}_{i(b)}^{\mathrm{T}}\boldsymbol{\theta}_{i(b)}^{*}\right)^2+o_p(1),$$

where $o_p(1)$ is uniform in $\boldsymbol{\omega} \in \Omega_n$. This means:

$$\lim_{n\to\infty} P\left(\mathcal{L}(\boldsymbol{\omega}) > \mathcal{L}(\mathbf{e}_{b^*})\right) \to 1.$$

Note the fact $\widehat{\boldsymbol{\omega}} = \inf_{\boldsymbol{\omega}\in\Omega_n} \mathcal{L}(\boldsymbol{\omega})$, which leads to:

$$\lim_{n\to\infty} P\left(\sum_{b\neq b^*} \widehat{\omega}_b = 0\right) \to 1.$$

Thus $\lim_{n\to\infty} P\left(\widehat{\omega}_{b^*} = 1\right) \to 1$.

Consider the more general scenario with multiple fitted candidate submodels. Note that if submodel $b^* \notin \mathcal{M}$, we have $\boldsymbol{\theta}^*_{i(b^*)} = \boldsymbol{\theta}^0_{i(b^*)}$. Further, for any submodel $b \notin \mathcal{M}$, $\mathbf{x}^{\mathrm{T}}_{i(b)}\boldsymbol{\theta}^*_{i(b)} = \mathbf{x}^{\mathrm{T}}_{i(b^*)}\boldsymbol{\theta}^*_{i(b^*)}$. So we can view all the fitted submodels as a new fitted submodel named as $b^\star$. Similarly, we can conclude that $\widehat{\boldsymbol{\omega}} = \mathbf{e}_{b^\star}$ in probability, which means that

$$\lim_{n\to\infty} P\left(\sum_{b\notin\mathcal{M}} \widehat{\omega}_b = 1\right) \to 1.$$

Thus, we complete the proof of Theorem 2.

5.1 *Remarks*

With Theorems 1 and 2, we have

$$\sup_i \|\widehat{\boldsymbol{\theta}}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| = O_p(\sqrt{|A_1|/|\mathcal{G}_{\min}|}),$$

and

$$\lim_{n\to\infty} P(\widehat{\boldsymbol{\omega}} \in \Omega^*) \to 1.$$

Note that for any submodel b that does not belong to $\mathcal{M}$, $\boldsymbol{\pi}_b^{\mathrm{T}}\boldsymbol{\theta}^*_{i(b)} = \boldsymbol{\theta}^0_i$. With Theorem 2, there exists an event $E = \left\{\sum_{b=1}^{B_n} \widehat{\omega}_b \boldsymbol{\pi}_b^{\mathrm{T}}\boldsymbol{\theta}^*_{i(b)} = \boldsymbol{\theta}^0_i\right\}$ which satisfies $\lim_{n\to\infty} P(E^c) \to 0$. Thus,

$$\begin{aligned}
&\sup_i \|\widehat{\boldsymbol{\theta}}_{i\widehat{\boldsymbol{\omega}}} - \boldsymbol{\theta}^0_i\| \\
=&\sup_i \left\| \sum_{b=1}^{B_n} \widehat{\omega}_b \boldsymbol{\pi}_b^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{i(b)} - \sum_{b=1}^{B_n} \widehat{\omega}_b \boldsymbol{\pi}_b^{\mathrm{T}} \boldsymbol{\theta}^*_{i(b)} + \sum_{b=1}^{B_n} \widehat{\omega}_b \boldsymbol{\pi}_b^{\mathrm{T}} \boldsymbol{\theta}^*_{i(b)} - \boldsymbol{\theta}^0_i \right\| \\
=&\sup_i \sup_b \|\widehat{\boldsymbol{\theta}}_{i(b)} - \boldsymbol{\theta}^*_{i(b)}\| \\
=&O_p(\sqrt{|A_1|/|\mathcal{G}_{\min}|}).
\end{aligned}$$

References

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

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112**, 410–423.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

6. Additional simulation

We conduct additional simulation to more comprehensively examine performance of the proposed approach.

