Two Tales of Variable Selection for High Dimensional Regression: Screening and Model Building

Cong Liu, Tao Shi and Yoonkyung Lee
Department of Statistics, The Ohio State University

Abstract

Variable selection plays an important role in high dimensional regression problems where a large number of variables are given as potential predictors of a response of interest. Typically, it arises at two stages of statistical modeling, namely screening and formal model building, with different goals. Screening aims at filtering out *irrelevant variables* prior to model building where a formal description of a functional relation between the variables screened for relevance and the response is sought. Accordingly, proper comparison of variable selection methods calls for evaluation criteria that reflect the differential goals: accuracy in ranking order of variables for screening and prediction accuracy for formal modeling.

Without delineating the difference in the two aspects, confounding comparisons of various screening and selection methods have often been made in the literature, which may lead to misleading conclusions. In this paper, we present comprehensive numerical studies for comparison of four commonly used screening and selection procedures: correlation screening (a.k.a. Sure Independence Screening), forward selection, LASSO and SCAD. By clearly differentiating screening and model building, we highlight the situations where the performance of these procedures might differ. In addition, we propose a new method for cross validation for LASSO. Furthermore, we discuss connections to relevant comparison studies that appeared in the recent literature to clarify different findings and conclusions.

1 Introduction

Variable selection plays an important role in high dimensional regression analysis. Recent developments in methodology have focused on the "large p small n" problems where the

AMS subject classifications: Primary-62J05; secondary-62J07. Keywords: Cross-Validation, Forward Selection, LASSO, ROC Curve, SCAD, SIS

number of predictive variables p is much larger than the sample size n. In this setting, sparsity is often assumed in the sense that only few variables (denoted by q with $q \ll p$) are supposed to be relevant to the response. The goal of variable selection is therefore to identify these *relevant variables* based on data.

Variable selection commonly arises at two different stages of modeling, namely screening and formal model building. Screening is for reducing the number of predictors to a moderate size m, which is usually comparable to the sample size, while trying to keep most of the relevant ones. No formal modeling is necessarily required for variable screening. In contrast, variable selection for model building involves construction of formal models and evaluation of their performance defined by a certain empirical criterion to select a single best model.

In practice, screening is often implemented as an initial step prior to model building, especially in large p small n problem. However, there is clear difference in the objectives of variable selection for screening and model building. For variable screening, it is desirable to order the variables accurately in terms of their relevance, putting relevant variables above irrelevant ones. With this ordering of all variables, one may eliminate a large number of variables at the bottom of the list from further investigation, additional data collection, or final model building. In contrast to screening, when variable selection is concerned in the model building process, a great emphasis is placed on the prediction accuracy of the model built on the final set of selected variables. Due to the difference in the objectives, the effectiveness of a given method in these two different settings should be studied separately.

In a typical screening process, a sequence of (usually nested) subsets of variables is generated by ranking the variables with a simple measure of association with the response or by varying a tuning parameter associated with a penalization method. For instance, correlation screening or Sure Independence Screening (SIS) in Fan and Lv (2008) ranks all the variables by the absolute value of empirical Pearson's correlation coefficient between the response and each predictor. As another example, the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) generates a sequence of variable subsets in a stepwise manner when the L_1 penalty parameter λ is varied from ∞ (corresponding to the null set of no variable) to 0 (corresponding to the "full" set of all variables). In each step, only one variable is added to the current subset.

Given a subset of candidate variables after screening, formal models are fitted. The model building process usually involves comparison of multiple candidate models and selection among them. For model selection, a criterion is used to evaluate the goodness of each candidate model. Popular choices of such criteria include prediction-error-based criteria such as Mallows' C_p or cross-validation (CV) error, and information-based criteria such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) or other variants. The quality of the final model is determined by both the algorithm used to generate candidate models and this selection criterion. The influence of the evaluation criteria on comparisons of variable selection procedures has often been overlooked in the literature, which could lead to misleading conclusions or claims about different model fitting procedures.

In this paper, we study the effectiveness of commonly used variable selection procedures in the screening stage and the model building stage separately. We present comprehensive simulation studies under a variety of settings encompassing a range of signal-to-noise ratio and correlation among the variables. Taking a broader perspective of a variable selection method as a procedure that generates a sequence of subsets of variables, the studies include four commonly used methods: Sure Independence Screening (SIS), Forward Selection (FS), LASSO, and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001; Fan and Peng, 2004). Differently from simply ranking all predictors by their absolute correlation coefficients with the response in SIS, the FS adds the variable that would reduce the residual sum of squares most, if added to the current variable subset for an ordinary least squares model, at each step. Screening consistency has been established for both methods in high dimensional settings; see Fan and Lv (2008) and Wang (2009). On the other hand, LASSO and SCAD generate a sequence of variable subsets by penalizing regression coefficients. We note that there are a class of related penalized regression methods such as the nonnegative garrote (Breiman, 1995; Yuan and Lin, 2007), the adaptive LASSO (Zou, 2006; Huang et al., 2008), the elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), and the Dantzig selector (Candes and Tao, 2007). Here LASSO and SCAD are taken as two representatives of this class of penalized regression methods.

We compare the screening performance of these procedures by investigating the quality of the whole sequence of variable subsets generated by each procedure. For these comparisons, we employ two different criteria: (a) the overall accuracy measured by the area under the curve (AUC) in the receiver operating characteristic (ROC) analysis, and (b) the number of relevant variables among the top m variables for a pre-determined set size m.

For model building, we examine the effect of model evaluation criteria on variable selection by considering widely applied cross-validation criterion and a variation of BIC criterion used in Wang (2009). In addition, we investigate the effect of the combination of each model fitting method and evaluation criteria. A series of papers (Leng et al., 2006; Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009) have suggested that LASSO coupled with minimum CV error tends to produce a model that includes a large number of irrelevant variables. To alleviate this issue, we propose to calculate CV error for LASSO by using OLS regression rather than LASSO fit for each tuning parameter λ, and include this LASSO-OLS hybrid method in the numerical study. We evaluate the performance of variable selection methods in the model building process in terms of model size, number of selected relevant variables and prediction accuracy.

Our numerical studies suggest that LASSO, when properly compared in the context of screening, has the potential to be more effective than the other three methods under a wide range of settings. We report the findings of the ROC analysis and fixed size screening in Section 2. For model building, the results show that the proposed λ-based LASSO-OLS hybrid method can reduce the size of final models substantially by eliminating many more irrelevant variables than the plain cross-validated LASSO, while keeping the majority of relevant variables. It generally outperforms other combinations considered in our study. We make comparisons of the accuracy in selection and prediction for the final models chosen by combinations of variable selection methods and evaluation criteria in Section 3, and further discuss differences in the findings from other studies in the literature. We conclude in Section 4.

2 Variable Screening

This section focuses on comparison of the screening performance of FS, LASSO, SCAD and SIS. The main objective of variable screening is to eliminate a majority of irrelevant variables, while keeping as many of the true predictors as possible, by reducing the size of candidate variables to a moderate number m. Thus, the goodness of a screening method largely hinges on the order in which relevant variables are added to the final subset and the choice of m. To assess the accuracy of the rank order of variables, we examine the receiver operating characteristics (ROC) of the sequence of variable subsets generated by a screening method.

