Web-based Supplementary Materials for "Model Confidence Bounds for Variable Selection"

Yang Li[1,2], Yuetian Luo[3], Davide Ferrari[4], Xiaonan Hu[5,6] and Yichen Qin[7,*]

[1]School of Statistics, Renmin University of China

[2]Center for Applied Statistics, Renmin University of China

[3]Department of Statistics, University of Wisconsin, Madison

[4]Faculty of Economics and Management, Free University of Bozen-Bolzano

[5]Department of Biostatistics, Yale University

[6]School of Mathematical Sciences, University of Chinese Academy of Sciences

[7]Department of Operations, Business Analytics, and Information Systems, University of Cincinnati

[*]*email:* yichen.qin@uc.edu

1 Web Appendix A: Additional Materials

1.1 Connection Between Algorithm 1 and Algorithm 2

We further conduct a toy example to graphically illustrate the connection between Algorithm 1 and Algorithm 2 in Web Figure 1. We simulate a data set (sample size $n = 150$) with ten

predictors and one response variable by $y_i = \sum_{j=1}^{3} \theta_j x_{i,j} + \epsilon_i$ with $\epsilon_i \sim N(0, 3)$. Using such a data set, we generate $B = 100$ bootstrap models. Then Algorithm 1 returns the MCB with LBM being X_1 and X_2 and UBM being X_1, X_2, X_3, X_7, X_{10}, X_4, and X_6.

We plot all the bootstrap models along with MCB given by Algorithm 1 in Web Figure 1. In the figure, each column represents one predictor and the order from left to right is based on the descending order of the frequency of each predictor in bootstrap models, $\overline{\pi}_j$, and their frequencies are denoted below the predictor names. Each row represents one bootstrap models. Each box (intersection of the column and row) represents one predictor in one bootstrap model with the color black/white indicating whether the predictor is selected/not selected in the bootstrap model.

All bootstrap models are divided into three groups which are separated by light blue lines. The top group (marked by vertical green bar) is regarded as underfitting models since these models miss some predictors in LBM. The middle group (purple bar) is regarded as models nested inside MCB. The bottom group (brown bar) is regarded as overfitting models since models contain some additional superfluous predictors apart from predictors in UBM. The order of bootstrap models in each group is based on the model ranking method introduced in the proof of Theorem 2. The horizontal red bar on the top which contains X_1 and X_2 represents the LBM. The horizontal blue bar which contains X_3, X_7, X_{10}, X_4, X_6 represent predictors that are in UBM but not in LBM (i.e., UBM-LBM).

In the figure, since the predictors are ordered by their selection frequencies (from left to right), we can see that the LBM by Algorithm 1 (red bar) is composed by the two most frequently selected predictors. UBM by Algorithm 1 (red bar + blue bar) is composed by the seven most frequently selected predictors. Therefore, it is obvious that both LBM and UBM follows the pattern/constrain imposed by $\widehat{m}_L(k) = \{u_1, \ldots, u_k\}$, $\widehat{m}_U(k, w) = \{u_1, \ldots, u_k, u_{k+1}, \ldots, u_{k+w}\}$ in Algorithm 2, which means Algorithm 2 will generate the exactly same MCB as Algorithm 1 in this case. Note that Algorithm 1 searches exhaustively

for all possible choices of LBM and UBM, whereas Algorithm 2 only searches within a much smaller set of models. However, since the MCB by Algorithm 1 already belongs to this smaller set in which Algorithm 2 searches, Algorithm 2 will guarantee to return the exactly same MCB as Algorithm 1. Therefore, Algorithms 1 and 2 are consistent with each other. This is generally true as long as predictors are not seriously correlated.

1.2 Additional Real Data Example

We present one additional real data example using a lung squamous cell carcinoma data set (Cancer Genome Atlas Research Network, 2012) to illustrate the proposed MCB. The dataset contains $n = 404$ patients with gene expressions available. The response variable is the logarithm of prognosis time. Among all patients, 129 of them died during follow-up, and their median follow-up time is 30 months. The rest 275 patients are censored, and their median follow-up time is 18 months. To reduce computational cost, we use $p = 100$ gene expressions which cover the top-ranked genes followed by the results in Chai et al. (2017). We adopt Algorithm 2 to construct the MCBs based on $B = 300$ bootstrap models, and apply the following model selection methods: Adaptive Lasso, Lasso, MCP, SCAD, and screening (which is based on the univariate linear regressions with p-values less than or equal to 0.05).

In Web Figure 2, we compare different model selection methods using the MUC curve. It can be seen that the CR increases with the MCBs width. When $w = 88$ and $w/p = 0.88$, the CR of MCP shows the corresponding MCB captures about 60% of the bootstrap models. Based on the shape of the MUC curve, we can evaluate the uncertainty of different methods, and conclude that screening and MCP are more stable than others for this dataset. In particular, for higher confidence levels, say 75% or above, screening has the smallest the width among all methods. We further visualize the MCBs returned by screening at confidence levels of 75% and 95% in Web Figure 3, and also report the single selected models based on the original dataset. As we can see, both 75%- and 95%- MCBs contain the single selected

model. As the confidence levels increases, the MCB becomes larger as well. Because of the weak signal in the data set, the LBMs are small and UBM are big. In fact, the UBM of 95% is the full model.

1.3 Bootstrap Algorithms

In the simulation studies, we have applied residual bootstrap for Adaptive Lasso, SCAD, MCP, LAD, and stepwise regression with BIC, and applied modified residual bootstrap for Lasso. The summary of bootstrap algorithms is below.

We first illustrate the residual bootstrap using Adaptive Lasso as an example. Let $\widehat{\boldsymbol{\theta}}$ be the Adaptive Lasso estimate of $\boldsymbol{\theta}$. Define the residual as $e_i = y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}$ for $i = 1, ..., n$. After centering the residuals as $e_i^* = e_i - \overline{e}_i$ where $\overline{e}_i = \sum_{i=1}^n e_i/n$, we sample from $\{e_i^*\}_{i=1}^n$ with replacement to form $\{e_i^{(b)}\}_{i=1}^n$, and generate the response variable as $y_i^{(b)} = \boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}} + e_i^{(b)}$. Using $\{y_i^{(b)}, \boldsymbol{x}_i\}_{i=1}^n$, we obtain the bootstrap estimate $\widehat{\boldsymbol{\theta}}^{(b)}$. We repeat the above procedure for B times and obtain a sequence of bootstrap estimates $\widehat{\boldsymbol{\theta}}^{(b)}$ for $b = 1, ..., B$.

