A MAJORIZATION-MINIMIZATION APPROACH TO VARIABLE SELECTION USING SPIKE AND SLAB PRIORS: SUPPLEMENTARY MATERIAL

SA. Some remarks on the log-sum function. We briefly discuss properties of the log-sum function stated in (3.7). First, note that $\log(1+\tau_3^{-1}|\beta_j|) = \log(\tau_3+|\beta_j|)-\log(\tau_3)$. By multiplying -1 to the sum $\sum_{j=1}^{p}\log(\tau_3+|\beta_j|)$ and let $\tau_3 \to 0$, one obtains the logarithm of the product of $1/|\beta_j|$ over $j = 1, 2, \cdots, p$. As pointed out by Tipping [7], the term $1/|\beta_j|$ is an improper version of Student's t density. A rather different way is to see the log-sum function $\sum_{j=1}^{p}\log(1+\tau_3^{-1}|\beta_j|)$ as a product of logarithm of the generalized Pareto density, which has a parametric form given by

$$p_{\mathrm{GP}}(z) = \frac{1}{a_2}\left(1+\frac{a_3(z-a_1)}{a_2}\right)^{-(1/a_3+1)}$$

for $z \in (a_1, \infty)$, $a_1 \in (-\infty, \infty)$, $a_2 \in (0, \infty)$, and $a_3 \in (-\infty, \infty)$. By multiplying -2 and adding a constant term $-p\log\tau_3$ to $\sum_{j=1}^{p}\log(1+\tau_3^{-1}|\beta_j|)$, it becomes $\log\prod_{j=1}^{p}\tau_3^{-1}(1+|\beta_j|/\tau_3)^{-2}$, which is a logarithm of the product of generalized Pareto densities with location parameter $a_1 = 0$, scale parameter $a_2 = \tau_3$, and shape parameter $a_3 = 1$.

The following two propositions state the relationships between the log-sum function and the l_0 and l_1 norms. The first one states that the error rate

between the log-sum function and the l_0 norm measured by an l_1 distance is of order $-\log\tau_3$ as $\tau_3 \to 0$. Proofs of the two propositions will be given in Appendix B.

PROPOSITION SA.1. *Define* $g_1(\beta_j;\tau_3) = \log(1+\tau_3^{-1}|\beta_j|)/\log(1+\tau_3^{-1})$ *and* $g_2(\beta_j) = \mathbb{I}(\beta_j \neq 0)$. *Then for* $\tau_3 \in [0,1)$, *there exists a positive constant* C_1 *such that*

$$||g_1(\beta;\tau_3) - g_2(\beta)||_1 \leq C_1\left(\frac{p}{-\log\tau_3}\right). \tag{SA.1}$$

A graphical representation of Proposition SA.1 can be found in Figure 1. The next proposition states that the log-sum function can do better in approximating the l_0 norm than the l_1 norm as $\tau_3 \to 0$. On the other hand, results in this proposition also implies that the log-sum function approaches to l_1 norm as $\tau_3 \to \infty$. Sriperumbudur et al. [6] gave another heuristic argument for this property.

PROPOSITION SA.2. *With the same notation used in Proposition SA.1, for* $\beta_j \neq 0$ *and* $s \in [0,1]$, *we have*

$$\lim_{\tau_3\to 0}\frac{g_1(\beta_j;\tau_3)}{|\beta_j|^s} = |\beta_j|^{-s},$$

and

$$\lim_{\tau_3\to \infty}\frac{g_1(\beta_j;\tau_3)}{|\beta_j|^s} = |\beta_j|^{1-s}.$$

Therefore, $\lim_{\tau_3 \to 0} g_1(\beta_j; \tau_3) = \mathbb{I}(\beta_j \neq 0)$ *and* $\lim_{\tau_3 \to \infty} g_1(\beta_j; \tau_3) = |\beta_j|$.

SB. Connection with other approaches. Fuchs [5], Donoho et al. [3], and Tropp [8] independently showed that under some regular conditions, regression coefficients estimated with the l_0 norm constraint can be approximated by those estimated with the l_1 norm constraint. The advantage of using the l_1 norm instead of the l_0 norm as a constraint on regression coefficients is that minimization with an l_1 norm constraint is a convex optimization problem while the minimization problem with an l_0 norm constraint is combinatorial in nature.

However, as shown by Fan and Li [4], the l_1 norm tends to provide larger penalty values to large coefficients and smaller penalty values to small coefficients. Therefore large coefficients tend to be biased estimated while zero-valued coefficients tend to be estimated with non-zero values. On the other hand, the l_0 norm provides equal penalty values to all coefficients, therefore is more likely to shrink small coefficients to zero and keep large coefficients unchanged. Candés et al. [1] proposed a reweighted l_1 approach for sparse recovery. They showed that the l_0 norm can be better approximated by the sum of some log functions than the conventional l_1 norm. Sriperumbudur et al. [6] further explored the idea and used a modified log-sum function to approximate the l_0 norm in solving sparse generalized eigenvalue problems

related to principal component analysis, canonical correlation analysis, and Fisher's discriminant analysis.

We have noticed that the use of binary indicators for variable selection has been studied by Yuan and Lin under the name of non-negative garrotte estimator [10]. However, the estimation procedure associated to the non-negative garrotte estimator is quite different from the BAVA-MIO estimation proposed in this paper. In Yuan and Lin's proposal, the non-negative garrotte estimation is carried out via a two stage procedure. In the first stage, least squares estimation is proposed to obtain an initial estimate for each regression coefficient. In the second stage, a soft-thresholding estimation is proposed to obtain an estimate for the binary indicator associated to each regression coefficient. Under the soft-thresholding estimation, the estimate for the binary indicator is continuous on the interval between 0 and 1. In this sense, the non-negative garrotte estimation can be seen as a shrinkage estimation on the least squares estimate.

We have also noticed that the objective function stated in (3.15) is similar to the one used to obtain the adaptive elastic net recently developed by Zou and Zhang [11]. However, the weights used in the adaptive elastic net objective are fixed while in (3.15) the weights are iteratively changed throughout the optimization procedure. In addition, the adaptive elastic net did not see

the l_1 norm with adaptive weights as an approximation to the l_0 norm.

APPENDIX A: DERIVATION OF THE SOFT-THRESHOLDING OPERATOR REPRESENTATION

In obtaining the BAVA-MIO estimator, the soft-thresholding operator (3.16) is used to build a coordinate descent algorithm for approximating the minimizer of the objective function (3.15). Generally speaking, a coordinate algroithm is an iteration procedure aiming to minimize an objective function coordinate-wisely. Here the word "coordinate-wisely" means that at each iteration only one coordinate of the minimizer is considered for optimization given that all other coordinates are fixed. Here we focus on the jth coordinate and derive an explicit form for the soft-thresholding operator (3.16). For simplicity, we drop the index m_1 in (3.16) and let $\rho = \rho_{\lambda,\kappa,\sigma^2\tilde{\phi}}$. Define $\beta_j^+ = \beta_j$ if $\beta_j \geq 0$ and $\beta_j^+ = 0$ if $\beta_j < 0$. Further define $\beta_j^- = \beta_j$ if $\beta_j \leq 0$ and $\beta_j^- = 0$ if $\beta_j > 0$. With the definitions given above, we can write $|\beta_j| = \beta_j^+ - \beta_j^-$ and $\beta_j = \beta_j^+ + \beta_j^-$. Since we only focus on the jth coordinate, we rewrite the l_2 loss $||y - X\beta||_2^2$ as $\sum_{i=1}^n (\tilde{r}_{i,-j} - x_{ij}\beta_j)^2$ for notation convenience, where $\tilde{r}_{i,-j} = y_i - \sum_{j' \neq j} x_{ij'}\tilde{\beta}_{j'}$ and $\tilde{\beta}_{j'}$'s are fixed constants. Note that given all other coordinates are fixed, the problem of minimizing the objective function (3.15) with respect to β_j is equivalent to

the following constrained optimization problem:

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^{n}\left(\tilde{r}_{i,-j} - x_{ij}(\beta_j^+ + \beta_j^-)\right)^2 \\ & \text{subject to} && (\beta_j^+ + \beta_j^-)^2 \le t_1, (\beta_j^+ - \beta_j^-) \le t_2, -\beta_j^+ \le 0, \beta_j^- \le 0. \end{aligned} \tag{A.1}$$