2.1 Preliminaries

In our study, we evaluate the screening performance with two criteria: (a) AUC, the area under a modified ROC curve, and (b) the number of relevant variables selected among the top m variables, denoted by T_m. The ROC curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) given a sequence of variable subsets. For each subset,

$$\begin{aligned} \text{TPR} &= \frac{\text{number of selected relevant variables}}{\text{total number of relevant variables}}, \qquad (2.1)\\ \text{FPR} &= \frac{\text{number of selected irrelevant variables}}{\text{total number of irrelevant variables}}. \qquad (2.2) \end{aligned}$$

For large p small n scenario considered in this paper, the definition of traditional ROC curve needs to be modified. In this case, screening methods like LASSO or FS can select at most n variables, which lowers the upper limit of FPR from 1 to $n/(p-q)$. To make the ROC plots comparable across different settings of p, q and n, we change the x-axis of ROC plots to the size of each subset, which increases from 0 to n regardless of p and q. Hence, as long as n is fixed, we can evaluate the screening performance of each method through the AUC of its modified ROC curve.

There are two factors that influence the AUC: (a) the number of true predictors selected by the full (or maximal) set, which decides the maximum TPR that is attainable, and (b) the rank order of the true predictors, which determines the rate of increase in TPR. A method which picks a majority of relevant variables in early stages is likely to yield a larger AUC.

2.2 Numerical Study

We conduct an extensive simulation study to compare the screening performance of the four methods in terms of both AUC and T_m under a variety of settings obtained by controlling several critical factors such as the signal-to-noise ratio and the covariance structure of the predictors. By doing so, we try to identify the settings favorable to each method and thus provide guidance to an appropriate choice of method.

2.2.1 Experimental Settings

We consider the following data generating model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{X} \equiv (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p)$ is an $n \times p$ design matrix, $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, ..., \beta_q, \beta_{q+1}, ..., \beta_p)^T$ is a vector of regression coefficients, $\mathbf{Y} \equiv (Y_1, Y_2, ..., Y_n)^T \in \mathbb{R}^n$ is a vector of responses, and $\boldsymbol{\epsilon} \equiv (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ with ε_i *i.i.d.* with $N(0, \sigma^2)$. Without loss of generality, assume that the first q ($q \ll p$) elements of $\boldsymbol{\beta}$ are non-zero and the last $(p - q)$ elements are zero. Let $\boldsymbol{\beta}_{(1)} \equiv (\beta_1, \beta_2, ..., \beta_q)^T$ and $\boldsymbol{\beta}_{(2)} \equiv (\beta_{q+1}, \beta_{q+2}, ..., \beta_p)^T = \mathbf{0}$. In the same way, $\mathbf{X}$ is divided accordingly into two parts with $\mathbf{X}_{(1)} \equiv (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_q)$ and $\mathbf{X}_{(2)} \equiv (\mathbf{X}_{q+1}, \mathbf{X}_{q+2}, ..., \mathbf{X}_p)$. Therefore, only a small subset of the predictors, $\mathbf{X}_{(1)}$, are *relevant* to the mean response.

The design matrix $\mathbf{X}$ is generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma_{p\times p})$. We fix the sample size n at 100 and vary the dimension p. In addition, we fix the number of true variables q at 10. Besides the dimension p that might affect screening performance, we primarily control three other factors listed below (see summary in Table 1) and study the difference of screening methods under combinations of the factor levels.

1. Correlation between variables (ρ)

   The covariance matrix Σ controls the correlation among the predictors. Generally, strong correlation among the predictors makes discrimination of relevant variables from irrelevant ones difficult. For simplicity, we adopt the compound symmetry structure for Σ. That is, all the predictors are equally correlated with each other. $\Sigma = [\sigma_{ij}]$ has diagonal elements $\sigma_{ii} = 1, i = 1, ..., p$ and equal off-diagonal elements $\sigma_{ij} = \rho$, for all $i \neq j$, $0 \leq \rho \leq 1$. The levels of ρ are set to $\{0, 0.3, 0.6, 0.9\}$.

2. Signal-to-noise ratio (SNR)

   The signal-to-noise ratio under regression setting is defined as

$$\text{SNR} = \frac{\text{Var}(X^T\boldsymbol{\beta})}{\text{Var}(\varepsilon)} = \frac{\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}}{\sigma^2}, \tag{2.4}$$

   which can be a dominant factor in many situations. Most methods perform fairly well in a very high SNR setting and almost equally poorly in the opposite setting. For meaningful comparisons, we consider three different levels of SNR from low (1) to medium (5) to high (10) by setting the values of σ^2 in (2.4).

3. Regression coefficients (BETA)

   This is a factor often overlooked by some related studies. Not only the sizes of coefficients, but also their signs can be important. The signs affect the correlation between the response $\mathbf{Y}$ and each predictor. Mainly focusing on the pattern of the signs, we set up three levels with two extreme cases and one in the middle: (i) "pure": all ten non-zero coefficients are $+1$'s; (ii) "half-half": five $+1$'s and five -1's; (iii) "mixed": $\boldsymbol{\beta}_{(1)} = (1, 1, -1, ..., -1)$. In addition, to investigate the impact of the size of coefficient, we consider (iv) "two-levels": two groups of positive coefficients have different values as given by $\boldsymbol{\beta}_{(1)} = (1, ..., 1, 0.5, ..., 0.5)$. Mixing positive and negative coefficients for positively correlated variables creates scenarios that make marginal regression and joint regression differ.

One hundred replicates of data are generated for each combination of all levels of the four factors listed in Table 1. For each replicate, we generate the sequence of subsets of selected variables using each of the four methods, and then record AUC and T_m accordingly. The averages of both AUC and T_m (with $m = 20, 40, 60, 80$) over 100 replicates are used to evaluate the screening performance.

Table 1: Treatment levels of four factors: p, ρ, SNR and BETA.

Factor	Levels
Number of variables (p)	200, 600, 1000
Correlation between variables (ρ)	0, 0.3, 0.6, 0.9
Signal-to-noise ratio (SNR)	1, 5, 10
Regression coefficients (BETA)	pure, half-half, mixed, two-levels

2.2.2 Results of ROC Analysis

We first report the overall screening performance with the AUC described in Section 2.1. Figure 1 shows ROC curves from a randomly selected sample, where $p = 1000$, BETA =“pure”, SNR = 10, and ρ varies from 0 to 0.9. In general, the AUC of all four methods decreases significantly when ρ increases. LASSO is much better than the other three methods in terms of AUC when $\rho = 0$ or 0.3, and it is more robust against the increase in ρ. No method produces satisfactory performance when $\rho = 0.9$. We observed similar results in other samples.

Due to the non-convexity of the SCAD penalty, we observed that the SCAD subset path is not very stable in general and it often drops some relevant variables already in the model as the penalty decreases. Differently from SIS, FS, and LASSO whose solutions are naturally indexed by the subset size, SCAD does not allow such a systematic parametrization of its solutions with subset size. Instead, we get solutions over a grid of penalty parameter values for the SCAD path using the R package *ncvreg*. As it does not provide a direct control over the subset size, when the maximum subset size n was not attainable for a range of penalty parameter values, we had to extrapolate from the largest subset size from SCAD to n for comparison with other methods. All these factors make SCAD less favorable for screening.

We calculate the AUC individually for each replicate and compute the average AUC for each combination of the four factors. We examine the pairwise differences of the average AUC among these four methods. Define $\Delta AUC_{LF} = AUC_{LASSO} - AUC_{FS}$, $\Delta AUC_{LS} = AUC_{LASSO} - AUC_{SIS}$, $\Delta AUC_{LSC} = AUC_{LASSO} - AUC_{SCAD}$ and so on. To summarize the results and to visually compare the screening performance of LASSO, SCAD, FS and SIS, we use boxplots of the pairwise differences. Figure 2 displays the main effects of the four factors.