We next illustrate the modified residual bootstrap using Lasso as an example. Let $\widehat{\boldsymbol{\theta}}$ be the Lasso estimate of $\boldsymbol{\theta}$. We first shrink $\widehat{\boldsymbol{\theta}}$ further toward zero to obtain $\widetilde{\boldsymbol{\theta}}$ by $\widetilde{\theta}_j = \widehat{\theta}_j I\{|\widehat{\theta}_j| > a\}$ for $j = 1, ..., p$ with $a = cn^{-\delta}$, $c \in (0, \infty)$, and $\delta \in (0, 1/2)$ (Chatterjee and Lahiri, 2011). We further calculate the residual using $\widetilde{\boldsymbol{\theta}}$ as $e_i = y_i - \boldsymbol{x}_i^T\widetilde{\boldsymbol{\theta}}$ for $i = 1, ..., n$. After centering the residuals as $e_i^* = e_i - \overline{e}_i$ where $\overline{e}_i = \sum_{i=1}^n e_i/n$, we sample from $\{e_i^*\}_{i=1}^n$ with replacement to form $\{e_i^{(b)}\}_{i=1}^n$, and generate the response variable as $y_i^{(b)} = \boldsymbol{x}_i^T\widetilde{\boldsymbol{\theta}} + e_i^{(b)}$. Using $\{y_i^{(b)}, \boldsymbol{x}_i\}_{i=1}^n$, we obtain the bootstrap estimate $\widehat{\boldsymbol{\theta}}^{(b)}$. We repeat the above procedure for B times and obtain a sequence of bootstrap estimates $\widehat{\boldsymbol{\theta}}^{(b)}$ for $b = 1, ..., B$.

1.4 Comparison of Algorithms

Using linear regression model, we simulate data according the six scenarios introduced in the manuscript under independent and correlated covariates. We present these MUCs of the correlated covariates in Web Figures 4. As we can see, Algorithms 1 and 2 perform approximately the same, verifying the validity of Algorithm 2. Since Algorithm 1 achieves the highest coverage rate possible, we conclude that Algorithm 2 performs (nearly) optimally as well.

1.5 Performance of MCB as $n \to \infty$

Using the linear regression model, we consider two scenarios: (a) $p = 15$, $p^* = 6$; (b) $p = 50$, $p^* = 8$. We set $B = 1000$ and $\sigma = 1$, and set $\rho = 0$ to simulate the cases of independent covariates. We use 10-fold cross-validated Adaptive Lasso as the model selection method. We increase the sample size n from 200 to 500 and explore the performance of MCB using MUC in Web Figures 5. In these figures, the MUC arches further towards the upper left corner with the increasing n. It makes sense because as n increases, there are fewer unique bootstrap models and the variation of these bootstrap models becomes less. Therefore, the MCB of the same width will be able to cover more bootstrap models and have a higher CR. Hence the MUC arches further towards the upper left corner. In the extreme case of $n \to \infty$, the Adaptive Lasso will always select the true model according to the oracle property. Therefore, all the bootstrap models will be the same, and MCB will contain only the true model. Such a phenomenon persists even in the case of correlated covariates. We consider three scenarios: (a) $\rho = 0.25$; (b) $\rho = 0.5$; (c) $\rho = 0.75$. The results are shown in Web Figure 6, which is consistent with our previous observation.

1.6 Comparison of CR and True CR

Many comparisons and calculation throughout the manuscript are based on the assumption that CR statistic (3) closely approximates the true model coverage rate TCR := $P(\widehat{m}_L \subseteq m^* \subseteq \widehat{m}_U)$. In this section, we will verify such a relationship between CR and TCR. Due the close relationship between MUC and CR (i.e., $\mathscr{P}_{\text{MUC}} = \{(w/p, \text{CR}(w)),\ 0 \leq w \leq p\}$), we can compare MUC with the true model uncertainty curve (TMUC, $\mathscr{P}_{\text{TMUC}} = \{(w/p, \text{TCR}(w)),\ 0 \leq w \leq p\}$) to assess the approximation. First, we consider the same linear regression model in three scenarios: (a) $p = 8$, $p^* = 3$, $n = 1000$, $\sigma = 10$; (b) $p = 20$, $p^* = 8$, $n = 1000$, $\sigma = 10$; and (c) $p = 20$, $p^* = 8$, $n = 300$, $\sigma = 6$. We set $B = 1000$. Web Figure 7 shows the comparison of MUC and TMUC. We can see that MUC closely approximates TMUC at different widths in all three scenarios. Next, we consider another four scenarios for correlated covariates where $\boldsymbol{x}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = [\Sigma_{ij}]_{p\times p}$ and $\Sigma_{ij} = \rho^{|i-j|}$: (a) $\rho = 0$; (b) $\rho = 0.25$; (c) $\rho = 0.5$; and (d) $\rho = 0.75$. We further set $p = 10$, $p^* = 5$, $\theta_j = 1$, $n = 100$, $\sigma = 6$. The results are shown in Web Figure 8, which are very similar to the previous study. Thus, we conclude that CR and MUC are good approximations for TCR and TMUC.

1.7 Integrated Score of MCB

As shown in the manuscript, to evaluate MCB, coverage probability and width are two separate criteria that are sometimes conflicting. Motivated by Gneiting and Raftery (2007), we propose an integrated score of MCB by combining coverage probability and width together as follows

$$s(\widehat{m}_L, \widehat{m}_U; m^*) = \underbrace{|\widehat{m}_U| - |\widehat{m}_L|}_{\text{MCB width}} + C_L \underbrace{\left|\widehat{m}_L \setminus (\widehat{m}_L \cap m^*)\right|}_{\substack{D_L\text{: number of} \\ \text{unimportant variables} \\ \text{included in LBM}}} + C_U \underbrace{\left|m^* \setminus (\widehat{m}_U \cap m^*)\right|}_{\substack{D_U\text{: number of} \\ \text{important variables} \\ \text{missing in UBM}}},$$

where m^*, $\widehat{m}_L$, and $\widehat{m}_U$ represent the true model, LBM, and UBM, respectively. We let $D_L = |\widehat{m}_L \setminus (\widehat{m}_L \cap m^*)|$ and $D_U = |m^* \setminus (\widehat{m}_U \cap m^*)|$. It is straightforward to see that D_L represents the number of unimportant variables (i.e. variables not in m^*) that are mistakenly included in LBM and D_U represents the number of important variables (i.e. variables in m^*) that are incorrectly missed out in UBM. The parameters C_L and C_U represent the penalty of including one more unimportant variable in LBM and missing one more important variable in UBM. Therefore, the last two terms in the score are the penalization for not capturing the true model in the MCB. Note that if $D_L = D_U = 0$, then the true model is nested between LBM and UBM. We can use such a criterion to evaluate different MCBs. The lower the score, the better the MCB. Note that to apply this score, we need to have m^* available. We illustrate the usage of this score through the simulation studies.

We simulate data according to the linear model with $n = 100$, $B = 1000$, $p^* = 5$, $p = 10$, $\theta_j = 1$, and $\sigma = 6$, and set $\rho = 0$, 0.25, 0.5, and 0.75 for the cases of independent and correlated covariates. We adopt Adaptive Lasso to construct MCB and present the average integrated scores, widths, D_L, D_U, cardinalities, and the coverage rates of MCBs at different confidence levels in Web Table 1. To calculate the score, we set $C_L = C_U = 5$.

As we can see, as the confidence level increases, the average width of MCB increases as well, so does the average cardinality. This is sensible because MCB needs to include more models and have a larger width in order to capture the true model with a higher probability. Meanwhile, as confidence level increases, D_L and D_U gradually decrease as it is less likely for MCB to miss the true model. Since the integrated score is the function of D_L, D_U, and width, the average score first decreases and then increases, leading to the MCB with lowest score occurring at around confidence level 70%. Lastly, we can observe that the true coverage rate is close to, and sometimes slightly higher than, the confidence level across the different settings.