(i) Consider Simulation 1 with $K = 2$, $\rho = 0.3$, $\beta = 2$, and pr=(0.5, 0.5). Consider the sequence with p =20, 40, 60, 80, 100. With the proposed approach, B_n is set as 2, 4, 6, 8, 10, correspondingly. We consider the direct application of double penalization without model averaging – this approach is referred to as "DP". Analysis is conducted on a server with 20 CPUs and 18G capacity per CPU. In Table A2, we present the mean, median, and SD of $\hat{K}$, as well as the percentage of $\hat{K}$ equal to the true number of subgroups. In addition, we also evaluate computer time. It is observed that when p =20 and 40, the DP approach has satisfactory performance. However, with a larger p, it cannot effectively identify the subgrouping structure. Its computer time increases very fast, compared to a much modest increase of the proposed approach. We acknowledge that, with more efficient coding, computer time of both approaches can be reduced. However, since the computation of the proposed approach is built on that of DP, similar trends would be expected. We further compare Accuracy, MSE, as well as TP and FP rates in Table A3, where we observe patterns similar to those in Table 2. This set of simulation re-establishes the advantage over DP and necessity of model averaging. In addition, with the proposed approach, as the analysis of a submodel is just an application of the DP approach, this simulation also suggests that the sizes of submodels should be kept moderate.

[Table 2 about here.]

[Table 3 about here.]

(ii) Consider settings similar to those above, with the difference that p =200, 400, and 600. Note that here the data dimensions are larger than the sample size and much larger than in some of the existing studies. With the proposed approach, B_n is set as 20, 40,

and 60, correspondingly. The identified number of subgroups is evaluated in Table A4. The subgrouping, estimation, and identification results are evaluated in Table A5. Here it is noted that when p is above 200, the *fmrs* software returns error messages. Thus, the KC approach is not included in comparison. The overall findings are similar to those above: the proposed approach can properly identify the number of subgroups, and it has performance superior to the Fitted and Underfitted approaches.

[Table 4 about here.]

[Table 5 about here.]

(iii) With the proposed approach, the random errors have been assumed to be sub-Gaussian. This assumption is common in high-dimensional regression studies and facilitates statistical developments. Here we use simulation to examine performance of the proposed approach when this assumption is violated. In particular, we consider Simulation 1 with $B_n = 10$ and various ρ and β values. Different from those above, the random errors are generated from a t distribution with degree of freedom 2. The results are presented in Tables A6 and A7. As expected, with heavier tailors, the results are inferior to those with normal errors. In particular, when the signal level is low, the proposed approach cannot effectively identify subgroups and important covariates. However, when the signal level is high, it can still have satisfactory performance, with patterns similar to those previously observed. It is noted that the proposed approach is not designed to be robust. When heavy-tailed distributions are present, robust goodness-of-fit measures would be needed. Such an exploration is beyond the scope of this study.

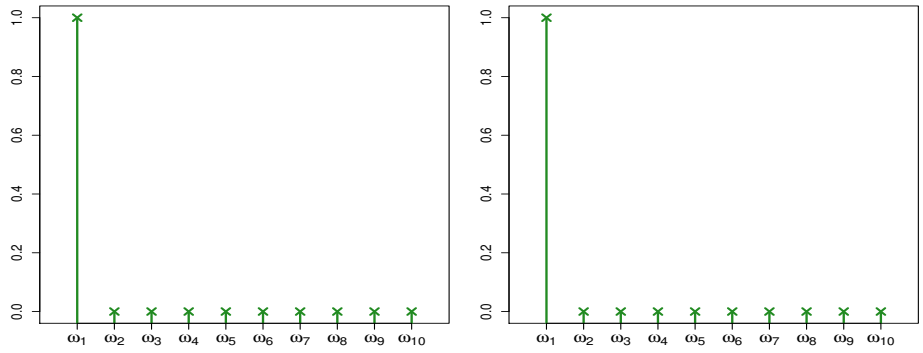
[Table 6 about here.]

[Table 7 about here.]