Evidently Figure 2 suggests that LASSO is better than FS, SCAD, and SIS for screening in terms of the AUC. In particular, in Figure 2(b) for LASSO and FS, an overwhelming majority of ΔAUC_{LF} are greater than 0, indicating that LASSO is superior to FS for screening across a broad range of settings. Very similar patterns are observed for LASSO and SCAD in Figure 2(c) as well. Although the pattern is less pronounced for ΔAUC_{LS} in Figure 2(a), we observe that ΔAUC_{LS} is positive on average in the most of the settings, and it increases in both p and SNR. In the high noise setting of SNR = 1, no method can effectively retrieve enough true variables. These results indicate that LASSO could be more effective for screening in practically tractable high dimensional problems. The comparisons among SCAD, SIS and FS, however, are less straightforward, as shown in Figures 2(d), 2(e), and 2(f). Each method has its own preferred settings, depending mainly on SNR. The AUC values of SCAD and FS are very close to each other in almost every setting whereas SIS outperforms both SCAD and FS when SNR = 1 and is less favorable when SNR = 5 and 10. Similar to

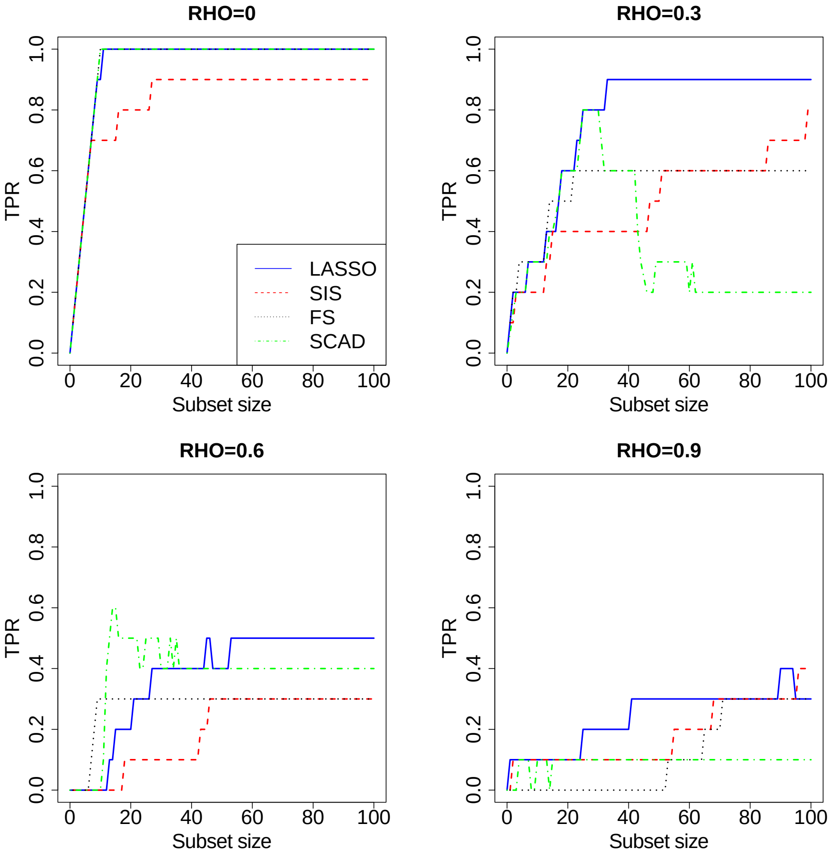
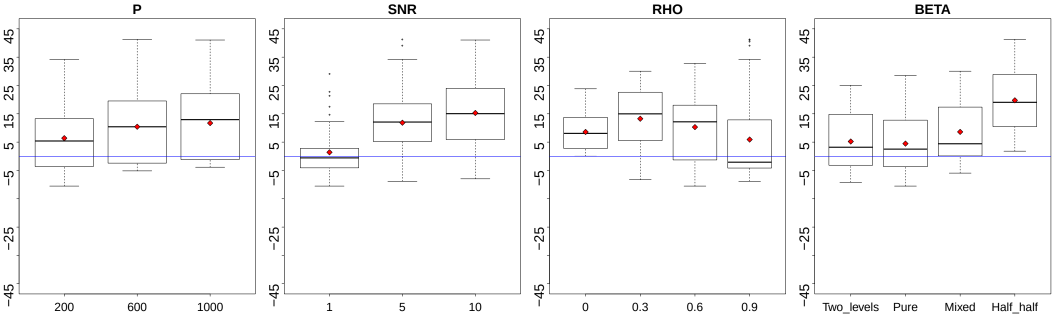
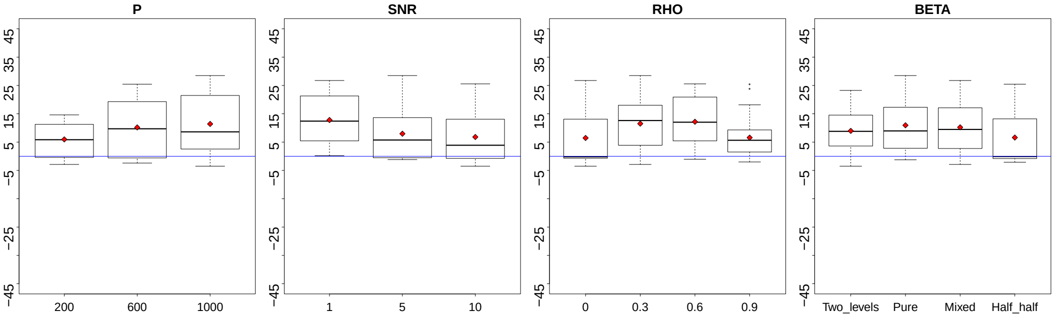


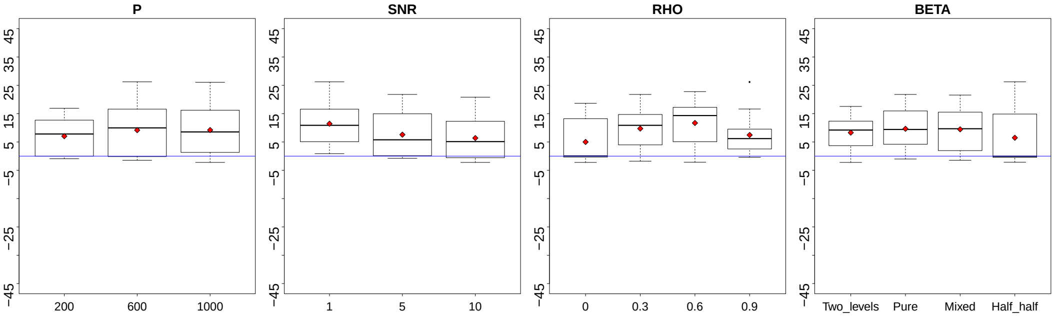
Figure 1: Sample ROC curves of LASSO, SIS, FS and SCAD for $p = 1000$, BETA =“pure”, SNR = 10 and $\rho = 0$, 0.3, 0.6 and 0.9.