The Lagrangian associated to problem (A.1) is given by

$$\begin{aligned} L(\beta_j^+, \beta_j^-, \lambda, \rho, \rho_1, \rho_2) &= \sum_{i=1}^{n}\left(\tilde{r}_{i-j} - x_{ij}(\beta_j^+ + \beta_j^-)\right)^2 \\ &\quad +\lambda[(\beta_j^+ + \beta_j^-)^2 - t_1] + \rho(\beta_j^+ - \beta_j^- - t_2) \\ &\quad +\rho_1(-\beta_j^+) + \rho_2\beta_j^-. \end{aligned}$$

The KKT conditions associated to problem (A.1) are given by

$$\begin{aligned} (\beta_j^+ + \beta_j^-)^2 - t_1 \le 0, \quad \lambda[(\beta_j^+ + \beta_j^-)^2 - t_1] &= 0, \\ \beta_j^+ - \beta_j^- - t_2 \le 0, \quad \rho(\beta_j^+ - \beta_j^- - t_2) &= 0, \\ -\beta_j^+ \le 0, \quad \rho_1(-\beta_j^+) &= 0, \\ \beta_j^- \le 0, \quad \rho_2\beta_j^- &= 0, \\ \lambda, \quad \rho, \quad \rho_1, \quad \rho_2 &\ge 0, \\ -2\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j} + 2\beta_j^+\sum_{i=1}^{n} x_{ij}^2 + 2\lambda\beta_j^+ + \rho - \rho_1 &= 0, \\ -2\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j} + 2\beta_j^-\sum_{i=1}^{n} x_{ij}^2 + 2\lambda\beta_j^- - \rho + \rho_2 &= 0. \end{aligned} \tag{A.2}$$

For the third line in the KKT conditions (A.2), the complementary slackness condition further implies that $\beta_j^+ > 0$ if and only if $\rho_1 = 0$, and $\beta_j^+ = 0$ if

and only if $\rho_1 > 0$. A similar argument can be applied to the fourth line in (A.2). With the arguments given above, the sixth and seventh lines in (A.2) jointly imply that

$$\left|\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j}\right| \leq \frac{\rho}{2} \tag{A.3}$$

if and only if $\beta_j = 0$, and otherwise if and only if $\beta_j \neq 0$. Now with condition (A.3) and the sixth line in (A.2), we can derive a closed form solution for β_j^+, which is given by

$$\tilde{\beta}_j^+ = \left(\sum_{i=1}^{n} x_{ij}^2 + \lambda\right)^{-1}\left(\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j} - \frac{\rho}{2}\right)_+. \tag{A.4}$$

A similar argument can be applied to derive a closed form solution for β_j^-, which is given by

$$\tilde{\beta}_j^- = \left(\sum_{i=1}^{n} x_{ij}^2 + \lambda\right)^{-1}\left(\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j} + \frac{\rho}{2}\right)_-. \tag{A.5}$$

Combining (A.4) and (A.5), we get

$$\tilde{\beta}_j = \left(\sum_{i=1}^{n} x_{ij}^2 + \lambda\right)^{-1} ST\left(\sum_{i=1}^{n} x_{ij}\tilde{r}_{i,-j}, \frac{\rho}{2}\right), \tag{A.6}$$

where $ST(a,b)$ is a soft-thresholding operator defined by $ST(a,b) = \text{sign}(a)(|a| - b)_+$.

APPENDIX B: PROOFS OF PROPOSITION SA.1 AND PROPOSITION SA.2

Proof of Proposition SA.1. Note that

$$\begin{aligned} g_1(\beta_j;\tau_3) - g_2(\beta_j) &= \frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})} - \mathbb{I}(\beta_j \neq 0) \\ &= \frac{\log(1+\tau_3^{-1}|\beta_j|) - \mathbb{I}(\beta_j \neq 0)\log(1+\tau_3^{-1})}{\log(1+\tau_3^{-1})} \\ &= \frac{\log(\tau_3+|\beta_j|) - \mathbb{I}(\beta_j \neq 0)\log(\tau_3+1)}{-\log(\tau_3)+\log(\tau_3+1)} \\ &\quad + \frac{(\mathbb{I}(\beta_j \neq 0)-1)\log(\tau_3)}{-\log(\tau_3)+\log(\tau_3+1)}. \end{aligned}$$

For the case of $\beta_j = 0$,

$$\frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})} - \mathbb{I}(\beta_j \neq 0) = \frac{\log(\tau_3)-\log(\tau_3)}{-\log(\tau_3)+\log(\tau_3+1)} = 0. \tag{B.1}$$

For the case of $\beta_j \neq 0$,

$$\frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})} - \mathbb{I}(\beta_j \neq 0) = \frac{\log(\tau_3+|\beta_j|)-\log(\tau_3+1)}{-\log(\tau_3)+\log(\tau_3+1)}. \tag{B.2}$$

The numerator in (B.2) is bounded from below and from above with $\beta_j \neq 0$ and $\tau_3 \in [0,1)$, therefore there exist two positive constants $C_2, C_3 \in \mathbb{R}$ such that

$$-\infty < -C_2 \leq \log(\tau_3+|\beta_j|) - \log(\tau_3+1) \leq C_3 < \infty,$$

for $j = 1, 2, \cdots, p$, and we can bound the numerator in a way that

$$|\log(\tau_3+|\beta_j|) - \log(\tau_3+1)| \leq C_1, \tag{B.3}$$

where $C_1 = C_2 \vee C_3$. Further note that for $\tau_3 \in [0, 1)$, the denominator in (B.2) is always greater than zero. Therefore by using (B.1), (B.3) and the fact that the denominator in (B.2) is a positive constant, we can bound $||g_1(\beta; \tau_3) - g_2(\beta)||_1$ in a way such that

$$\begin{aligned} ||g_1(\beta;\tau_3) - g_2(\beta)||_1 &= \sum_{j=1}^{p} \left| \frac{\log(1+\tau_3^{-1}|\beta_j|)}{\log(1+\tau_3^{-1})} - \mathbb{I}(\beta_j \neq 0) \right| \\ &\leq \sum_{j=1}^{p} C_1 \left| \frac{1}{-\log\tau_3 + \log(\tau_3+1)} \right| \\ &\leq C_1 \left(\frac{p}{-\log\tau_3} \right), \end{aligned}$$

which completes the proof. □

Proof of Proposition SA.2. By direct calculation, we have

$$\lim_{\tau_3 \to 0} \frac{g_1(\beta_j;\tau_3)}{|\beta_j|^s} = \frac{1}{|\beta_j|^s} \times \lim_{\tau_3 \to 0} \frac{-\log(\tau_3) + \log(\tau_3 + |\beta_j|)}{[-\log(\tau_3) + \log(\tau_3+1)]} = \frac{1}{|\beta_j|^s} \times 1 = |\beta_j|^{-s}.$$

On the other hand,

$$\begin{aligned} \lim_{\tau_3 \to \infty} \frac{g_1(\beta_j;\tau_3)}{|\beta_j|^s} &= \frac{1}{|\beta_j|^s} \times \lim_{\tau_3 \to \infty} \left[\tau_3 \log\left(\frac{\tau_3 + |\beta_j|}{\tau_3} \right) \right] \left[\tau_3 \log\left(\frac{\tau_3+1}{\tau_3} \right) \right]^{-1} \\ &= |\beta_j|^{-s} \times \frac{\log\exp(|\beta_j|)}{\log\exp(1)} \\ &= |\beta_j|^{1-s}, \end{aligned}$$

which completes the proof. □

APPENDIX C: PROOF OF THEOREM 5.1

We will use the following notations in the proof. For two vector $a = (a_1, a_2, \cdots, a_p)$ and $b = (b_1, b_2, \cdots, b_p)$, the notation $|a| \leq |b|$ means pairwise

inequalities hold for elements in a and b, i.e. $|a_j| \leq |b_j|$ for $j = 1, 2, \cdots, p$. Similar operations are applicable to $|a| > |b|$, $|a| \geq |b|$, $|a| < |b|$ and the function $\max(a, b)$. In addition, let $1^*_{p^*}$ denote the p^*-dimensional vector with entries all equal to 1.