2 Web Appendix B: Proofs

Proof of Theorem 1. The result follows by a straightforward application of the law of total probability. Let $\mathcal{A} = \{\widehat{m}_L \subseteq m^* \subseteq \widehat{m}_U\}$ be the event that the true model m^* is nested between the lower and upper bound models. Then

$$\begin{aligned} P(\mathcal{A}^c) &= P(\mathcal{A}^c|\widehat{m} = m^*)P(\widehat{m} = m^*) + P(\mathcal{A}^c|\widehat{m} \neq m^*)P(\widehat{m} \neq m^*) \\ &\leq P(\{\widehat{m}_L \subseteq \widehat{m}\}^c \cup \{\widehat{m} \subseteq \widehat{m}_U\}^c) + o(1) \\ &\leq P(\widehat{m}_L \supsetneq \widehat{m}) + P(\widehat{m}_U \not\supseteq \widehat{m}) + o(1), \end{aligned}$$

where the probabilities in the last expression concern the event that $\widehat{m}_L$ overestimates m^* and $\widehat{m}_U$ underestimates $\widehat{m}_L$. Let $\widehat{m}^{(b)}$ denote a bootstrap model and note that

$$\begin{aligned} P(\widehat{m}_L \supsetneq \widehat{m}) &= P(\widehat{m}_L \supsetneq \widehat{m}|\widehat{m}^{(b)} \not\subseteq \widehat{m}_L)P(\widehat{m}^{(b)} \not\subseteq \widehat{m}_L) \\ &\quad + P(\widehat{m}_L \supsetneq \widehat{m}|\widehat{m}^{(b)} \subseteq \widehat{m}_L)P(\widehat{m}^{(b)} \subseteq \widehat{m}_L) \\ &\leq P(\widehat{m}^{(b)} \not\subseteq \widehat{m}_L) + P(\widehat{m}^{(b)} \supsetneq \widehat{m}). \end{aligned} \tag{1}$$

Similarly, $P(\widehat{m}_U \not\supseteq \widehat{m}) \leq P(\widehat{m}^{(b)} \supsetneq \widehat{m}_U) + P(\widehat{m}^{(b)} \not\supseteq \widehat{m})$. Combining this with (1) implies

$$\begin{aligned} P(\mathcal{A}^c) &\leq P(\widehat{m}^{(b)} \not\subseteq \widehat{m}_L) + P(\widehat{m}^{(b)} \supsetneq \widehat{m}_U) + P(\widehat{m}^{(b)} \supsetneq \widehat{m}) + P(\widehat{m}^{(b)} \not\supseteq \widehat{m}) + o(1) \\ &\leq \alpha + 2P(\widehat{m}^{(b)} \neq \widehat{m}) + o(1). \end{aligned}$$

The desired result follows from Assumption A.2 □

Proof of Theorem 2. Denote the MCB by Algorithm 1 as $\{\widehat{m}_{L,1}, \widehat{m}_{U,1}\}$, and the MCB by Algorithm 2 as $\{\widehat{m}_{L,2}, \widehat{m}_{U,2}\}$, which have the same width of w. Since Algorithm 1 exhaustively searches for all the possible MCB of width w to maximize the CR, it returns the MCB with

the highest CR. Hence,

$$\widehat{r}(\widehat{m}_{L,1}, \widehat{m}_{U,1}) \geq \widehat{r}(\widehat{m}_{L,2}, \widehat{m}_{U,2}). \tag{2}$$

For the MCB by Algorithm 1, suppose the length of its LBM is a (the length of the LBM is the number of variables that LBM contains), and suppose the MCB width is w. Then there are totally $\binom{p}{a}\binom{p-a}{w}$ possible MCB (with the same LBM length and MCB width) that could be generated by Algorithm 1. They can be denoted as a sequence of MCB,

$$\left\{\mathcal{G}^{(i)} : \mathcal{G}^{(i)} = \{\widehat{m}_L^i(a,w), \widehat{m}_U^i(a,w)\}, \quad i = 1, \cdots, \binom{p}{a}\binom{p-a}{w}\right\},$$

where $\widehat{m}_L^i(a,w)$ and $\widehat{m}_U^i(a,w)$ denote all possible MCB with the LBM length a and width w. Define the CR of $\mathcal{G}^{(i)}$ as

$$\widehat{r}(\mathcal{G}^{(i)}) = \frac{\sum_{b=1}^{B} I(\widehat{m}_L^i(a,w) \subseteq \widehat{m}^{(b)} \subseteq \widehat{m}_U^i(a,w))}{B}.$$

Due to the nature of Algorithm 1, we have

$$\widehat{r}(\widehat{m}_{L,1}, \widehat{m}_{U,1}) = \max_i \{\widehat{r}(\mathcal{G}^{(i)})\}. \tag{3}$$

Let $\Pi = (u_1, \ldots, u_p)$ be the arrangement of indices $\{1, \ldots, p\}$ induced by the ordered selection frequencies $\overline{\pi}_{u_1} \geq \cdots \geq \overline{\pi}_{u_p}$. The ordering Π induces a natural ranking of predictors in terms of their selection frequency. For the MCB returned by Algorithm 2, it maximizes the CR among all MCB with width w and with the following conditions,

$$\widehat{m}_{L,2} = \{u_1, \cdots, u_l\}, \quad \widehat{m}_{U,2} = \{u_1, \cdots, u_l, \cdots, u_{l+w}\}.$$

Suppose the MCB by Algorithm 2 has LBM length of a^* and width of w, hence can be

expressed as

$$\widehat{m}_{L,2} = \{u_1, \cdots, u_{a^*}\}, \quad \widehat{m}_{U,2} = \{u_1, \cdots, u_{a^*+w}\}.$$

where $a^* = \arg\max_{0 \le l \le p-w} \widehat{r}(\{u_1, \cdots, u_l\}, \{u_1, \cdots, u_{l+w}\})$.