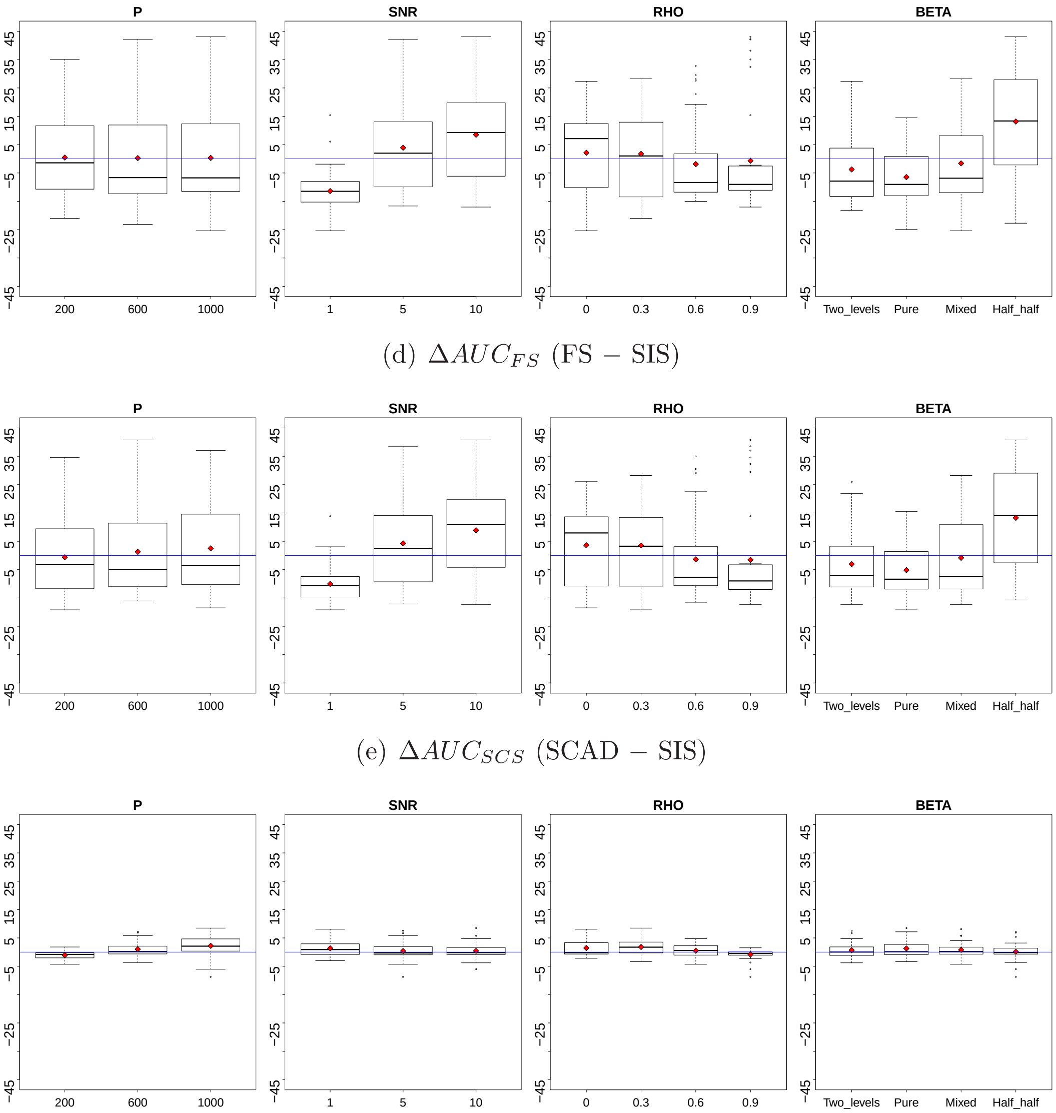
(a) ΔAUC_{LS} (LASSO − SIS)



(b) ΔAUC_{LF} (LASSO − FS)



(c) ΔAUC_{LSC} (LASSO − SCAD)



(d) ΔAUC_{FS} (FS − SIS)

(e) ΔAUC_{SCS} (SCAD − SIS)

(f) ΔAUC_{SCF} (SCAD − FS)

Figure 2: Boxplots of pairwise differences in AUC against the four factors of p, SNR, ρ, and BETA for the four methods, LASSO, SIS, FS, and SCAD. The red diamond represents the mean value.

ΔAUC_{LS}, ΔAUC_{FS} and ΔAUC_{SCS} increase steadily in SNR, and their averages stay above 0.

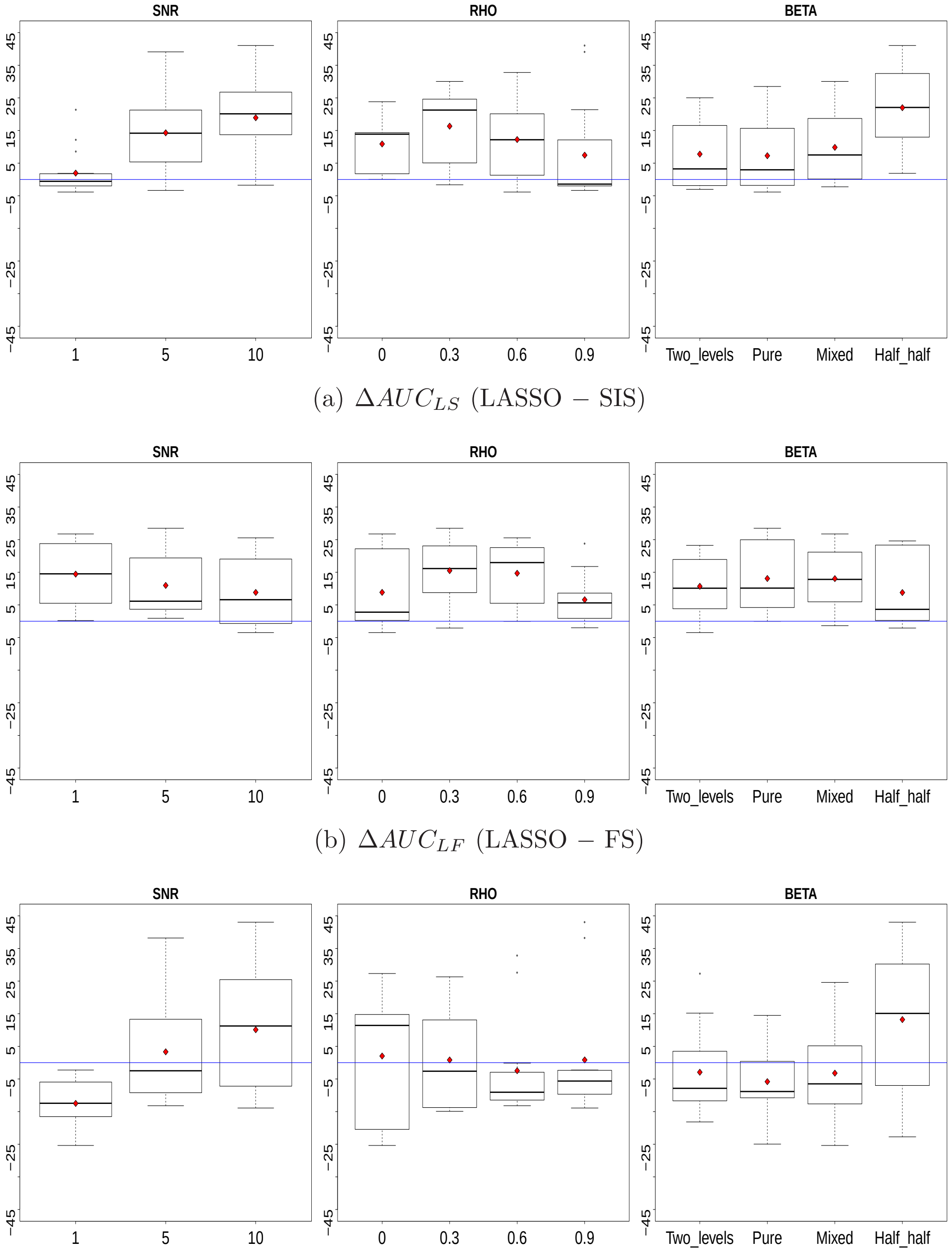
Besides the effect of SNR, high correlation seems to produce rather peculiar results for ΔAUC_{FS} in Figure 2(d) and ΔAUC_{SCS} in Figure 2(e). When $\rho = 0.6$ or 0.9, the central box is well below the horizontal line of zero, but the mean difference is dragged up to almost zero by several positive outliers. This phenomenon is associated with one special setting of BETA, "half-half". Under this setting, relevant variables are divided into two groups of the same size with different signs in the true coefficients. As expected, SIS based on marginal correlation is much less effective than LASSO, SCAD and FS in the setting. For other three methods, all the relevant variables are much more likely to be selected in early stages alternately from the two groups given the compound symmetry covariance matrix with a positive correlation. This particular setting may not be of practical importance, but nonetheless shows the difference of SIS from LASSO, SCAD and FS.