Proof of Theorem 5.1. Define $\widehat{w}^{\tau_3} = \widehat{\beta}^{\tau_3} - \beta$. The sign consistency implies that if $\beta_j > 0$, then $\widehat{w}_j^{\tau_3} = \widehat{\beta}_j^{\tau_3} - \beta_j > -\beta_j$ should hold; if $\beta_j < 0$, $\widehat{w}_j^{\tau_3} = \widehat{\beta}_j^{\tau_3} - \beta_j < -\beta_j$ should hold; if $\beta_j = 0$, $\widehat{w}_j^{\tau_3} = \widehat{\beta}_j^{\tau_3} - \beta_j = 0$ should hold. In addition, it can be shown that $\widehat{w}^{\tau_3}$ is the minimizer of the following function

$$L(w) = ||\epsilon - Xw||_2^2 + \lambda||w + \beta||_2^2 + \rho \sum_{j=1}^{p} \frac{\log(1 + \tau_3^{-1}|w + \beta_j|)}{\log(1 + \tau_3^{-1})},$$

where $\epsilon = y - X\beta$. It means that $\widehat{w}^{\tau_3}$ is the solution for the following subgradient equations:

$$2 \begin{pmatrix} X_{S_0}^T X_{S_0} & X_{S_0}^T X_{S_0^c} \\ X_{S^c}^T X_{S_0} & X_{S_0^c}^T X_{S_0^c} \end{pmatrix} \begin{pmatrix} \widehat{w}_{S_0}^{\tau_3} \\ \widehat{w}_{S_0^c}^{\tau_3} \end{pmatrix} - 2 \begin{pmatrix} X_{S_0}^T \epsilon \\ X_{S_0^c}^T \epsilon \end{pmatrix}$$
$$+2\lambda \begin{pmatrix} \widehat{\beta}_{S_0}^{\tau_3} \\ \widehat{\beta}_{S_0^c}^{\tau_3} \end{pmatrix} \begin{pmatrix} \text{sign}(\widehat{\beta}_{S_0}^{\tau_3})/(\tau_3 + |\widehat{\beta}_{S_0}^{\tau_3}|) \\ \text{sign}(\widehat{\beta}_{S_0^c}^{\tau_3})/(\tau_3 + |\widehat{\beta}_{S_0^c}^{\tau_3}|) \end{pmatrix} \frac{\rho}{\log(1 + \tau_3^{-1})} = 0.$$

Then following conditions are necessary and sufficient for event E_{0,τ_3}:

$$E_1 = \Big\{ \beta : X_{S_0}^T X_{S_0} \widehat{w}_{S_0}^{\tau_3} - X_{S_0}^T \epsilon + \lambda(\widehat{w}_{S_0}^{\tau_3} + \beta_{S_0}) = -\frac{\rho \cdot \text{sign}(\beta_{S_0})}{2(\tau_3 + |\widehat{w}_{S_0}^{\tau_3} + \beta_{S_0}|) \log(1 + \tau_3^{-1})} \Big\},$$

and

$$E_3 = \Big\{\beta \quad : \quad -\frac{\rho}{2\tau_3 \log(1+\tau_3^{-1})} 1^*_{|S_0^c|} \leq X_{S_0^c}^T X_{S_0} \widehat{w}_{S_0}^{\tau_3} - X_{S_0^c}^T \epsilon \\ \leq \frac{\rho}{2\tau_3 \log(1+\tau_3^{-1})} 1^*_{|S_0^c|} \Big\}.$$

We will restrict our discussion on the following case:

$$E_2 = \Big\{\beta : |\widehat{w}_{S_0}^{\tau_3}| < |\beta_{S_0}| \Big\}. \tag{C.1}$$

Remember that $S_0 = \{j : \beta_j \neq 0\}$, therefore if (C.1) holds, then $\widehat{\beta}_{S_0}^{\tau_3} \neq 0$ will hold. To see why it is, let us consider the case when $\beta_j > 0$. Given that $\beta_j > 0$, the event E_2 implies $-\beta_j < \widehat{w}_j^{\tau_3} = \widehat{\beta}_j^{\tau_3} - \beta_j < \beta_j$, which further implies $\widehat{\beta}_j^{\tau_3}$ can not be zero. Moreover, $\widehat{\beta}_j^{\tau_3}$ has some value greater than zero. The same argument can be applied to the case when $\beta_j < 0$. Technically we express $|\widehat{\beta}_{S_0}^{\tau_3}| = \delta$ and given (C.1), $\delta > 0$ almost surely.

To continue our proof, we first solve the equations in E_1 to obtain a representation for $\widehat{w}_S^{\tau_3}$ in terms of C_{SS_0} and D_{S_0}. The representation is given by

$$\widehat{w}_S^{\tau_3} \quad = \quad n^{-1} C_{SS_0}^{-1} \Big(n^{1/2} D_{S_0} - \frac{\rho \cdot \text{sign}(\beta_{S_0})}{2(\tau_3 + \delta)\log(1+\tau_3^{-1})} - \lambda \beta_{S_0} \Big). \tag{C.2}$$

Then by plugging (C.2) on the left hand side of E_2, we obtain the following

inequality:

$$\begin{aligned}\left|n^{-1/2}C_{SS_0}^{-1}D_{S_0} \;-\; n^{-1}C_{SS_0}^{-1}\left(\frac{\rho\cdot\mathrm{sign}(\beta_{S_0})}{2(\tau_3+\delta)\log(1+\tau_3^{-1})}+\lambda\beta_{S_0}\right)\right| \\ \leq\; n^{-1/2}|C_{SS_0}^{-1}D_{S_0}| \\ +\frac{\rho}{2n}\left|C_{SS_0}^{-1}\left(\frac{\mathrm{sign}(\beta_{S_0})}{(\tau_3+\delta)\log(1+\tau_3^{-1})}+\frac{2\lambda}{\rho}\beta_{S_0}\right)\right|\end{aligned} \tag{C.3}$$

If the right hand side of (C.3) is smaller than $|\beta_{S_0}|$, then E_2 will hold. Denote the event by E_2'. Equivalently, E_2' can be written as

$$\begin{aligned}E_2' \;=\; &\Big\{\beta : |C_{SS_0}^{-1}D_{S_0}| < n^{1/2}|\beta_{S_0}| \\ &-\frac{\rho}{2n^{1/2}}\left|C_{SS_0}^{-1}\left(\frac{\mathrm{sign}(\beta_{S_0})}{(\tau_3+\delta)\log(1+\tau_3^{-1})}+\frac{2\lambda}{\rho}\beta_{S_0}\right)\right|\Big\}.\end{aligned}$$

Obviously, $E_2' \subseteq E_1 \cap E_2$. On the other hand, by plugging (C.2) in the middle term of E_3 and taking absolute value on it, we obtain the following inequality:

$$\begin{aligned}&\Big|n^{1/2}C_{S^cS_0}C_{SS_0}^{-1}D_{S_0} \\ &-C_{S^cS_0}C_{SS_0}^{-1}\left(\frac{\rho\cdot\mathrm{sign}(\beta_{S_0})}{2(\tau_3+\delta)\log(1+\tau_3^{-1})}+\lambda\beta_{S_0}\right)-n^{1/2}D_{S_0^c}\Big| \\ \leq\; &n^{1/2}\left|C_{S^cS_0}C_{SS_0}^{-1}D_{S_0}-D_{S_0^c}\right| \\ +\; &\left|C_{S^cS_0}C_{SS_0}^{-1}\left(\frac{\rho\cdot\mathrm{sign}(\beta_{S_0})}{2(\tau_3+\delta)\log(1+\tau_3^{-1})}+\lambda\beta_{S_0}\right)\right|.\end{aligned} \tag{C.4}$$

If the right hand side of (C.4) is smaller than $\rho[2\tau_3\log(1+\tau_3^{-1})]^{-1}1^*_{|S_0^c|}$, then

E_3 will hold. Denote the event by E_3'. Equivalently, E_3' can be written as

$$\begin{aligned} & E_3' \\ = & \left\{\beta : |C_{S^cS_0}C_{SS_0}^{-1}D_{S_0} - D_{S_0^c}| \leq \frac{\rho}{2\tau_3 \log(1+\tau_3^{-1})n^{1/2}}\right. \\ \times & \left.\left(1_{|S_0^*|}^* - \left|C_{S^cS_0}C_{SS_0}^{-1}\left(\frac{\tau_3 \cdot \text{sign}(\beta_{S_0})}{(\tau_3+\delta)} + \frac{2\tau_3\log(1+\tau_3^{-1})\lambda}{\rho}\beta_{S_0}\right)\right|\right)\right\}. \end{aligned} \tag{C.5}$$

Obviously $E_3' \subseteq E_1 \cap E_3$.