We further define

$$\widehat{m}^a_{L,2} = \{u_1, \cdots, u_a\}, \quad \widehat{m}^a_{U,2} = \{u_1, \cdots, u_{a+w}\}. \tag{4}$$

Due to the nature of Algorithm 2, we have

$$\widehat{r}(\widehat{m}_{L,2}, \widehat{m}_{U,2}) \ge \widehat{r}(\widehat{m}^a_{L,2}, \widehat{m}^a_{U,2}). \tag{5}$$

Next, we prove the following lemma:

Lemma 1. *For any* $i = 1, \cdots, \binom{p}{a}\binom{p-a}{w}$, *when* $B \to \infty$, *we have*

$$\widehat{r}(\widehat{m}^a_{L,2}, \widehat{m}^a_{U,2}) \ge \widehat{r}(\mathcal{G}^{(i)}) + o_p(1). \tag{6}$$

Proof of Lemma 1. For any MCB, $\mathcal{G}^{(i)} = \{\widehat{m}^i_L(a,w), \widehat{m}^i_U(a,w)\}$, with LBM length a and width w, there are at most 2^w unique models m^i_k, $k = 1, ..., 2^w$, that satisfy $\widehat{m}^i_L(a,w) \subseteq m^i_k \subseteq \widehat{m}^i_U(a,w)$. Let us denote them as $m^i_1, \cdots, m^i_{2^w}$. Let $\#(m^i_k) = \sum_{b=1}^B I(m_k = \widehat{m}^{(b)})$ be the appearing frequency of model m^i_k in B bootstrap models $\widehat{m}^{(b)}$, $b = 1, ..., B$. Then we have

$$\widehat{r}(\mathcal{G}^{(i)}) = \frac{\sum_{b=1}^B I(\widehat{m}^i_L(a,w) \subseteq \widehat{m}^{(b)} \subseteq \widehat{m}^i_U(a,w))}{B} = \frac{\#(m^i_1) + \cdots + \#(m^i_{2^w})}{B}.$$

Next, we define a order for all possible models, namely Model Rank (MR), according to the following two criteria:

1. The model containing less variables is in front of the model containing more variables.

2. For the models containing the same number of variables, we rank these models according to the sum of the selected frequency of included variables by descending order.

For example, let $\Delta(X_j) = \sum_{b=1}^{B} I(j \in \widehat{m}^{(b)})$ be the selected frequency of variable X_j in the B bootstrap models $\widehat{m}^{(b)}$ and $0 \leq \Delta(X_j) \leq B$. Suppose there are four variables in total, $X_1, \cdots, X_4$, and assume no ties, $\Delta(X_1) > \Delta(X_2) > \Delta(X_3) > \Delta(X_4)$. Then we let model $\{1\}$ be in front of model $\{1,2\}$. And we let model $\{1,2\}$ be in front of model $\{1,3\}$. Note that models $\{1\}$ and $\{1,2\}$ and $\{1,3\}$ mean the models selecting $\{X_1\}$ and $\{X_1, X_2\}$ and $\{X_1, X_3\}$, respectively.

Thus, according to the MR order, these 2^w unique models m_k^i which satisfy $\widehat{m}_L^i(a,w) \subseteq m_k^i \subseteq \widehat{m}_U^i(a,w)$ can be ranked as $m_1^{i,c}, \cdots, m_{2^w}^{i,c}$. Similarly, the 2^w unique models m_k, $k = 1, \cdots, 2^w$ which satisfy $\widehat{m}_{L,2}^a \subseteq m_k \subseteq \widehat{m}_{U,2}^a$ can be ranked as $m_1^d, \cdots, m_{2^w}^d$. It is obvious that $m_1^d = \widehat{m}_{L,2}^a = \{u_1, \cdots, u_a\}$ and $m_{2^w}^d = \widehat{m}_{U,2}^a = \{u_1, \cdots, u_a, \cdots, u_{a+w}\}$.

Let $P(\widehat{m}^{(b)})$ be the probability that model $\widehat{m}^{(b)}$ is generated by the chosen model selection method. Let $P(X_j)$ be the probability that variable X_j is selected by the chosen model selection method. By the bootstrap validity, when $B \to \infty$, we have

$$\begin{aligned} \frac{\#(\widehat{m}^{(b)})}{B} &= P(\widehat{m}^{(b)}) + o_p(1), \\ \frac{\Delta(X_j)}{B} &= P(X_j) + o_p(1). \end{aligned}$$

Meanwhile, when Assumption A.3 is satisfied, we have

$$\frac{\#(\widehat{m}^{(b)})}{B} = \Pi_{j \in \widehat{m}^{(b)}} \frac{\Delta(X_j)}{B} \Pi_{j \notin \widehat{m}^{(b)}} (1 - \frac{\Delta(X_j)}{B}) + o_p(1).$$

For each $k = 1, \cdots, 2^w$, we know that both $m_k^{i,c}$ and m_k^d contain the same number of variables (for any $i = 1, \cdots, \binom{p}{a}\binom{p-a}{w}$). Suppose they contain q_k variables and $a \leq q_k \leq a+w$.

For model $m_k^{i,c}$, the selected q_k variables can be ranked as $X_1^{i,c}, \cdots, X_{q_k}^{i,c}$ according to their selection frequencies by descending order. Similarly, for model m_k^d, the selected q_k variables can be ranked as $X_1^d, \cdots, X_{q_k}^d$. Because of the nature of $\widehat{m}_{L,2}^a$ and $\widehat{m}_{U,2}^a$ (i.e., Equation 4) and the MR order, for any $h = 1, \cdots, q_k$, we have

$$\Delta(X_h^{i,c}) \leq \Delta(X_h^d).$$

For example, assume that there are four variables: $\Delta(X_1) > \Delta(X_2) > \Delta(X_3) > \Delta(X_4)$, and three selected variables: $m_k^{i,c} = \{1, 3, 4\}$, and $m_k^d = \{1, 2, 3\}$. Thus, we have

$$\frac{\#(\widehat{m}_k^{i,c})}{B} = \frac{\Delta(X_1)}{B}\left(1 - \frac{\Delta(X_2)}{B}\right)\frac{\Delta(X_3)}{B}\frac{\Delta(X_4)}{B} + o_p(1),$$

while

$$\frac{\#(\widehat{m}_k^d)}{B} = \frac{\Delta(X_1)}{B}\frac{\Delta(X_2)}{B}\frac{\Delta(X_3)}{B}\left(1 - \frac{\Delta(X_4)}{B}\right) + o_p(1).$$

Since $\Delta(X_1) > \Delta(X_2) > \Delta(X_3) > \Delta(X_4)$, we have $1 - \Delta(X_1)/B < 1 - \Delta(X_2)/B < 1 - \Delta(X_3)/B < 1 - \Delta(X_4)/B$. Therefore, $\#(\widehat{m}_k^{i,c})/B \leq \#(\widehat{m}_k^d)/B + o_p(1)$.

The same argument can be easily generalized for p variables with q_k selected variables, and we have

$$\frac{\#(\widehat{m}_k^{i,c})}{B} \leq \frac{\#(\widehat{m}_k^d)}{B} + o_p(1), \quad \forall k.$$

Furthermore, for $i = 1, \cdots, \binom{p}{a}\binom{p-a}{w}$, since

$$\widehat{r}(\mathcal{G}^{(i)}) = \frac{\#(\widehat{m}_1^{i,c}) + \cdots + \#(\widehat{m}_{2^w}^{i,c})}{B},$$

and

$$\widehat{r}(\widehat{m}_{L,2}^a, \widehat{m}_{U,2}^a) = \frac{\#(\widehat{m}_1^d) + \cdots + \#(\widehat{m}_{2^w}^d)}{B}.$$

Then we have $\widehat{r}(\widehat{m}_{L,2}^a, \widehat{m}_{U,2}^a) \geq \widehat{r}(\mathcal{G}^{(i)}) + o_p(1)$. □

By Equation (3) and Lemma 1, we have

$$\widehat{r}(\widehat{m}^a_{L,2}, \widehat{m}^a_{U,2}) \geq \widehat{r}(\widehat{m}_{L,1}, \widehat{m}_{U,1}) + o_p(1).$$

By Equation (5), we further have

$$\widehat{r}(\widehat{m}_{L,2}, \widehat{m}_{U,2}) \geq \widehat{r}(\widehat{m}_{L,1}, \widehat{m}_{U,1}) + o_p(1). \tag{7}$$

Combining Equations (2) and (7), we have

$$|\widehat{r}(\widehat{m}_{L,1}, \widehat{m}_{U,1}) - \widehat{r}(\widehat{m}_{L,2}, \widehat{m}_{U,2})| = o_p(1).$$

□

References

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525.