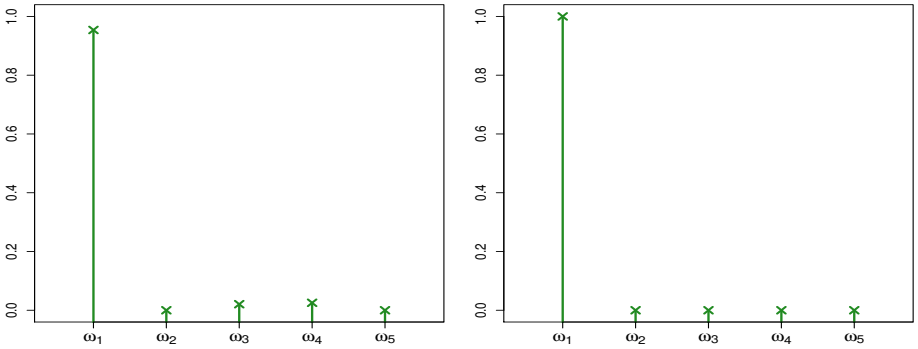
(iv) To evaluate the sensitivity of the analysis results to tuning parameter selection, we

consider Simulation 1 with $K = 2$, $B_n = 10$, $\rho = 0.3$, $\alpha = 2$ and pr $= (0.5, 0.5)$. For each simulated replicate, we first find the optimal tunings using the BIC. Then we fix one tuning parameter at its optimal value and increase/decrease the value of the other tuning to the next value in our grid search. With 100 replicates, the results are presented in Table A8. For example, the row with λ_1 "Increased" and λ_2 "Fixed" means that, for each replicate, the value of λ_1 is increased from its optimal, and the value of λ_2 is fixed at its optimal. The summary results are close to their counterparts in Tables 1 and 2, suggesting reasonable stability.

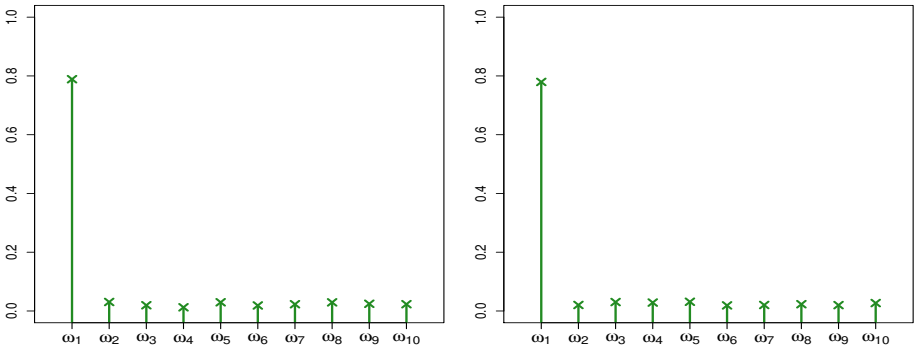
[Table 8 about here.]



(a) $K = 2$, $B_n = 10$



(b) $K = 2$, $B_n = 5$



(c) $K = 3$, $B_n = 10$

Figure A1: Simulation: average weights of the B_n candidate models. (a) and (b): pr=(0.5, 0.5) on the left and pr=(0.3, 0.7) on the right. (c): pr=(1/3, 1/3, 1/3) on the left and pr=(0.3, 0.3, 0.4) on the right.

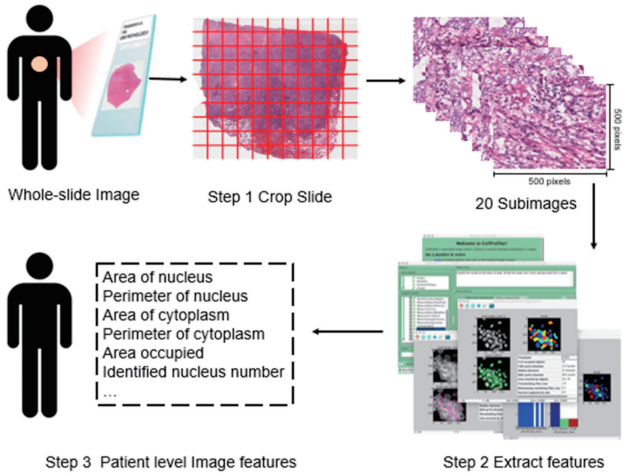


Figure A2: Data analysis: pipeline for extracting imaging features.