Note that E_2 is a restricted event, and since E_1 and E_3 are necessary and sufficient for E_{0,τ_3}, we have $E_1 \cap E_2 \cap E_3 \subseteq E_{0,\tau_3}$. Now by using the fact that $\mathbb{P}(E_1 \cap E_2 \cap E_3) \geq \mathbb{P}(E_2' \cap E_3')$, we have

$$\begin{aligned} \mathbb{P}(E_{0,\tau_3}) &\geq \mathbb{P}(E_2' \cap E_3') \\ &= 1 - \mathbb{P}(E_2'^{,c} \cup E_3'^{,c}) \\ &\geq 1 - \left[\mathbb{P}(E_2'^{,c}) + \mathbb{P}(E_3'^{,c})\right] \end{aligned}$$

Here $E_2'^{,c}$ and $E_3'^{,c}$ are the complements of E_2' and E_3', respectively. To bound the probability $\mathbb{P}(E_2'^{,c}) + \mathbb{P}(E_3'^{,c})$, first note that by Assumption 2 and the assumption that $p \propto n^\alpha$, we have $||D_{S_0}||_\infty \leq ||D_{S_0}||_1 = n^{-1/2}||X_{S_0}^T\epsilon|| \leq n^{-1/2}||X^T\epsilon||_1 = O(n^{-1/2}p) = O(n^{\alpha-1/2})$. Then

$$||C_{SS_0}^{-1}D_{S_0}||_\infty \leq \Lambda_{\min}(C_{SS_0})^{-1}||D_{S_0}||_\infty < c_{10}n^{\alpha-1/2}$$

by Assumption 1, where c_{10} is a positive finite constant. In addition, by Assumption 3, the second term on the right hand side of the inequality in

E_2' will goes to 0 as $n \to \infty$. Therefore for $E_2'^{,c}$, we have

$$\begin{aligned}\mathbb{P}(E_2'^{,c}) &\leq \mathbb{P}\bigg(||C_{SS_0}^{-1} D_{S_0}||_\infty \geq n^{1/2} \max_j |\beta_j|\bigg) \\ &\leq \mathbb{P}\bigg(c_{10} n^{\alpha - 1/2} \geq n^{1/2} \max_j |\beta_j|\bigg) \to 0 \end{aligned} \tag{C.6}$$

for $0 < \alpha < 1$ as $n \to \infty$. On the other hand, for the term on the left hand side of the ineuqality in E_3', we have

$$\begin{aligned} & ||C_{S^c S_0} C_{SS_0}^{-1} D_{S_0} - D_{S_0^c}||_\infty \\ \leq \;& ||C_{S^c S_0} D_{S_0}||_\infty \Lambda_{\min}(C_{SS_0})^{-1} + ||D_{S_0^c}||_\infty \\ \leq \;& n^{-1} ||D_{S_0^c}||_\infty \frac{\Lambda_{\max}(X_{S_0}^T X_{S_0})}{n^{-1}[\Lambda_{\min}(X_{S_0}^T X_{S_0}) + \lambda]} + ||D_{S_0^c}||_\infty, \end{aligned} \tag{C.7}$$

which is also bounded from above by Assumption 1 and Assumption 2. Further note that since $\tau_3 \propto n^{-1}$, therefore for $\tau_3 = c_8 n^{-1}$ with $0 < c_8 < \infty$ we have

$$\tau_3 \log(1 + \tau_3^{-1}) n^{1/2} = \frac{\log(1 + n/c_8)^{c_8}}{n^{1/2}} \to 0 \tag{C.8}$$

as $n \to \infty$. In addition, the last term on the right hand side of the inequality in E_3' also approaches to 0 as $n \to \infty$ under the assumption that $\tau_3 \propto n^{-1}$. This event guarantees that the quantity on the right hand side of the inequality in E_3' will remain non-negative as $n \to \infty$. Therefore by using

(C.7), we can bound the probability of $E_3^{\prime,c}$ by

$$\begin{aligned} & \mathbb{P}(E_3^{\prime,c}) \\ \leq \; & \mathbb{P}\bigg(||D_{S_0^c}||_\infty \frac{\Lambda_{\max}(X_{S_0}^T X_{S_0})}{\Lambda_{\min}(X_{S_0}^T X_{S_0}) + \lambda} + ||D_{S_0^c}||_\infty > \frac{\rho}{2\tau_3 \log(1+\tau_3^{-1}) n^{1/2}}\bigg), \end{aligned} \tag{C.9}$$

which will approach to 0 for $0 < \alpha < 1/2$ as $n \to \infty$ by (C.8) and the assumption that $||D_{S_0^c}||_\infty \leq n^{-1/2}||X^T\epsilon||_1 = O(n^{\alpha-1/2})$. By (C.6) and (C.9), when $0 < \alpha < 1/2$ and $\tau_3 \propto n^{-1}$, we have $\mathbb{P}(E_2^{\prime,c}) + \mathbb{P}(E_3^{\prime,c}) \to 0$, therefore $\mathbb{P}(E_{0,\tau_3}) \to 1$ as $n \to \infty$, which completes the proof. □

The following proposition summarizes the invariance of the BAVA-MIO estimator under the Irrepresentable Condition.

COROLLARY C.1. *Assume that $\delta > 0$, $\tau_3 \propto n^{-1}$ and*

$$1^*_{|S_0^c|} \cdot 0 < \left| C_{S^c S_0} C_{SS_0}^{-1} \left(\frac{\text{sign}(\beta_{S_0})}{(\tau_3 + \delta)} \right) \right| < 1^*_{|S_0^c|} \cdot \infty. \tag{C.10}$$

Then given Assumptions 1 and 3 hold, for estimator defined in (5.1), the inequality stated in $E_3^{\prime}$ will hold under the following condition:

$$1^*_{|S_0^c|} \leq \left| C_{S^c S_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0}) \right| < 1^*_{|S_0^c|} \cdot \infty. \tag{C.11}$$

Proof of Corollary C.1. We start the proof by defining

$$\begin{aligned} \alpha_1 &= \max\left(1, \frac{|C_{S^c S_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})|}{|C_{S^c S_0} C_{SS_0}^{-1} [\text{sign}(\beta_{S_0})(\tau_3 + \delta)^{-1}]|}\right). \\ \alpha_2 &= \max\left(1, \frac{|C_{S^c S_0} C_{SS_0}^{-1} [\text{sign}(\beta_{S_0})(\tau_3 + \delta)^{-1}]|}{|C_{S^c S_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})|}\right). \end{aligned}$$

Given Assumption 1, we have $\alpha_1 < \infty$, and given (C.11), we have $\alpha_2 < \infty$. The second term on the right hand side of the inequality in E_3' can be bounded from below in a way such that

$$\begin{aligned} 1^*_{|S_0^c|} \quad - \quad & \left| C_{S^cS_0} C_{SS_0}^{-1} \left(\frac{\tau_3 \cdot \text{sign}(\beta_{S_0})}{(\tau_3 + \delta)} + \frac{2\tau_3 \log(1+\tau_3^{-1})\lambda}{\rho} \beta_{S_0} \right) \right| \\ \geq \quad & 1^*_{|S_0^c|} - \left| C_{S^cS_0} C_{SS_0}^{-1} \left(\frac{\tau_3 \cdot \text{sign}(\beta_{S_0})}{(\tau_3 + \delta)} \right) \right| \\ & - \frac{2\tau_3 \log(1+\tau_3^{-1})\lambda}{\rho} |C_{S^cS_0} C_{SS_0}^{-1} \beta_{S_0}| \end{aligned} \tag{C.12}$$