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E., et al. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econometrics and Statistics* **4,** 105–120.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association* **106,** 608–625.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

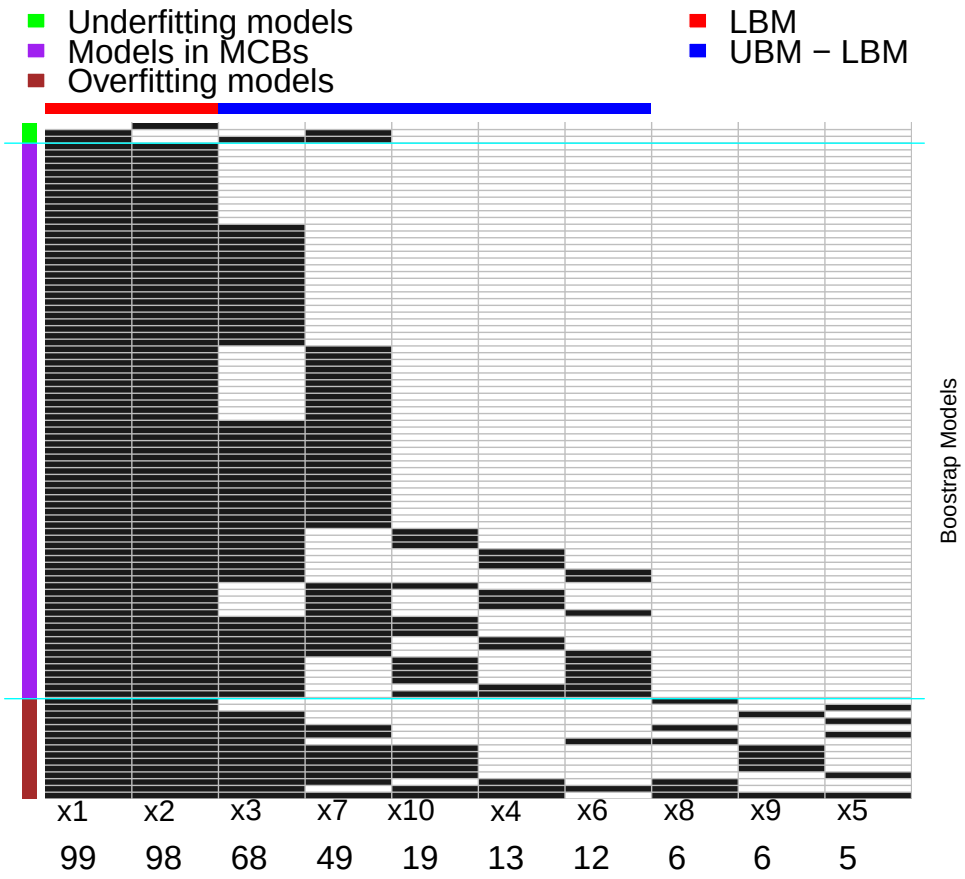


Figure 1: Illustration of forming MCB using Algorithms 1 and comparison to Algorithm 2: Data are generated from the linear regression model $Y \sim N_n(X\boldsymbol{\theta}, 3I_n)$ with $n = 150$, $B = 100$, $\boldsymbol{\theta} = (1, 1, 1, 0, 0, 0, 0, 0, 0, 0)^T$. Each column denotes one predictor and each row represents one bootstrap model. Each box (intersection of the column and row) represents one predictor in one bootstrap model with black/white indicating the predictor is selected/not selected. Bootstrap models are divided into three groups, underfitting (green), inside MCB (purple), overfitting (brown) by two light blue lines. Predictors are divided into LBM (red), UBM-LBM (blue), and others. The Adaptive Lasso is used as the variable selection method.

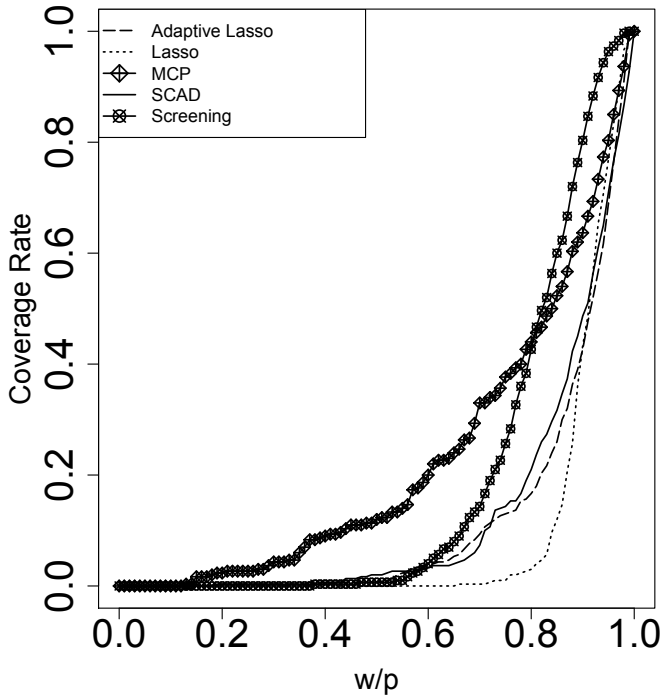


Figure 2: MUC of five model selection methods (Adaptive Lasso, Lasso, MCP, SCAD and Screening) when applying to the lung squamous cell carcinoma data. $B = 300$ bootstrap samples are generated from the original dataset. All of the tunning parameters are chosen by 10-fold cross-validation.