Table A1: Simulation 2 and 3: accuracy rate of correctly identifying subgroup membership (Accuracy), estimation mean squared error, and TP and FP rates.

				pr= $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$						pr= $(0.3, 0.3, 0.4)$					
K	ρ	β	Index	Proposed	Oracle	KC	True	Fitted	Underfitted	Proposed	Oracle	KC	True	Fitted	Underfitted
3	0	2	Accuracy	0.829		0.406	0.861	0.720	0.542	0.852		0.522	0.869	0.759	0.544
			MSE	8.123	0.221	74.901	12.031	36.948	89.083	10.053	0.228	74.562	13.234	82.962	77.901
			TP	1	1	0.983				1	1	0.475			
			FP	0.359	0.015	0.930				0.394	0.017	0.208			
	0.3	2	Accuracy	0.906		0.51	0.890	0.744	0.557	0.904		0.514	0.869	0.745	0.548
			MSE	5.156	0.230	78.021	8.625	65.912	95.319	7.233	0.227	75.584	25.129	102.875	98.346
			TP	1	1	0.419				1	1	0.442			
			FP	0.430	0.015	0.230				0.474	0.017	0.243			
	0.7	2	Accuracy	0.930		0.442	0.896	0.750	0.597	0.909		0.435	0.896	0.810	0.566
			MSE	3.646	0.359	92.502	54.346	235.571	157.144	4.859	0.346	91.819	93.420	95.073	218.966
			TP	1	1	0.973				0.998	1	0.980			
			FP	0.489	0.017	0.966				0.54	0.018	0.964			
				pr= $(0.5, 0.5)$						pr= $(0.3, 0.7)$					
				Proposed	Oracle	KC	True	Fitted	Underfitted	Proposed	Oracle	KC	True	Fitted	Underfitted
2	0	1	Accuracy	0.766		0.503	0.793	0.746	0.506	0.759		0.559	0.849	0.806	0.527
			MSE	0.661	0.035	4.963	0.575	2.920	5.693	1.807	0.058	5.083	1.283	0.647	5.920
			TP	0.935	1	0.398				0.870	1	0.488			
			FP	0.008	0.011	0.178				0.009	0.021	0.164			
		2	Accuracy	0.890		0.503	0.891	0.840	0.508	0.924		0.550	0.913	0.892	0.530
			MSE	0.340	0.058	19.520	1.063	4.440	21.922	0.459	0.065	20.166	1.336	2.157	21.748
			TP	1	1	0.438				1	1	0.515			
			FP	0.003	0.011	0.205				0.007	0.014	0.202			
	0.3	1	Accuracy	0.826		0.503	0.837	0.734	0.520	0.830		0.543	0.892	0.861	0.545
			MSE	0.322	0.043	5.318	0.646	3.630	6.650	1.168	0.059	5.377	0.070	0.952	6.662
			TP	0.980	1	0.285				0.898	1	0.405			
			FP	0.006	0.011	0.184				0.008	0.017	0.214			
		2	Accuracy	0.925		0.499	0.911	0.876	0.516	0.924		0.566	0.937	0.896	0.544
			MSE	0.193	0.059	19.436	11.352	11.379	26.920	0.294	0.076	19.915	0.074	5.102	26.14
			TP	1	1	0.360				1	1	0.492			
			FP	0.003	0.011	0.219				0.005	0.014	0.209			
	0.7	1	Accuracy	0.764		0.507	0.884	0.844	0.631	0.799		0.563	0.918	0.899	0.669
			MSE	1.619	0.084	5.833	5.878	3.349	13.486	1.89	0.112	5.704	0.152	0.534	10.591
			TP	0.754	1	0.285				0.800	1	0.350			
			FP	0.030	0.015	0.207				0.02	0.017	0.21			
		2	Accuracy	0.936		0.507	0.930	0.891	0.620	0.954		0.534	0.955	0.907	0.679
			MSE	0.185	0.105	22.925	1.491	7.799	36.316	0.297	0.163	23.076	24.099	222.613	138.869
			TP	1	1	0.332				1	1	0.382			
			FP	0.016	0.011	0.265				0.022	0.016	0.281			