Since our primary interest lies in the "large p small n" setting, we take a closer look at the $p = 1000$ case. The boxplots of pairwise difference of AUC against other three factors for $p = 1000$ are displayed in Figure 3. Note that SCAD is not included in this comparison as its performance is similar to that of FS as shown in Figure 2(f). The boxplots in Figure 3 are similar to those in Figure 2 except for a few noticeable differences. As we expected, as a screening tool, LASSO is even more superior to FS (SCAD) and SIS in the AUC when $p = 1000$ and there is no clear winner between SIS and FS (SCAD). Compared with Figure 2(c), the notable changes are that not only the central box of ΔAUC_{FS} for $\rho = 0.9$, but also the one for $\rho = 0.6$ is entirely below zero, which suggests that FS suffers more from high correlation than SIS when p is relatively large, due to its greediness.

We note that Genovese et al. (2009) also conducted similar ROC analysis in their study, where the selection accuracy is measured by the Hamming distance. They showed that SIS (marginal regression in their terms, fitting OLS models with the variables selected by SIS) is competitive with LASSO as a screening tool and sometimes has a much lower prediction error (measured by MSE) under the setting of $p = 500$ and $q = 100$ (much less sparse than our setting). In addition, the norm of regression coefficients was set to two levels of 0.5 and 5, and it was shown that SIS enjoys significantly lower MSE only when the norm is 0.5. As suggested by Tibshirani (1996), the performance of LASSO can be degraded significantly when a large number of small effects exist.

2.2.3 Results of Fixed Size Screening

Next, we turn to the comparison of results of fixed-size screening in terms of T_m, the number of relevant variables among the top m variables, which directly measures the quality of a



(a) ΔAUC_{LS} (LASSO − SIS)

(b) ΔAUC_{LF} (LASSO − FS)

(c) ΔAUC_{FS} (FS − SIS)

Figure 3: Boxplots of pairwise differences in AUC against the three factors of SNR, ρ, and BETA for the three methods, LASSO, FS and SIS. p is fixed at 1000. The red diamond represents the mean value.

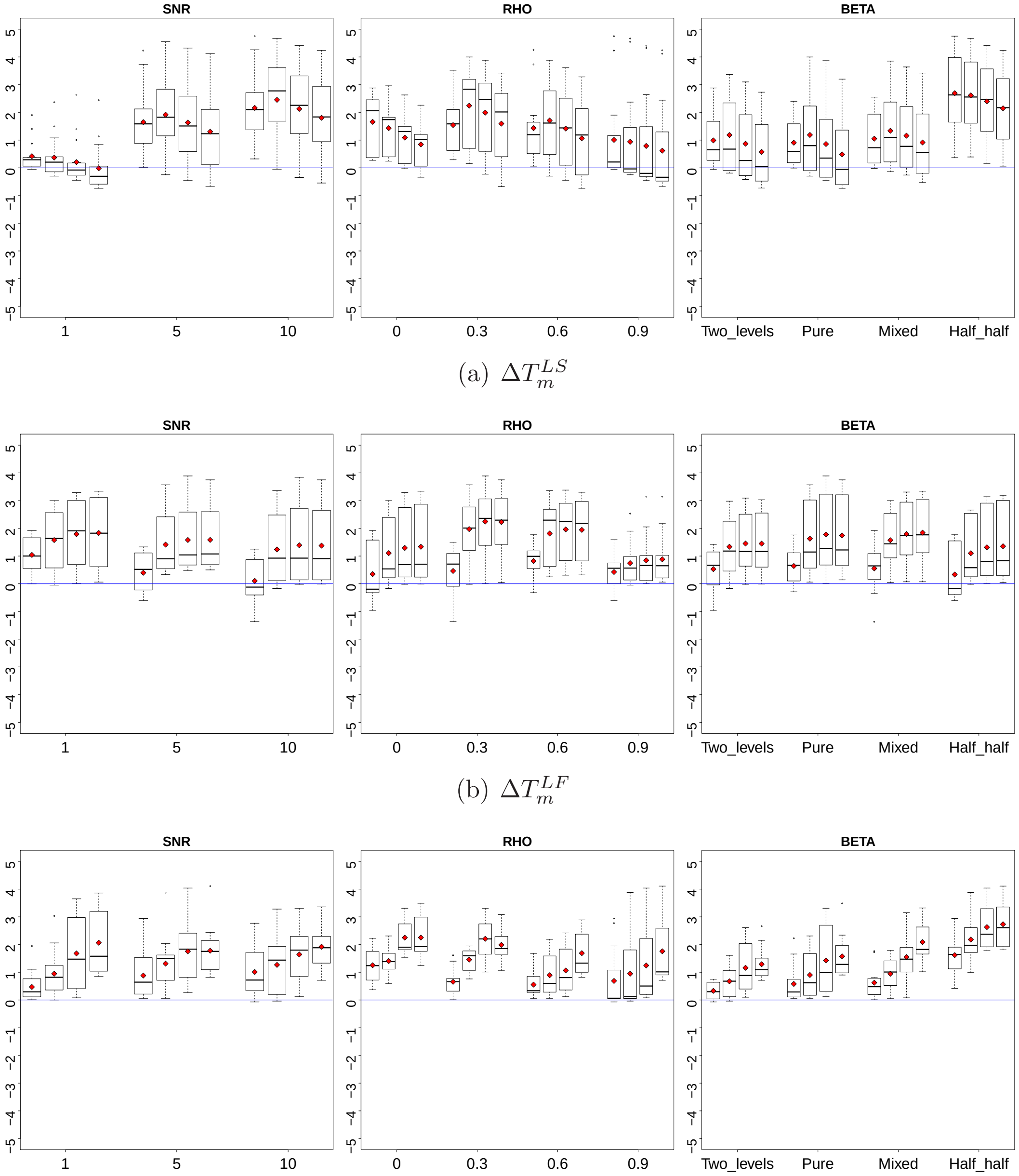
post-screening subset. Similar notations are used for pairwise differences in T_m as for the differences in AUC, e.g. $\Delta T_{20}^{LF} = T_{20}^{LASSO} - T_{20}^{FS}$. Figure 4 shows the boxplots of pairwise differences in T_m between LASSO and other three methods against the main factors of ρ, SNR and BETA when $p = 1000$.