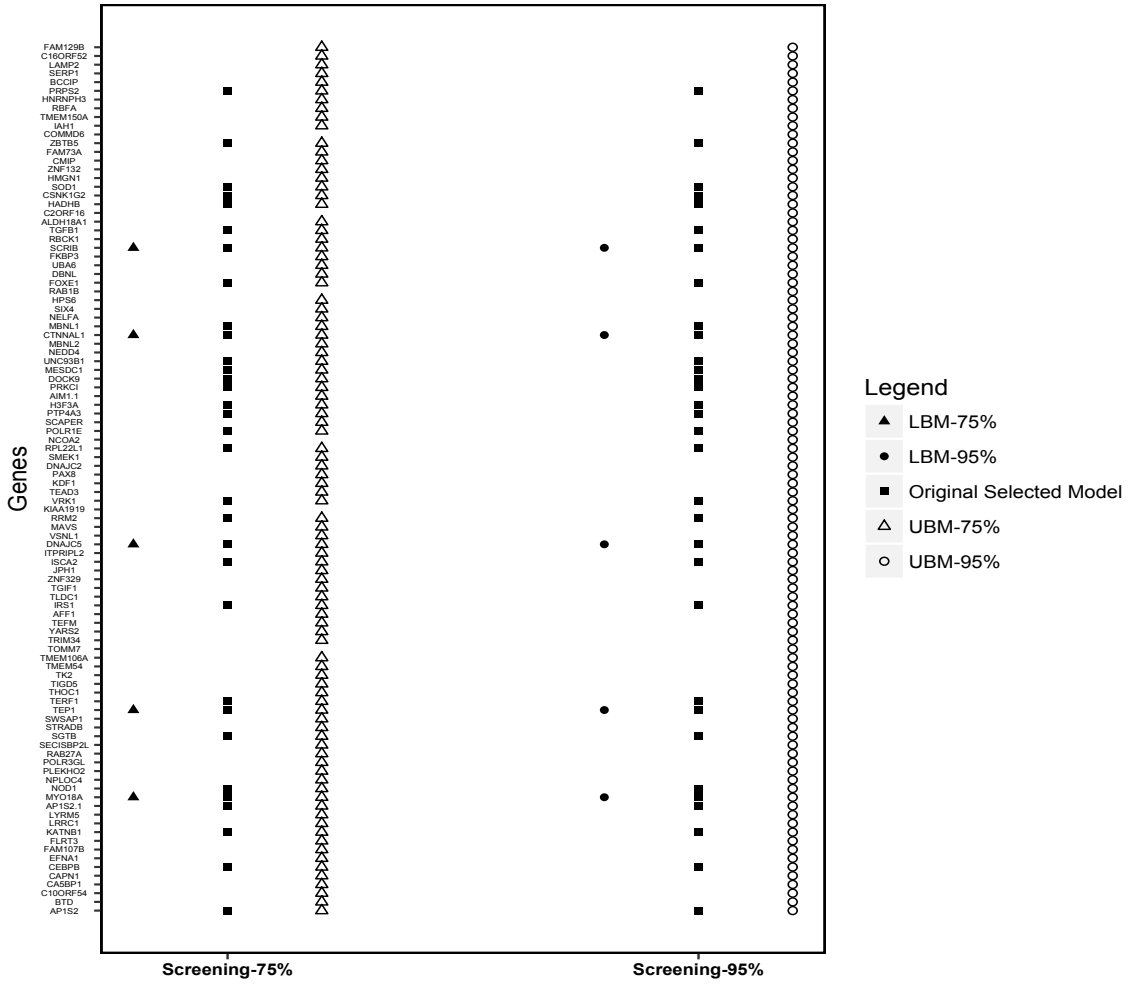
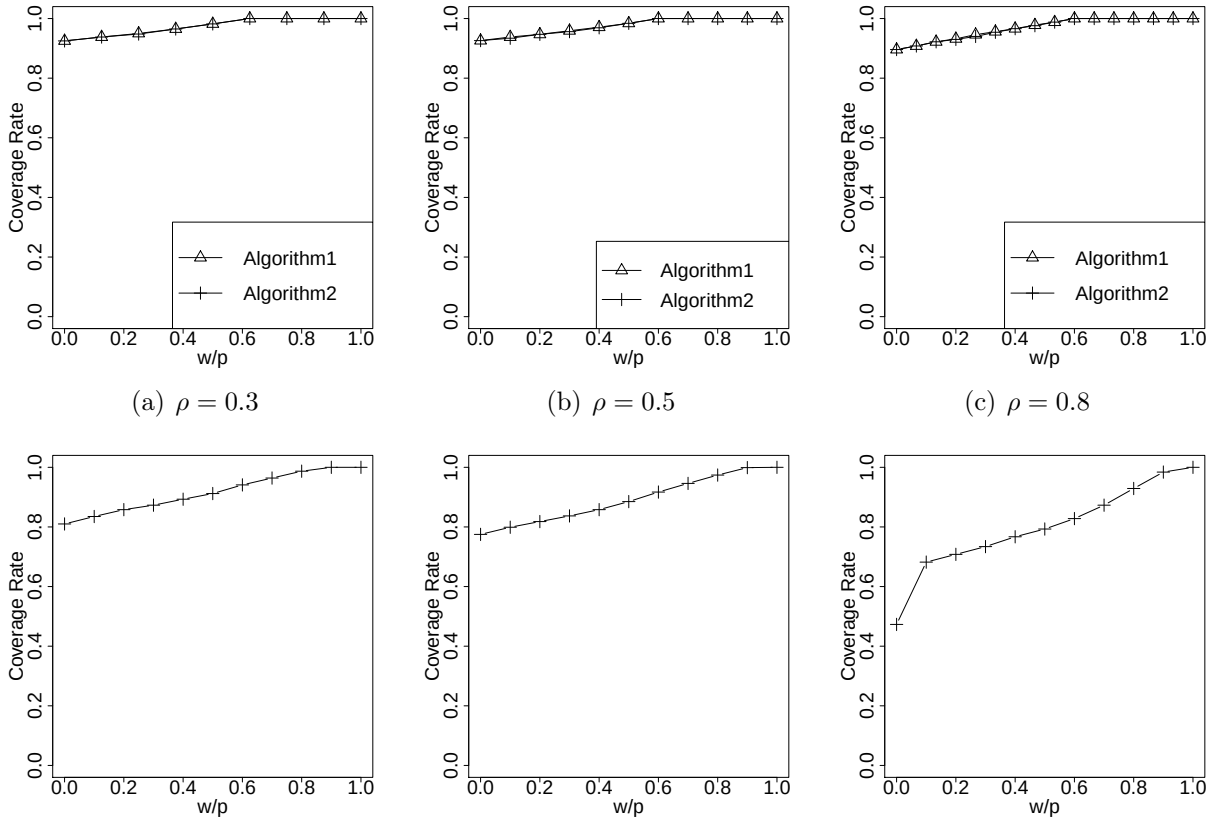


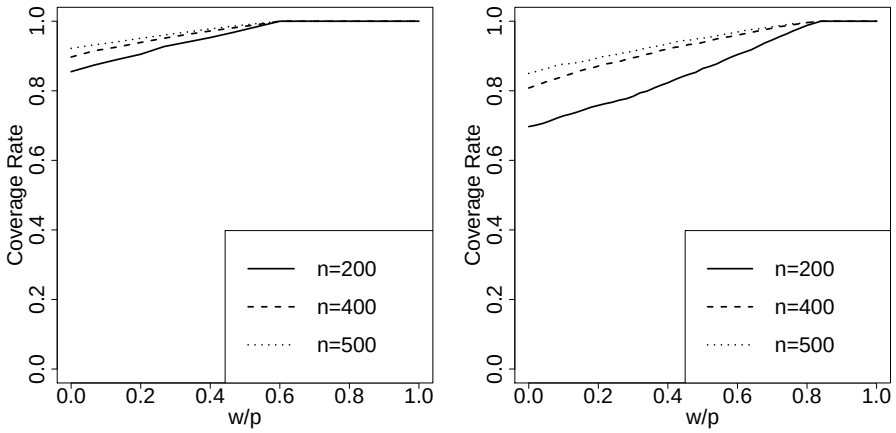
Figure 3: MCBs of screening method at 75% and 95% confidence levels. $B = 300$ bootstrap samples are generated. The original selected model by screen is also marked.