The final term on the right hand side of (C.12) approaches to zero as $n \to \infty$ given that $\tau_3 \propto n^{-1}$ and Assumption 3 holds. The first two terms on the right hand side of (C.12) can be bounded in a way such that

$$\begin{aligned} & 1^*_{|S_0^c|} - \left| C_{S^cS_0} C_{SS_0}^{-1} \left(\frac{\tau_3 \cdot \text{sign}(\beta_{S_0})}{(\tau_3 + \delta)} \right) \right| \\ \geq \quad & 1 - \tau_3 \alpha_1 \left| C_{S^cS_0} C_{SS_0}^{-1} \left(\frac{\text{sign}(\beta_{S_0})}{(\tau_3 + \delta)} \right) \right| \\ = \quad & 1^*_{|S_0^c|} - \tau_3 \max \left(|C_{S^cS_0} C_{SS_0}^{-1} [\text{sign}(\beta_{S_0})(\tau_3 + \delta)^{-1}]|, \right. \\ & \left. |C_{S^cS_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})| \right) \\ = \quad & 1^*_{|S_0^c|} - \tau_3 \max \left(\frac{|C_{S^cS_0} C_{SS_0}^{-1} [\text{sign}(\beta_{S_0})(\tau_3 + \delta)^{-1}]|}{|C_{S^cS_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})|}, 1 \right) \\ & \times |C_{S^cS_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})| \\ = \quad & 1^*_{|S_0^c|} - \tau_3 \alpha_2 |C_{S^cS_0} C_{SS_0}^{-1} \text{sign}(\beta_{S_0})|, \end{aligned} \tag{C.13}$$

Given that $\tau_3 \propto n^{-1}$, the second term on the right hand side of (C.13) approaches to $1^*_{|S_0^c|} \cdot 0$ as $n \to \infty$, which implies that E_3' can still hold under (C.11), i.e. the Irrepresentable Condition is violated. □

APPENDIX D: PROOF OF THEOREM 5.2

We first prove the following Lemma.

LEMMA D.1. *Assume $S \neq S_0$. Then given Assumptions 4 to 6 hold, there exist some positive constants n^* and ξ such that*

$$\log f(y^n|\ \mathcal{M}_{S_0}) - \log f(y^n|\ \mathcal{M}_S) + \log f(\mathcal{M}_{S_0}) - \log f(\mathcal{M}_S) > n\xi.$$

for $n > n^*$.

Proof of Lemma D.1. For the Bayesian hierarchical model stated in (3.1), the Bayes' factor between $\mathcal{M}_S$ and $\mathcal{M}_{S_0}$ is given by

$$\begin{aligned}
& \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) \\
= \ & \frac{f(y^n|\ \gamma, \tau_1, \tau_2, \lambda)}{f(y^n|\ \gamma_0, \tau_1, \tau_2, \lambda)} \\
= \ & \frac{|\lambda^{-1} X_{S_0}^T X_{S_0} + I_{|S_0|}|^{1/2}}{|\lambda^{-1} X_S^T X_S + I_{|\mathcal{S}|}|^{1/2}} \frac{\Gamma((n+2\tau_1)/2)}{\Gamma((n+2\tau_1)/2)} \\
& \times \frac{\left[(y^n)^T \left(X_{S_0} X_{S_0}^T + \lambda I_n\right)^{-1} y^n + 2\lambda^{-1}\tau_2\right]^{[(n+2\tau_1)/2]}}{\left[(y^n)^T \left(X_S X_S^T + \lambda I_n\right)^{-1} y^n + 2\lambda^{-1}\tau_2\right]^{[(n+2\tau_1)/2]}} \\
= \ & \left[\frac{(y^n)^T \left(X_{S_0} X_{S_0}^T + \lambda I_n\right)^{-1} y^n + 2\lambda^{-1}\tau_2}{(y^n)^T \left(X_S X_S^T + \lambda I_n\right)^{-1} y^n + 2\lambda^{-1}\tau_2}\right]^{[(n+2\tau_1)/2]} \\
& \times H(X, \gamma, \gamma_0, \lambda), \qquad \text{(D.1)}
\end{aligned}$$

where $H(X, \gamma, \gamma_0, \lambda)$ is some function that collects the remainder terms. It can be shown that under Assumption 1, $H(X, \gamma, \gamma_0, \lambda)$ is bounded. By Assumption 6, we can bound the term $(y^n)^T (X_{S_0} X_{S_0}^T + \lambda I_n)^{-1} y^n$ from above in a way such that

$$\begin{aligned} (y^n)^T (X_{S_0} X_{S_0}^T + \lambda I_n)^{-1} y^n &\leq \frac{(y^n)^T y^n}{\Lambda_{\min}(X_{S_0} X_{S_0}^T) + \lambda} \\ &\leq \frac{(y^n)^T y^n}{\Lambda_{\min}(X_{S_0} X_{S_0}^T)} \\ &< (y^n)^T (X_S X_S^T + \lambda I_n)^{-1} y^n, \end{aligned} \tag{D.2}$$

which further implies that

$$\frac{(y^n)^T (X_{S_0} X_{S_0}^T + \lambda I_n)^{-1} y^n}{(y^n)^T (X_S X_S^T + \lambda I_n)^{-1} y^n} < 1. \tag{D.3}$$

The inequality (D.3) implies that the logarithm of Bayes's factor (D.1) is bounded away from above for all n. Define

$$\begin{aligned} K_1 &= \frac{1}{2} \log \left(\frac{(y^n)^T (X_{S_0} X_{S_0}^T + \lambda I_n)^{-1} y^n + 2\lambda^{-1} \tau_2}{(y^n)^T (X_S X_S^T + \lambda I_n)^{-1} y^n + 2\lambda^{-1} \tau_2} \right) \\ K_2 &= \tau_1 K_1 + \log H(X, \gamma, \gamma_0, \lambda). \end{aligned}$$

We can express the logarithm of Bayes' factor (D.1) in terms of K_1 and K_2 as

$$\begin{aligned} \log \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) &= \log f(y^n | \, \mathcal{M}_S) - \log f(y^n | \, \mathcal{M}_{S_0}) \\ &= n K_1 + K_2. \end{aligned} \tag{D.4}$$

Since $K_1 < 0$ and K_2 is bounded, (D.4) will converge to $-\infty$ as $n \to \infty$. From (D.4) we can see that

$$\begin{aligned} & \log f(y^n|\ \mathcal{M}_{S_0}) - \log f(y^n|\ \mathcal{M}_S) + \log f(\mathcal{M}_{S_0}) - \log f(\mathcal{M}_S) \\ = \ & -nK_1 - K_2 + \log \frac{f(\mathcal{M}_{S_0})}{f(\mathcal{M}_S)} \\ = \ & n\Big(- K_1 - \frac{K_2 - \log f(\mathcal{M}_{S_0}) + \log f(\mathcal{M}_S)}{n} \Big). \end{aligned}$$

Let $K_3 = K_2 - \log f(\mathcal{M}_{S_0}) + \log f(\mathcal{M}_S)$. Note that $-K_1 > 0$, therefore, for $K_3 < 0$, we let ξ^* such that $0 < \xi^* < -K_1$. For $K_3 > 0$, we choose some $n^* < n$ so that $-K_1 - K_3/n > -K_1 - K_3/n^* = \xi^{**} > 0$. Let $\xi = \xi^* \wedge \xi^{**}$ and n^* satisfying the condition given above, we complete the proof. □

Proof of Theorem 5.2. Note that the posterior probability $\mathbb{P}(\mathcal{M}_{S_0}|\ y^n)$ can be expressed in terms of Bayes' factors as

$$\text{(D.5)} \qquad \mathbb{P}(\mathcal{M}_{S_0}|\ y^n) \ = \ 1 - \frac{\sum_{S' \neq S_0} \mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n) f(\mathcal{M}_{S'})}{\sum_{S \in \mathcal{S}} \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) f(\mathcal{M}_S)}.$$

The first step for proving the theorem is to bound the tail probability (D.5) from below. We focus on deriving an upper bound for the second term on the right hand side of (D.5). One trick to derive this upper bound is to derive a lower bound for the denominator and an upper bound for the numerator. Define

$$\mathcal{S}_1 = \Big\{ S : \log \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) \geq -\frac{n\xi}{2} \Big\}.$$