As illustrated in Figure 4, LASSO again dominates other three methods by a significant margin when SNR = 5 and 10. One interesting exception is ΔT_{20}^{LF}, implying that FS occasionally selects more relevant variables in very early steps than LASSO, arguably under high SNR and low correlation scenario. However, LASSO quickly catches up and surpasses FS as m increases. Agreeing with the results in ROC analysis, the settings favorable to SIS remain to be either of low SNR or high correlation, and FS generally does better than SIS in the other settings. The close agreement between AUC and T_m shows that LASSO in general does have some advantage in variable screening over SIS, FS and SCAD, for a reasonably large m.

In addition, the trend in T_m with increasing m may provide some insights into how to choose m in practice. The boxplots of T_m in Figure 5 reveal distinct patterns of T_m as m increases. For LASSO, a significant improvement in T_m occurs when m increases from 20 to 40, and increasing m further does not seem to result in improvement in T_m as indicated by the almost identical boxplots for $m = 40$, 60, and 80. Hence a moderate m is a reasonable choice for LASSO. By contrast, the improvement in T_m^{FS} with increasing m is rather trivial. If any true variables were selected by FS, it is most likely that the correct selection is likely to occur in very early steps. Therefore a better choice of m should be relatively small compared to n. For SIS, unlike T_m^{LASSO} and T_m^{FS}, T_m^{SIS} constantly increases in m with a notable margin. This makes large m a preferable choice for SIS. Lastly, similar to T_m^{FS}, T_m^{SCAD} is relatively stable with increasing m. However, large m does not necessarily yield large T_m^{SCAD} because SCAD may drop relevant variables as the penalty decreases.

In spite of the fact that even T_{80}^{SIS} is usually much smaller than both T_{80}^{LASSO} and T_{80}^{FS} (when SNR is high), the method SIS itself possesses two unrivaled advantages. First, the selection is solely based on the order of absolute marginal correlation between the response and each individual predictor, so m can exceed n (up to p) when $p > n$. Second, the computational cost is much smaller for SIS than both LASSO and FS. With these two advantages combined, SIS can potentially be adopted as a computationally efficient screening tool, reducing huge p to a moderate size m (possibly greater than n). Other selection procedure can then be applied to the reduced variable set.

Fan and Lv (2008) (Section 4.2) also compared the screening performance of SIS and LASSO through a simulation study, where the performance was evaluated by the percentages of all relevant variables included in models fitted after screening. The results showed that LASSO only slightly edges SIS in most cases. Their comparisons, however, raise some

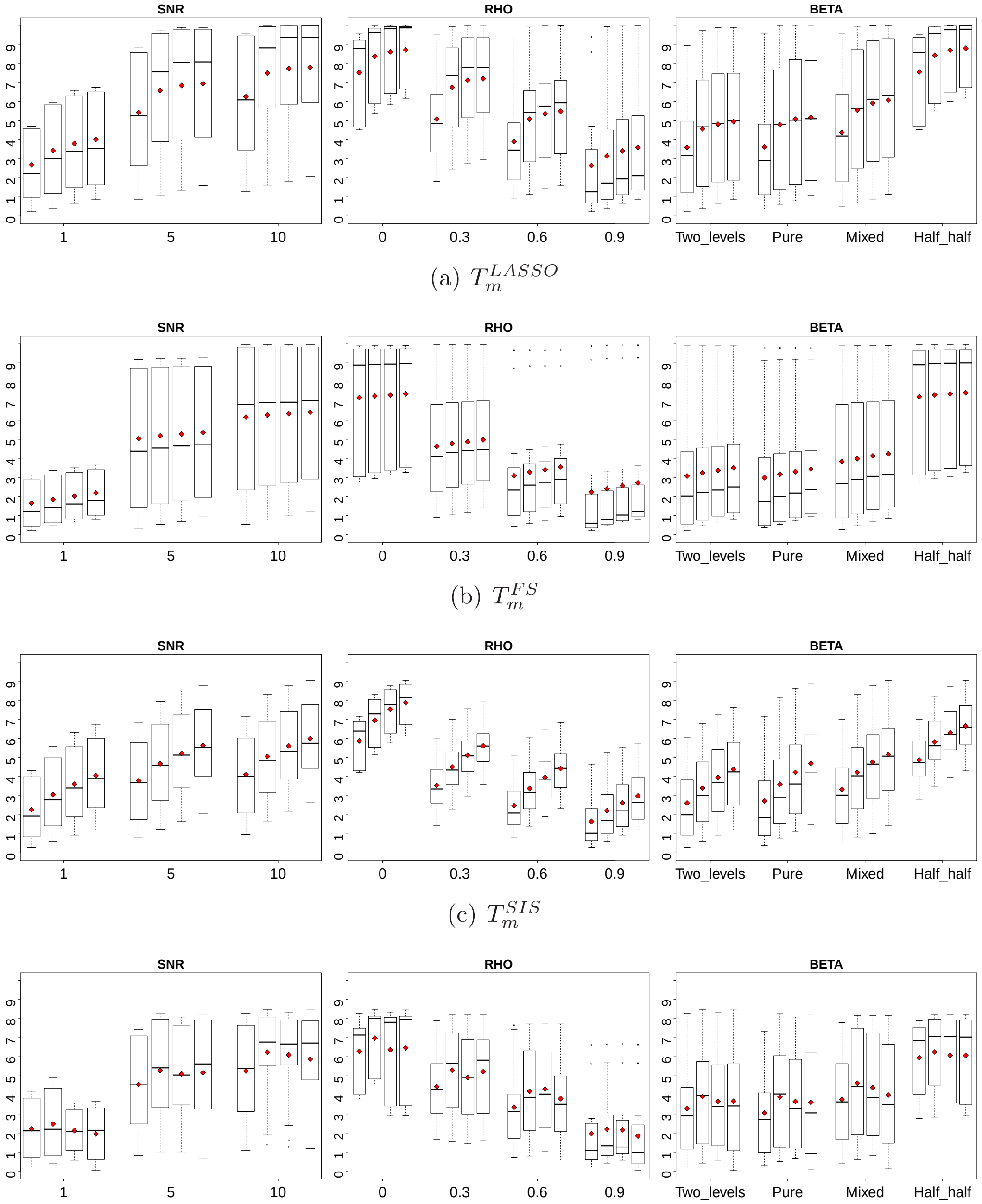


(a) ΔT_m^{LS}

(b) ΔT_m^{LF}

(c) ΔT_m^{LSC}

Figure 4: Boxplots of pairwise differences in the number of selected true variables against the three factors of SNR, ρ, and BETA between LASSO and the other three methods, FS, SIS and SCAD, under fixed subset size of $m = 20$, 40, 60 and 80, respectively (from left to right). p is fixed at 1000. The red diamond represents the mean value.



(a) T_m^{LASSO}

(b) T_m^{FS}

(c) T_m^{SIS}

(d) T_m^{SCAD}

Figure 5: Boxplots of the number of selected true variables when the subset size is fixed at $m = 20$, 40, 60 and 80 respectively, for LASSO, FS, SIS, and SCAD, against the main factors of SNR, ρ, and BETA. $p = 1000$, and the red diamond represents the mean value.

concerns. First, the post-screening size was fixed at n for SIS, while the LASSO model was chosen by minimizing CV, which leads to a model size much smaller than n in the comparison. This difference puts the two methods on a quite different footing. Aside from this, as pointed out by Levina and Zhu (2008) in their discussion paper, the SNRs of all examples in Fan and Lv (2008) are extremely high (from 40 to 200), where any method should have a reasonably good performance given a moderate sample size n.