(a) $\rho = 0.3$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

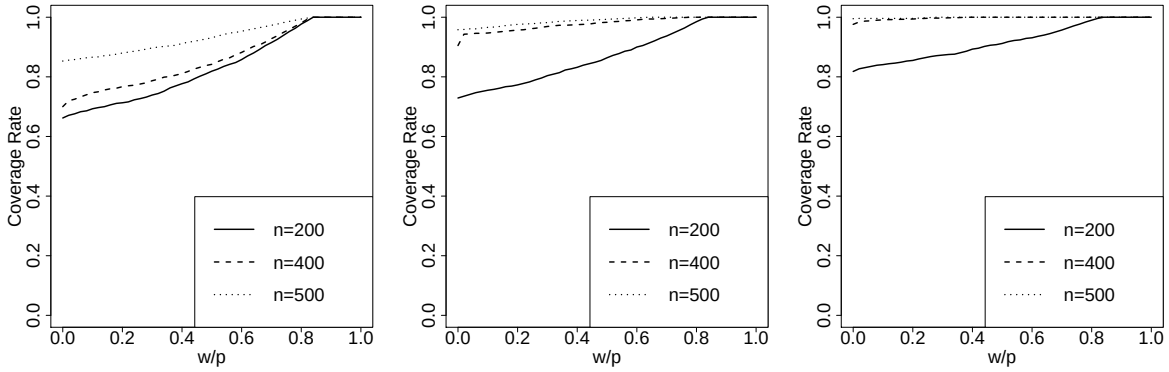
(d) $p = 50$ (e) $p = 100$ (f) $p = 200$

Figure 4: MUCs of Algorithms 1 and 2 for different scenarios with correlated covariates (i.e. $\rho = 0.5$). The top row shows the results of Algorithms 1 and 2: (a) $p = 8$, $p^* = 3$; (b) $p = 10$, $p^* = 4$; (c) $p = 15$, $p^* = 6$. The second row shows only Algorithm 2 because Algorithm 1 is infeasible in these cases: (d) $p = 50$, $p^* = 8$; (e) $p = 100$, $p^* = 10$; (f) $p = 200$, $p^* = 12$. The Adaptive Lasso is used as the variable selection method.



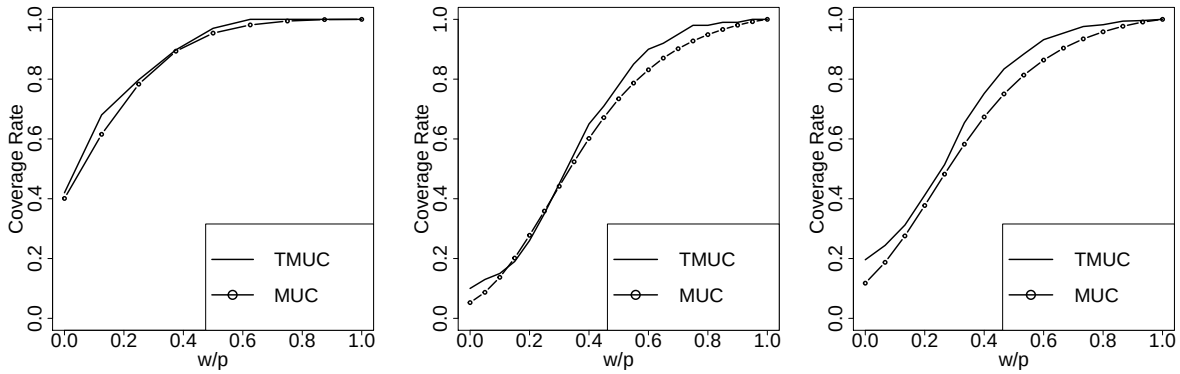
(a) $p = 15$ (b) $p = 50$

Figure 5: Performance of MCB for different sample sizes using independent covariates. Panel: (a) $p = 15$, $p^* = 6$; (b) $p = 50$, $p^* = 8$. The Adaptive Lasso is used as the model selection method.



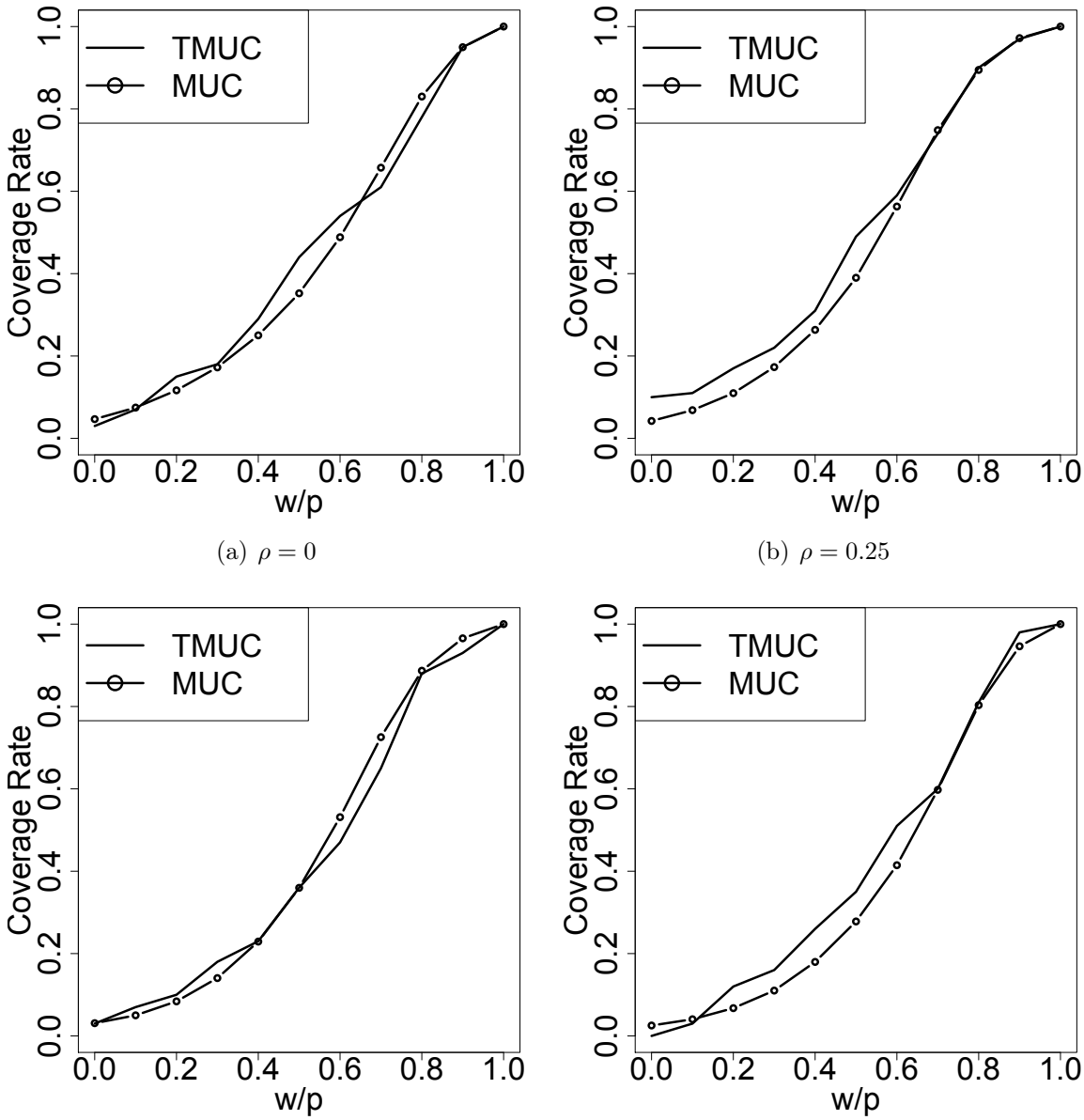
(a) $\rho = 0.25$ (b) $\rho = 0.5$ (c) $\rho = 0.75$