Now for the denominator, we have

$$\begin{aligned}
&\sum_{S\in\mathcal{S}} \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) f(\mathcal{M}_S) \\
= \;& \sum_{S\in\mathcal{S}_1} \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) f(\mathcal{M}_S) + \sum_{S\in\mathcal{S}_1^c} \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) f(\mathcal{M}_S) \\
\overset{\text{(D.6)}}{\geq} \;& |\mathcal{S}_1| \sum_{S\in\mathcal{S}_1} \frac{1}{|\mathcal{S}_1|} \exp[\log \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_S)].
\end{aligned}$$

Since (D.6) is a sum of convex functions, a lower bound can be derived in a way such that

$$\begin{aligned}
&|\mathcal{S}_1| \sum_{S\in\mathcal{S}_1} \frac{1}{|\mathcal{S}_1|} \exp[\log \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_S)] \\
\geq \;& |\mathcal{S}_1| \exp\left\{ \frac{1}{|\mathcal{S}_1|} \sum_{S\in\mathcal{S}_1} \log \mathrm{BF}(\mathcal{M}_S, \mathcal{M}_{S_0}; y^n) + \frac{1}{|\mathcal{S}_1|} \sum_{S\in\mathcal{S}_1} \log f(\mathcal{M}_S) \right\} \\
\overset{\text{(D.7)}}{=} \;& |\mathcal{S}_1| \exp\left\{ -\frac{1}{|\mathcal{S}_1|} \sum_{S\in\mathcal{S}_1} \log \frac{f(y^n|\,\mathcal{M}_{S_0})}{f(y^n|\,\mathcal{M}_S)} + \frac{1}{|\mathcal{S}_1|} \sum_{S\in\mathcal{S}_1} \log f(\mathcal{M}_S) \right\}.
\end{aligned}$$

Now we turn to derive an upper bound for the numerator. Define

$$\tilde{S} = \arg \max_{S'\in\mathcal{S}\setminus S_0} \left\{ \log \mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_{S'}) \right\}.$$

An upper bound for the numerator can be derived in a way that

$$\begin{aligned}
&\sum_{S'\neq S_0} \exp\left\{ \log \mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_{S'}) \right\} \\
\leq \;& 2^p \max_{S'\in\mathcal{S}\setminus S_0} \exp\left\{ \log \mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_{S'}) \right\} \\
\leq \;& 2^p \exp\left(\max_{S'\in\mathcal{S}\setminus S_0} \left\{ \log \mathrm{BF}(\mathcal{M}_{S'}, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_{S'}) \right\} \right) \\
\overset{\text{(D.8)}}{=} \;& 2^p \exp\left\{ \log \mathrm{BF}(\mathcal{M}_{\tilde{S}}, \mathcal{M}_{S_0}; y^n) + \log f(\mathcal{M}_{\tilde{S}}) \right\}.
\end{aligned}$$

Further define

$$c_3 = \frac{1}{|\mathcal{S}_1|}\exp\Big\{-\frac{1}{|\mathcal{S}_1|}\sum_{S\in\mathcal{S}_1\setminus S_0}\log f(\mathcal{M}_S)\Big\}. \tag{D.9}$$

Here $|\mathcal{S}_1|$ is the number of elements in $\mathcal{S}_1$. Obviously $0 \le c_3 < \infty$. Now with (D.7), (D.8), (D.9), the result of Lemma (D.1) and the assumption that $p \propto n^\alpha$, the tail probability on the right hand side of (D.5) can be bounded from below in a way such that

$$\begin{aligned}
& 1 - \frac{\sum_{S'\neq S_0}\exp[\log \mathrm{BF}(\mathcal{M}_{S'},\mathcal{M}_{S_0};y^n)+\log f(\mathcal{M}_{S'})]}{\sum_{S\in\mathcal{S}}\exp[\log \mathrm{BF}(\mathcal{M}_S,\mathcal{M}_{S_0};y^n)+\log f(\mathcal{M}_S)]} \\
\ge\ & 1 - 2^p c_3 \exp\Big\{-\log\frac{f(y^n|\ \mathcal{M}_{S_0})}{f(y^n|\ \mathcal{M}_{\tilde{S}})} \\
& + \log\frac{f(\mathcal{M}_{\tilde{S}})}{f(\mathcal{M}_{S_0})} + \frac{1}{|\mathcal{S}_1|}\sum_{S\in\mathcal{S}_1}\log\frac{f(y^n|\ \mathcal{M}_{S_0})}{f(y^n|\ \mathcal{M}_S)}\Big\} \\
\ge\ & 1 - c_3\exp\Big\{p\log 2 - \Big[\log\frac{f(y^n|\ \mathcal{M}_{S_0})}{f(y^n|\ \mathcal{M}_{\tilde{S}})} - \log\frac{f(\mathcal{M}_{\tilde{S}})}{f(\mathcal{M}_{S_0})}\Big] + \frac{n\xi}{2}\Big\} \\
\ge\ & 1 - c_3\exp\Big(c_{11}n^\alpha\log 2 - \frac{n\xi}{2}\Big), \\
=\ & 1 - c_3\exp\Big\{-\frac{n^\alpha}{2}(n^{1-\alpha}\xi - c_{11}\log 4)\Big\},
\end{aligned} \tag{D.10}$$

where c_{11} is a finite positive constant. Now by (D.10), we can see that for $0 < \alpha < 1$, the probability $\mathbb{P}(\mathcal{M}_{S_0}|\ y^n) \to 1$ as $n \to \infty$, therefore the proof is completed. □

APPENDIX E: PROOF OF THEOREM 5.3

Proof of Theorem 5.3. First define the ridge estimator $\widehat{\beta}^{\text{ridge}}$ by

$$\widehat{\beta}^{\text{ridge}} = \arg\min \left\{ ||y - X\beta||_2^2 + \lambda ||\beta||_2^2 \right\}$$

Note that the ridge estimator has a closed form solution that $\widehat{\beta}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T y$. The definitions of $\widehat{\beta}^{\tau_3}$ and $\widehat{\beta}^{\text{ridge}}$ imply that

$$\begin{aligned} ||y - X\widehat{\beta}^{\tau_3}||_2^2 + \lambda ||\widehat{\beta}^{\tau_3}||_2^2 \quad - \quad & ||y - X\widehat{\beta}^{\text{ridge}}||_2^2 - \lambda ||\widehat{\beta}^{\text{ridge}}||_2^2 \\ = \quad & (\widehat{\beta}^{\tau_3})^T (X^T X + \lambda I_p) \widehat{\beta}^{\tau_3} \\ & - (\widehat{\beta}^{\text{ridge}})^T (X^T X + \lambda I_p) \widehat{\beta}^{\text{ridge}} \\ & - 2y^T X \widehat{\beta}^{\tau_3} + 2y^T X \widehat{\beta}^{\text{ridge}}. \end{aligned} \tag{E.1}$$

The last two terms on the right hand side of (E.1) can be rearranged as

$$\begin{aligned} - 2y^T X \widehat{\beta}^{\tau_3} \quad + \quad & 2y^T X \widehat{\beta}^{\text{ridge}} \\ = \quad & -2(\widehat{\beta}^{\text{ridge}})^T (X^T X + \lambda I_p) \widehat{\beta}^{\tau_3} \\ & + 2(\widehat{\beta}^{\text{ridge}})^T (X^T X + \lambda I_p) \widehat{\beta}^{\text{rigde}}. \end{aligned}$$