Table A2: Simulation: mean, median, standard deviation (SD) of $\widehat{K}$, percentage (per) of $\widehat{K}$ equal to the true number of subgroups, and computer time (minutes), with $p = 20, 40, 60, 80, 100$.

p	B_n	method	mean	median	SD	per	time
20	2	Proposed	2	2	0	1	9.2
		DP	2	2	0	1	11.5
40	4	Proposed	2	2	0	1	12.3
		DP	2	2	0	1	169.8
60	6	Proposed	2	2	0	1	14.5
		DP	1.25	1	0.50	0.25	176.0
80	8	Proposed	2	2	0	1	40.8
		DP	1.50	1.5	0.71	0.50	789.6
100	10	Proposed	2	2	0	1	48.0
		DP	1	1	0	0	915.7

Table A3: Simulation: accuracy rate of correctly identifying subgroup memberships (Accuracy), mean squared error, TP and FP rates, with $p = 20, 40, 60, 80, 100$.

p	B_n	Index	Proposed	DP	Oracle	KC	True	Fitted	Underfitted
20	2	Accuracy	0.928	0.895		0.697	0.920	0.870	0.521
		MSE	0.471	0.410	0.065	11.916	0.047	2.995	25.774
		TP	1	1	1	0.968			
		FP	0.011	0.024	0.032	0.589			
40	4	Accuracy	0.923	0.705		0.545	0.906	0.853	0.519
		MSE	0.517	1.203	0.067	24.613	1.707	1.846	26.151
		TP	1	1	1	0.614			
		FP	0.020	0.240	0.022	0.485			
60	6	Accuracy	0.927	0.499		0.503	0.933	0.937	0.515
		MSE	0.572	27.175	0.056	29.869	0.053	0.546	25.919
		TP	1	1	1	0.969			
		FP	0	0.980	0.029	0.982			
80	8	Accuracy	0.913	0.498		0.496	0.895	0.913	0.542
		MSE	0.470	25.850	0.059	26.041	0.045	0.050	18.494
		TP	1	1	1	0.500			
		FP	0	0.940	0.013	0.566			
100	10	Accuracy	0.914	0.496		0.495	0.922	0.886	0.495
		MSE	0.577	23.509	0.026	31.438	0.035	1.045	28.642
		TP	1	1	1	0.375			
		FP	0.003	0.980	0.010	0.427			

Table A4: Simulation: mean, median, standard deviation (SD) of $\widehat{K}$, and percentage (per) of $\widehat{K}$ equal to the true number of subgroups, with $p = 200, 400, 600$.

p	B_n	mean	median	SD	per
200	20	2	2	0	1
400	40	2.02	2	0.141	0.98
600	60	1.98	2	0.141	0.98

Table A5: Simulation: accuracy rate of correctly identifying subgroup memberships (Accuracy), mean squared error, TP and FP rates, with $p = 200, 400, 600$.

p	B_n	Index	Proposed	Oracle	True	Fitted	Underfitted
200	20	Accuracy	0.907		0.913	0.821	0.543
		MSE	0.516	0.315	0.357	3.448	28.611
		TP	1.000	1.000			
		FP	0.029	0.018			
400	40	Accuracy	0.851		0.835	0.796	0.519
		MSE	0.868	0.449	0.209	5.610	19.753
		TP	1.000	1.000			
		FP	0.052	0.011			
600	60	Accuracy	0.807		0.830	0.805	0.541
		MSE	0.952	0.562	0.603	3.742	27.670
		TP	1.000	1.000			
		FP	0.067	0.009			

Table A6: Simulation: mean, median, standard deviation (SD) of $\widehat{K}$, and percentage (per) of $\widehat{K}$ equal to the true number of subgroups, with the random errors generated from a t distribution.