2.3 Summary

It is customary to view LASSO as a method for model building rather than screening. Our simulation study, however, shows that LASSO as a screening method is better than FS, SIS and SCAD by a noticeable margin in terms of AUC and T_m, given a reasonably large SNR. It can capture most relevant variables with a moderate subset size. Meanwhile, Forward Selection includes many relevant variables in very early stages, but with increasing m, it becomes less effective than LASSO in general. On the other hand, SIS has slight advantage only when SNR is low, where no methods perform reasonably well, and in all other situations, it is much worse than the other three. However, SIS has the unique advantage of selecting more than n variables and is computationally much faster than any penalized or stepwise regression methods. So it could be useful for screening when p is extremely large and n is comparatively very small.

3 Model Building

In this section, we compare the quality of final models built by different combinations of subset generation (now as variable selection) methods, model fitting methods and model evaluation criteria. A good model should have both high selection accuracy and prediction accuracy. For better interpretability, it should also be as concise as possible.

3.1 Preliminaries

For model evaluation and selection, we include both prediction-based criteria and information-based criteria in our study. For the former, we focus on widely used cross-validation (CV). For the latter, we consider a variant of the BIC proposed by Wang (2009), which is designed for high dimensional setting.

Cross-validation is commonly used in model evaluation due to its generality and ease of implementation. For example, it is used as the default criterion for LASSO in the least angle regression (LARS) algorithm for selection of the penalty parameter λ. Despite wide applications of CV in practice, it is well known that selecting a model for a penalized procedure by minimizing CV error tends to favor a very large model. LASSO as a shrinkage method tends to include many irrelevant variables with small non-zero coefficients when the model is chosen by minimizing CV error; see Fan and Lv (2008); Leng et al. (2006); Meinshausen and Bühlmann (2006); Meinshausen and Yu (2009). It is partly explained by the fact that shrinkage of the coefficients for relevant variables generally yields larger CV error than the unpenalized counterpart. Reducing the penalty parameter may lessen the extent of shrinkage, which leads to reduced CV error at the cost of inclusion of many variables with negligible coefficients.

Another important issue when cross-validation is used in the large p small n case is that the sequence of subsets generated by using the entire data tends to be very data-dependent. Treating the subset sequence as given and cross-validating the models fitted based on the subsets would make CV error fairly close to the in-sample error, which in turn, makes cross-validation favor larger models. This selection bias is a critical issue in high dimensional setting, especially for FS.

In this paper, we propose a variant of cross-validation algorithm for LASSO and call it a λ-based LASSO-OLS hybrid method (abbreviated as LO_λ or simply LO). In essence, the hybrid method uses LASSO for subsets generation only, fits models by OLS, and calculates CV error accordingly. Cross-validation based on the OLS models is expected to produce a relatively smaller model than the one picked by the standard CV for LASSO. A related idea can be found in Meinshausen (2007). The hybrid method is described in Table 2.

In the proposed algorithm, note that the variables selected by $\mathcal{M}_j^{[-i]}$ in Step 2 can be quite different from those selected by $\mathcal{M}_j$ on the full data set. This feature is common to the standard LASSO cross validation and the hybrid method. In contrast with standard CV, this property mitigates the selection bias of cross-validation toward larger models due to strong data-dependence of the subsets sequence to some extent.

As an alternative to cross validation, we consider a variant of the BIC proposed in Wang (2009), which modifies the classical BIC with an extra penalty term on the number of parameters p for high dimensional setting. It is defined as

$$BIC(\mathcal{M}) = \log\frac{1}{n}RSS(\mathcal{M}) + \frac{1}{n}|\mathcal{M}|(\log n + 2\log p), \tag{3.1}$$

where $|\mathcal{M}|$ is the size of the model $\mathcal{M}$, and $RSS(\mathcal{M})$ is the residual sum of squares based on the OLS model $\mathcal{M}$. It is closely related to the extended BIC (EBIC) criterion proposed by Chen and Chen (2008), and the authors proved that the selection consistency is guaranteed under regularity conditions when p increases in any arbitrary polynomial rate of n. For simplicity, we use the EBIC to refer this adjusted BIC criterion hereafter.

Table 2: A λ-based LASSO-OLS hybrid algorithm (LO_λ)

Step 1 :	Use LASSO to generate the sequence of variable subsets corresponding to the null model $\mathcal{M}_0$ to the full model $\mathcal{M}_t$ of size n. Record all the corresponding values of the penalty parameter λ, λ_j for $j = 0, 1, ..., t$.
Step 2 :	For K-fold cross-validation, split the data into approximately K equal parts. For every λ_j, apply LASSO with $\lambda = \lambda_j$ to the training data without the i^{th} part and denote the model by $\mathcal{M}_j^{[-i]}$ for $i = 1, ..., K$. Then fit an OLS model with those variables selected by $\mathcal{M}_j^{[-i]}$ and record the OLS estimate $\hat{\boldsymbol{\beta}}_j^{[-i]}$.
Step 3:	For every λ_j, calculate the mean squared error (MSE) of the new OLS estimates as cross validation error $CV(\lambda_j)$.
Step 4:	Find the minimizer λ_{j^*} of $CV(\lambda_j)$ and report the OLS model based on the variables selected by $\mathcal{M}_{j^*}$ as the final model.

Table 3 summarizes all the combinations of variable selection (or subset generation), model fitting methods, and evaluation criteria to be compared in numerical studies along with their labels. For instance, the standard LASSO with CV is viewed as a procedure that uses LASSO for both variable selection and fitting and denoted by LL (LASSO+LASSO). Along with the standard approach of cross validation, we employ the heuristic "one-standard error" rule suggested by Hastie et al. (2001). It picks the most parsimonious model whose CV error is within one standard error of the minimum CV error.

3.2 Numerical Study

Taking the results of screening in Section 2.2 as an input for model building, we compare the performance of all combinations listed in Table 3 under the same simulation setup as in Section 2.2.1. EBIC is calculated in (3.1) based on the OLS fit while CV error is calculated via ten-fold cross-validation. Here we focus on the $p = 1000$ scenario. Given the sequence of subsets generated by each variable selection method in Section 2.2, we further apply fitting algorithms and evaluation criteria to build and select final models.

We evaluate the goodness of the final models in terms of selection accuracy and prediction accuracy. The selection accuracy is measured by the number of selected true predictors T and model size $|\mathcal{M}|$, and the prediction accuracy is reflected by Mean Squared Prediction Errors (MSPE, see Appendix for more details). The average of each accuracy measure over

Table 3: A summary of variable selection, fitting methods, and evaluation criteria

Method Label	Variable Selection (Subset Generation)	Fitting	Evaluation Criteria
LL_{min}	LASSO	LASSO	Minimize CV error
LL_{1se}			Apply "one-se rule" to CV error
LO_{min}		OLS	Minimize CV error
LO_{1se}			Apply "one-se rule" to CV error
LO_{EBIC}			Minimize EBIC
FS_{min}	FS		Minimize CV error
FS_{1se}			Apply "one-se rule" to CV error
FS_{EBIC}			Minimize EBIC
SIS_{min}	SIS		Minimize CV error
SIS_{1se}			Apply "one-se rule" to CV error
SIS_{EBIC}			Minimize EBIC
$SCAD_{min}$	SCAD	SCAD	Minimize CV error
$SCAD_{1se}$			Apply "one-se rule" to CV error
$SCAD_{EBIC}$			Minimize EBIC

the 100 replicates is reported.