Therefore

$$\begin{aligned} ||y - X\widehat{\beta}^{\tau_3}||_2^2 + \lambda ||\widehat{\beta}^{\tau_3}||_2^2 \quad - \quad & ||y - X\widehat{\beta}^{\text{ridge}}||_2^2 - \lambda ||\widehat{\beta}^{\text{ridge}}||_2^2 \\ = \quad & (\widehat{\beta}^{\tau_3} - \widehat{\beta}^{\text{ridge}})^T (X^T X + \lambda I_p) (\widehat{\beta}^{\tau_3} - \widehat{\beta}^{\text{ridge}}), \end{aligned}$$

and in turn,

$$
\begin{aligned}
& (\Lambda_{\min}(X^TX)+\lambda)||\widehat{\beta}^{\tau_3}-\widehat{\beta}^{\text{ridge}}||_2^2 \\
\leq\ & (\widehat{\beta}^{\tau_3}-\widehat{\beta}^{\text{ridge}})^T(X^TX+\lambda I_p)(\widehat{\beta}^{\tau_3}-\widehat{\beta}^{\text{ridge}}) \\
=\ & ||y-X\widehat{\beta}^{\tau_3}||_2^2+\lambda||\widehat{\beta}^{\tau_3}||_2^2 \\
& -||y-X\widehat{\beta}^{\text{ridge}}||_2^2-\lambda||\widehat{\beta}^{\text{ridge}}||_2^2 \\
\leq\ & \rho_1\sum_{j=1}^{p}\log\left(\frac{1+\tau_3^{-1}|\widehat{\beta}^{\text{ridge}}|}{1+\tau_3^{-1}|\widehat{\beta}^{\tau_3}|}\right), \qquad \text{(E.2)}
\end{aligned}
$$

by definition of $\widehat{\beta}^{\tau_3}$, where $\rho_1 = \rho[\log(1+\tau_3^{-1})]^{-1}$. Rearranging (E.2) we have

$$
||\widehat{\beta}^{\tau_3}-\widehat{\beta}^{\text{ridge}}||_2^2 \leq \frac{\rho_1}{\Lambda_{\min}(X^TX)+\lambda}\sum_{j=1}^{p}\log\left(\frac{1+\tau_3^{-1}|\widehat{\beta}_j^{\text{ridge}}|}{1+\tau_3^{-1}|\widehat{\beta}_j^{\tau_3}|}\right). \qquad \text{(E.3)}
$$

Let $\widehat{S}$ denote the estimated index set of $\widehat{\beta}_j^{\tau_3} \neq 0$. By using the inequality $\log\theta \leq \theta - 1$ for $\theta > 0$, we can bound the right hand side of (E.3) in a way such that

$$
\begin{aligned}
\rho_1\sum_{j=1}^{p}\log\left(\frac{1+\tau_3^{-1}|\widehat{\beta}_j^{\text{ridge}}|}{1+\tau_3^{-1}|\widehat{\beta}_j^{\tau_3}|}\right) \leq\ & \rho_1\sum_{j\in\widehat{S}}\left(\frac{\tau_3+|\widehat{\beta}_j^{\text{ridge}}|}{\tau_3+|\widehat{\beta}_j^{\tau_3}|}-1\right) \\
& +\rho\sum_{j\in\widehat{S}^c}\frac{\log(1+\tau_3^{-1}|\widehat{\beta}_j^{\text{ridge}}|)}{\log(1+\tau_3^{-1})} \\
=\ & \rho\sum_{j\in\widehat{S}}\frac{|\widehat{\beta}_j^{\text{ridge}}|-|\widehat{\beta}_j^{\tau_3}|}{(\tau_3+|\widehat{\beta}^{\tau_3}|)\log(1+\tau_3^{-1})} \\
& +\rho\sum_{j\in\widehat{S}^c}\frac{\log(1+\tau_3^{-1}|\widehat{\beta}_j^{\text{ridge}}|)}{\log(1+\tau_3^{-1})}.
\end{aligned}
$$

(E.4)

Note that the second term on the right hand side of (E.4) approaches to $\rho|\widehat{S}^c|$ as $\tau_3 \to 0$ with $\widehat{\beta}_j^{\text{ridge}} \neq 0$ for $j \in \widehat{S}^c$. In addition, the term

$$\frac{1}{(\tau_3 + |\widehat{\beta}^{\tau_3}|)\log(1+\tau_3^{-1})} \to 0$$

as $\tau_3 \to 0$ given that $\widehat{\beta}_j^{\tau_3} \neq 0$ for $j \in \widehat{S}$. Therefore

$$\lim_{\tau_3 \to 0} \rho_1 \sum_{j=1}^{p} \log\left(\frac{1+\tau_3^{-1}|\widehat{\beta}_j^{\text{ridge}}|}{1+\tau_3^{-1}|\widehat{\beta}_j^{\tau_3}|}\right) \quad \leq \quad \rho|\widehat{S}^c|. \tag{E.5}$$

By using the inequality (E.5) with $\widehat{\beta}_{\text{BMIO}} = \lim_{\tau_3 \to 0} \widehat{\beta}^{\tau 3}$ and $|\widehat{S}^c| \leq p$, we have

$$||\widehat{\beta}^{\text{ridge}} - \widehat{\beta}_{\text{BMIO}}||_2^2 \leq \frac{\rho p}{\Lambda_{\min}(X^T X) + \lambda} \tag{E.6}$$

In addition, as shown in the Theorem 1 of Zou and Zhang [11], the quantity $||\widehat{\beta}^{\text{ridge}} - \beta_0||_2^2$ can be bounded by

$$||\widehat{\beta}^{\text{ridge}} - \beta_0||_2^2 \leq \frac{2\lambda^2||\beta_0||_2 + 2p\Lambda_{\max}(X^T X)\sigma^2}{(\Lambda_{\min}(X^T X) + \lambda)^2} \tag{E.7}$$

By using the results in (E.6), (E.7) and Assumption 1, we can bound the quantity $\mathbb{E}_Y[||\widehat{\beta}_{\text{BMIO}} - \beta_0||_2^2]$ in a way such that

$$\begin{aligned}
& \mathbb{E}_Y[||\widehat{\beta}_{\text{BMIO}} - \beta_0||_2^2] \\
\leq \; & 2\mathbb{E}_Y[||\widehat{\beta}_{\text{BMIO}} - \widehat{\beta}^{\text{ridge}}||_2^2] + 2\mathbb{E}_Y[||\widehat{\beta}^{\text{ridge}} - \beta_0||_2^2] \\
\leq \; & \frac{2\rho p(\Lambda_{\min}(X^T X) + \lambda)}{(\Lambda_{\min}(X^T X) + \lambda)^2} + \frac{4\lambda^2||\beta_0||_2^2 + 4p\Lambda_{\max}(X^T X)\sigma^2}{(\Lambda_{\min}(X^T X) + \lambda)^2} \\
\leq \; & \frac{2\rho p n c_2 + 4\lambda^2||\beta_0||_2^2 + 4pnc_2\sigma^2}{(nc_1 + \lambda)^2} \\
\leq \; & \frac{2\rho p c_2 + 4n^{-1}\lambda^2||\beta_0||_2^2 + 4pc_2\sigma^2}{nc_1^2}.
\end{aligned} \tag{E.8}$$

Further by using Markov's inequality, the probability $\mathbb{P}(||\widehat{\beta}_{\text{BMIO}}-\beta_0||_2^2 > \xi_n)$ can be bounded by

$$\begin{aligned}\mathbb{P}\Big(||\widehat{\beta}_{\text{BMIO}}-\beta_0||_2^2 > \xi_n\Big) &\leq \frac{2\rho p c_2 + 4n^{-1}\lambda^2||\beta_0||_2^2 + 4pc_2\sigma^2}{nc_1^2\xi_n} \\ &\leq \frac{2p\rho c_2 + 4n^{-1}\lambda^2 pc_4 + 4pc_2c_5}{nc_1^2\xi_n} \\ &\leq \frac{2p\rho c_2 + 4\lambda^2 pc_4 + 4pc_2c_5}{nc_1^2\xi_n} \qquad \text{(E.9)}\end{aligned}$$

for $n > 1$. We let $c_7 = (2\rho c_2 + 4\lambda^2 c_4 + 4c_2c_5)/c_1^2$, then with the assumption that $p \propto n^{\alpha}$, the right hand side of (E.9) becomes $c_7 p(\xi_n n)^{-1} = c_7 c_{12} n^{\alpha-1}\xi_n^{-1}$, where c_{12} is a positive finite constant. Then (E.9) can be re-expressed as

$$\mathbb{P}\Big(||\widehat{\beta}_{\text{BMIO}}-\beta_0||_2^2 > \xi_n\Big) \leq c_{13}\exp\{-\log(n^{1-\alpha}\xi_n)\}, \qquad \text{(E.10)}$$

where $c_{13} = c_7c_{12}$. If $\xi_n \propto n^{-\alpha^*}$, then $n^{1-\alpha}\xi_n \propto n^{1-(\alpha+\alpha^*)}$. Then under the condition $0 < \alpha^* < \alpha < 1/2$, the term $n^{1-(\alpha+\alpha_*)} \to \infty$ as $n \to \infty$. Therefore if $0 < \alpha^* < \alpha < 1/2$, the right hand side of (E.10) will approach to 0, which completes the proof. □