Figure 6: Performance of MCB for different sample sizes using correlated covariates. Panel: (a) $p = 50$, $p^* = 8$, $\rho = 0.25$; (b) $p = 50$, $p^* = 8$, $\rho = 0.5$; (c) $p = 50$, $p^* = 8$, $\rho = 0.75$. The Adaptive Lasso is used as the model selection method.



(a) $p = 8$, $p^* = 3$, $\sigma = 10$ (b) $p = 20$, $p^* = 8$, $\sigma = 10$ (c) $p = 20$, $p^* = 8$, $\sigma = 6$

Figure 7: Comparison of MUC and TMUC. We consider three scenarios: (a) $p = 8$, $p^* = 3$, $n = 1000$, $\sigma = 10$, (b) $p = 20$, $p^* = 8$, $n = 1000$, $\sigma = 10$, and (c) $p = 20$, $p^* = 8$, $n = 300$, $\sigma = 6$. We set $B = 1000$. 500 replications are sampled to compute the TCR. The Adaptive Lasso is used as the variable selection method.



(a) $\rho = 0$

(b) $\rho = 0.25$

(c) $\rho = 0.5$

(d) $\rho = 0.75$

Figure 8: Comparison of MUC and TMUC. We consider four scenarios of covariates: (a) $\rho = 0$; (b) $\rho = 0.25$; (c) $\rho = 0.5$; (d) $\rho = 0.75$;. The Adaptive Lasso is used as the variable selection method.

Confidence Level	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	99%
									$\rho = 0$										
Width	2.04	3.03	3.78	4.43	4.97	5.48	5.90	6.31	6.64	6.97	7.28	7.58	7.82	8.14	8.38	8.64	8.95	9.29	9.75
D_L	0.48	0.37	0.32	0.30	0.29	0.26	0.26	0.25	0.25	0.23	0.22	0.20	0.18	0.14	0.12	0.08	0.05	0.02	0.00
D_U	1.29	1.05	0.87	0.71	0.59	0.47	0.37	0.29	0.21	0.15	0.11	0.07	0.05	0.01	0.00	0.00	0.00	0.00	0.00
Integrated Score	10.91	10.13	9.73	9.50	9.37	9.13	9.07	9.02	8.91	8.89	8.92	8.91	8.97	8.93	8.99	9.04	9.20	9.39	9.76
Cardinality	5.84	11.75	18.86	29.55	42.00	59.74	77.63	101.87	126.62	156.06	193.41	230.91	269.50	340.22	399.87	469.50	566.53	695.04	900.10
Coverage Rate	0.11	0.19	0.26	0.33	0.39	0.46	0.50	0.56	0.62	0.67	0.71	0.76	0.79	0.85	0.89	0.92	0.95	0.98	1.00
									$\rho = 0.25$										
Width	2.77	3.58	4.13	4.63	5.01	5.38	5.68	5.95	6.23	6.43	6.67	6.92	7.12	7.34	7.64	7.94	8.28	8.75	9.47
D_L	0.98	0.91	0.86	0.81	0.76	0.72	0.67	0.62	0.54	0.49	0.44	0.36	0.31	0.27	0.20	0.15	0.10	0.06	0.01
D_U	0.40	0.28	0.19	0.14	0.12	0.07	0.05	0.03	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Integrated Score	9.67	9.53	9.38	9.38	9.41	9.33	9.28	9.20	9.03	8.93	8.87	8.72	8.67	8.69	8.64	8.69	8.78	9.05	9.52
Cardinality	8.76	15.01	21.81	30.22	39.46	50.16	61.76	73.58	89.66	102.24	121.25	142.37	165.12	191.17	235.46	290.56	361.60	498.18	770.82
Coverage Rate	0.27	0.31	0.35	0.37	0.39	0.42	0.45	0.48	0.53	0.58	0.62	0.69	0.72	0.76	0.81	0.85	0.90	0.94	0.99
									$\rho = 0.5$										
Width	3.12	3.87	4.40	4.84	5.26	5.56	5.84	6.12	6.37	6.62	6.85	7.07	7.27	7.54	7.86	8.12	8.46	8.96	9.60
D_L	1.18	1.09	1.00	0.92	0.85	0.80	0.75	0.67	0.59	0.52	0.46	0.42	0.36	0.28	0.21	0.16	0.10	0.04	0.01
D_U	0.25	0.15	0.09	0.07	0.05	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Integrated Score	10.27	10.07	9.85	9.79	9.76	9.71	9.69	9.52	9.37	9.22	9.15	9.17	9.07	8.94	8.91	8.92	8.96	9.16	9.65
Cardinality	10.55	17.34	25.11	33.24	44.14	54.37	65.07	78.78	93.92	111.58	129.95	150.78	173.38	209.41	262.27	312.45	398.08	550.14	825.34
Coverage Rate	0.19	0.24	0.29	0.32	0.37	0.39	0.41	0.47	0.52	0.57	0.60	0.63	0.68	0.74	0.80	0.84	0.91	0.96	0.99
									$\rho = 0.75$										
Width	3.72	4.50	5.04	5.49	5.86	6.19	6.46	6.74	6.97	7.19	7.43	7.67	7.92	8.12	8.36	8.65	8.97	9.34	9.81
D_L	1.26	1.11	1.01	0.94	0.88	0.83	0.75	0.68	0.60	0.54	0.47	0.38	0.31	0.26	0.19	0.13	0.08	0.05	0.01
D_U	0.36	0.23	0.16	0.10	0.07	0.05	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Integrated Score	11.82	11.20	10.89	10.69	10.61	10.59	10.36	10.24	10.02	9.94	9.78	9.57	9.47	9.42	9.31	9.30	9.37	9.59	9.86
Cardinality	15.50	26.55	37.71	51.36	65.82	82.08	99.20	118.98	138.62	161.41	190.14	225.15	265.86	303.49	358.91	443.90	550.66	705.54	929.28
Coverage Rate	0.17	0.23	0.26	0.31	0.34	0.37	0.42	0.47	0.52	0.56	0.61	0.69	0.73	0.77	0.83	0.88	0.93	0.95	0.99

Table 1: Illustration of integrated score, width, penalties, cardinality and true coverage rate of MCB at different confidence levels.