ρ	β	mean	median	SD	per	mean	median	SD	per
			pr= $(0.5, 0.5)$				pr= $(0.3, 0.7)$		
0	1	1.25	1	0.50	0.25	1.25	1	0.46	0.25
	2	2	2	0	1.00	2	2	0	1
0.3	1	1.44	1	0.54	0.40	1.32	1	0.47	0.32
	2	2.02	2	0.25	0.94	2	2	0	1
0.7	1	1.50	1	0.58	0.50	2	2	1	0.57
	2	2.02	2	0.16	0.98	2	2	0	1

Table A7: Simulation: accuracy rate of correctly identifying subgroup memberships (Accuracy), mean squared error, TP and FP rates, with the random errors generated from a t distribution.

			pr= $(0.5, 0.5)$						pr= $(0.3, 0.7)$					
ρ	β	Index	Proposed	Oracle	KC	True	Fitted	Underfitted	Proposed	Oracle	KC	True	Fitted	Underfitted
0	1	Accuracy	0.498		0.497	0.694	0.659	0.507	0.608		0.531	0.702	0.694	0.526
		MSE	5	0.430	7.494	0.364	0.472	7.964	5.082	0.968	6.588	1.190	4.517	7.282
		TP	0	1	0.250				0.797	0.938	0.516			
		FP	0.065	0.034	0.271				0.042	0.030	0.288			
	2	Accuracy	0.805		0.500	0.810	0.751	0.514	0.842		0.549	0.843	0.799	0.555
		MSE	1.376	0.291	29.328	1.290	5.627	27.198	1.831	0.324	27.319	1.497	8.502	27.435
		TP	1	1	0.295				1	1	0.492			
		FP	0.054	0.015	0.238				0.077	0.022	0.253			
0.3	1	Accuracy	0.521		0.501	0.724	0.644	0.516	0.585		0.545	0.779	0.719	0.551
		MSE	5.068	0.399	7.794	3.082	5.066	9.619	6.096	0.751	7.177	1.141	14.549	14.243
		TP	0.245	0.995	0.252				0.570	0.965	0.348			
		FP	0.087	0.023	0.218				0.066	0.032	0.198			
	2	Accuracy	0.826		0.503	0.831	0.788	0.549	0.868		0.533	0.866	0.850	0.579
		MSE	1.901	0.332	29.023	6.403	16.981	26.417	1.819	0.581	29.425	4.582	1.871	33.600
		TP	0.980	1	0.300				1	0.998	0.430			
		FP	0.204	0.014	0.260				0.179	0.017	0.258			
0.7	1	Accuracy	0.504		0.501	0.794	0.754	0.614	0.602		0.549	0.814	0.787	0.661
		MSE	5.962	0.926	7.482	1.363	28.818	8.638	7.257	2.662	6.983	2.728	4.322	11.748
		TP	0.375	0.969	0.250				0.500	0.857	0.357			
		FP	0.099	0.035	0.293				0.099	0.027	0.302			
	2	Accuracy	0.862		0.505	0.879	0.838	0.658	0.888		0.525	0.886	0.869	0.694
		MSE	2.575	0.801	29.849	1.029	17.209	133.790	3.102	2.628	30.145	150.783	5.803	38.736
		TP	1	1	0.375				1	0.965	0.410			
		FP	0.333	0.018	0.326				0.342	0.021	0.311			

Table A8: Mean, median, standard deviation (SD) of $\widehat{K}$, percentage (per) of $\widehat{K}$ equal to the true number of subgroups. Accuracy rate of correctly identifying subgroup memberships (Accuracy), mean squared error, and TP and FP rates.

		K							
λ_1	λ_2	mean	median	SD	per	Accuracy	MSE	TP	FP
Increased	Fixed	1.98	2	0.141	0.98	0.908	0.748	0.972	0.004
Fixed	Increased	1.75	2	0.439	0.80	0.904	1.108	1	0.005
Decreased	Fixed	2	2	0	1	0.919	0.573	1	0.163
Fixed	Decreased	2.12	2	0.328	0.88	0.880	0.940	1	0.059