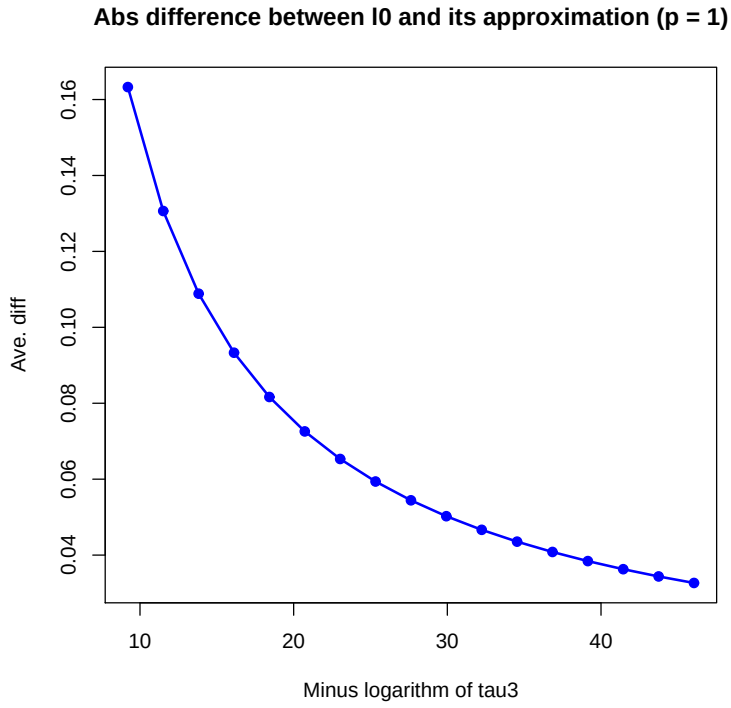

FIG 1. *The absolute difference between the l_0 norm and its log-sum approximation.*

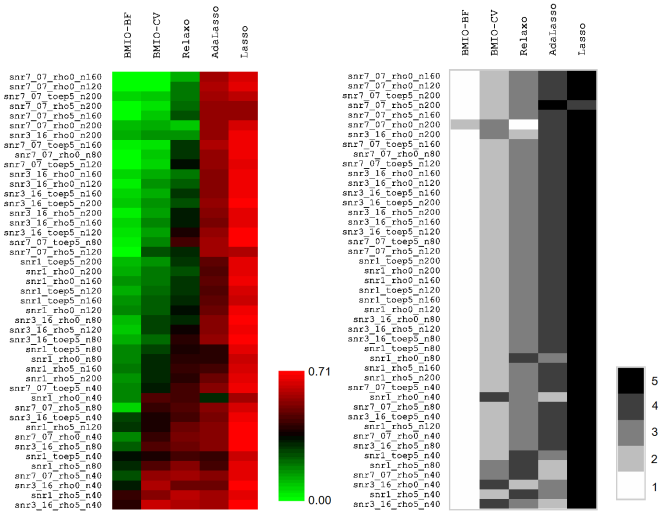

FIG 2. *Heapmaps based on GAP (the generalized association plots) for the S-FPR (left) and rankings of the S-FPR (right) of the five estimation approaches under the 45 simulation scenarios.*

The heatmaps are generated by using the graphical software GAP (Generalized Associated Plots), which was developed by Wu, Tien and Chen [9] as companion software to [2]. The GAP-based heatmaps further suggest that using BAVA-MIO estimation can lead to more accurate variable selection.

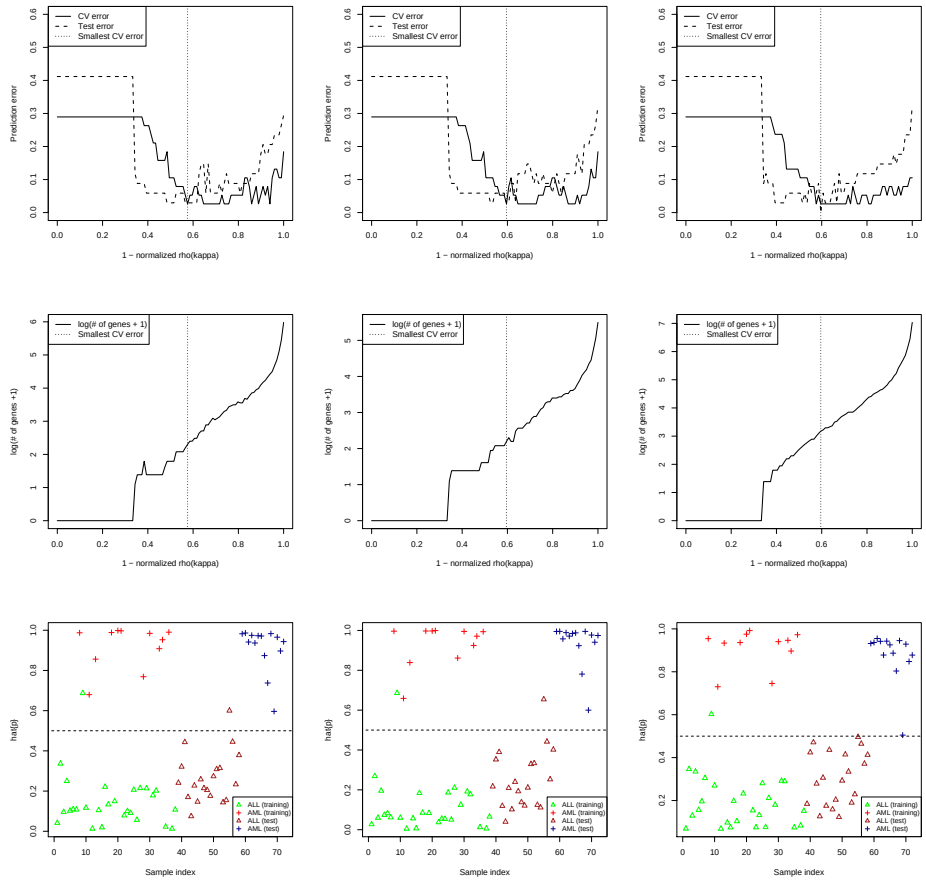

FIG 3. *Results of the BAVA-MIO estimation for Golub's gene expression data. The training and test errors, and the number of included variables at logarithm scale along the regularization path under the three BAVA-MIO estimations are shown in the first panel and second panel, respectively, while the estimated label probabilities for the total 72 patients under the three BAVA-MIO estimations are shown in the bottom panel. Left:* $\lambda^* = 0.05$*; Center* $\lambda^* = 0.1$*; Right* $\lambda^* = 0.5$*. Top: the CV error and test error against the tuning parameter; Middle: logarithm of the number of selected genes against the tuning parameter; Bottom: scatter plot for estimated label probabilities. The vertical dash line in each plot in the top two panels indicates where the tuning parameter is selected.*

REFERENCES

[1] Candés, E. J., Wakin, M. B. and Boyd, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications* **14** 877-905.

[2] Chen, C.-H. (2002). Generalized association plots: information visualization via iteratively generated correlation matrices. *Statistica Sinica* **12** 7-29.

[3] Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory* **52** 6-18.

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

[5] Fuchs, J. J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transcations on Information Theory* **51** 3601-3608.

[6] Sriperumbudur, B. K., Torres, D. A. and Lanckriet, G. R. G. (2009). A D.C. programming approach to the sparse generalized eigenvalue problem. *http://arxiv.org/abs/0901.1504*.

[7] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1** 211-244.

[8] Tropp, J. A. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory* **52** 1030-1051.

[9] Wu, H.-M., Tien, Y.-J. and Chen, C.-H. (2010). GAP: a graphical environment for matrix visualization and cluster analysis. *Computational Statistics and Data Analysis* **54** 767-778.

[10] Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology) **69** 143-161.

[11] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37** 1733-1751.