Figure 6 provides an overview of the entire selection process for each method listed in Table 3 on a randomly chosen replicate from the scenario with SNR = 10, BETA = "pure" and $\rho = 0$. In this scenario, every selection method should be able to capture most of the true predictors. CV error (averaged over 20 splits), EBIC and MSPE are calculated for every candidate model and plotted against the model size in each panel. The models selected by "min", "one-se" and "EBIC" are marked by symbols of different shapes and colors. A red vertical line is drawn when a relevant variable is added. The pattern shown in the figure is typical among 100 replicates.

Several observations are made from the overview plots. First, the CV error curves of FS and LL are very flat, yet slowly decreasing as the number of variables increases. Compared to LL, CV seems to work well for LO and selects much fewer irrelevant variables while retaining all the true predictors. Second, the effect of "one-se" rule is only significant for LL for this setup. Third, EBIC works well with FS in this high SNR case. However, when combined with other selection methods, the final models tend to be too small to include many of the relevant variables. We also point out that EBIC eventually gets below 0 when combined with FS in general. Hence one needs to set an upper bound ($m = 50$ in our simulation) to avoid selecting the full model. Lastly, we observe that the CV error and EBIC curves of SCAD fluctuate with decreasing penalty. Due to the non-convexity of the SCAD penalty, many variables entering the model in early stages can be dropped later. Nonetheless, the final SCAD model chosen by minimizing CV tends to be small in size with decent number of relevant variables. These observations are also confirmed by the summaries of the results over the 100 replicates reported in Table 4, with complete results presented in Table 5. The SIS results are not listed in Table 4 because the only scenario where SIS (combined with any criterion) has slight advantage over the other three methods is SNR =1 and $\rho = 0$. Besides, as discussed in Section 2.2.2, BETA ="half-half" is a very special setting that favors both LASSO and FS. We concentrate on the other three settings of BETA. Because of their similarities, in Table 4, the results are averaged over BETA for brevity.

Another notable observation from Figure 6 is that the CV error curve for a sequence of models may not be taken as a proxy for the corresponding MSPE curve. For example, the CV error and MSPE for the model sequence generated by Forward Selection procedure (upper left panel) exhibit a very different pattern after all relevant variables are included. This discrepancy between CV curve and MSPE curve is even more obvious for the sequence of models generated by SIS, whereas it is less severe for LL and LO. The numerical result calls for further investigation of the effectiveness and validity of the cross-validation procedure in high dimensional setting as it is widely used in practice for model evaluation and selection. In addition, CV error is often taken as a performance measure in the absence of independent test data for comparison of competing methods. The observed discrepancy cautions against this practice.

Next we discuss findings about CV from the simulation study in detail with graphical displays of the case with SNR = 10 in Figure 7. In terms of selection accuracy, LO selects about the same number of relevant variables as LL and FS with a considerably smaller model in general. On the other hand, LO almost always select more relevant variables than SCAD and the model size is only slightly larger, as illustrated by the left and middle panels of Figure 7. When the minimum value of CV criterion is used to pick the final model, LO_{min} selects significantly smaller models than both FS_{min} and LL_{min}, while missing only less than one relevant variable on average. The average loss in the number of selected relevant variables T is in the range of 0.5 to 1.3. The best performance of LO_{min} is achieved when SNR $= 10$ and $\rho = 0$, where the average model size is 23 with 9.5 relevant variables. LO_{min} eliminates almost 40 more irrelevant variables than LL_{min}, let alone FS_{min}. By contrast, FS_{min} tends to select extremely large models (of nearly the same size as training data for cross-validation). The conciseness of the model selected by LO is also confirmed in other simulation scenarios summarized in Table 4.

The right panels of Figure 7 show the average MSPE for all levels of ρ in the case of SNR $= 5$ and 10. In terms of the prediction accuracy of the final model, the models selected by LO have MSPE values close to those picked by LL and smaller MSPE when $\rho = 0$. As correlation among predictors increases, LL turns to perform slightly better, due to the increased model size. Overall, the model selected by LO is as competitive as the one chosen by LL in terms of prediction accuracy.

In our study, we observed that EBIC typically yields a model of very small size, and its effectiveness depends heavily on the combination of SNR and the relative relationship among n, p and q. The models selected by EBIC miss at least 4 or more true predictors due to its preference for a small model size, as highlighted in Figure 6. The only exception is the scenario with SNR $= 10$ and $\rho = 0$, where FS_{EBIC} selects almost perfectly a model with both the model size $|\mathcal{M}|$ and the number of selected relevant variables T close to 10.

Instead of selecting a model with the minimum value of CV criterion, applying the "one-standard error" rule can further decrease the model size by a notable margin for LASSO, but as a result, it would miss out slightly less than one relevant variable on average. Figure 7 gives a summary of the comparison between the minimum rule and the "one-standard error" rule, while Table 4 contains detailed results of $|\mathcal{M}|$, T and $MSPE$ of the models selected by the "one-standard error" rule. It is clear that this "one-se" rule is only effective for LL and LO, but not for FS and $SCAD$. On average, the differences in $|\mathcal{M}|$ and T between LO_{min} and LO_{1se} are roughly 6 to 10 and 0.5 to 1, respectively, under all combinations of factor levels, while the difference in $MSPE$ is almost trivial. In general, it seems safe to apply this rule if the goal is to establish a model as concise as possible.

4 Conclusions

In this paper, we have carefully distinguished two aspects of variable selection: screening and model building and have conducted comprehensive numerical studies to compare the performance of four methods: LASSO, SIS, FS and SCAD for screening and model building separately, with EBIC and CV error as evaluation criteria for selecting final models.

Our studies indicate that generally LASSO has a better screening performance with a notable margin over SIS, FS and SCAD, given a reasonably large signal-to-noise ratio. For model building, it is demonstrated that the proposed LASSO-OLS hybrid method can significantly reduce the number of irrelevant variables while keeping most of the relevant ones with competitive MSPE, bettering the standard LASSO with CV. Meanwhile, EBIC is found to be an unstable criterion in high-dimensional setting with its effectiveness being sensitive to the change in values of p and n, when SNR is weak and dependence among predictors is present.

The studies also call for some caution in using cross validation for model evaluation and selection in high dimensional setting. The result shows that there could be large discrepancy between the CV error and the MSPE for certain model building and selection procedures. More studies are needed to investigate the effectiveness of CV in different settings numerically and to formalize proper conditions for the validity of CV theoretically.

Lastly, we have clarified different findings and conclusions from relevant comparisons in the literature, pointing out the underlying reasons. Differential merits discussed in the studies can be used in practice to choose appropriate variable selection methods and evaluation criteria for a broad range of situations.

Acknowledgments

Shi's research is partially supported by NSF grants DMS-1007060 and DMS-1308458. Lee's research is supported in part by NSA grant H98230-10-1-0202 and NSF grant DMS-1209194. The authors would like to thank Dr. Angela Dean and Dr. Jason Hsu for helpful discussion.

Appendix: MSPE Calculation for a Final Model

The Mean Squared Prediction Error (MSPE) serves as an evaluation criterion for the prediction accuracy of a model. It is defined as

$$\text{MSPE} = \frac{1}{n}\,\text{E}[(\mathbf{Y}_{new} - \hat{\mathbf{Y}})^T(\mathbf{Y}_{new} - \hat{\mathbf{Y}})|\mathbf{X}], \tag{4.1}$$

where $\mathbf{Y}_{new} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}$ is a vector of n new observations at fixed design points $\mathbf{X}$, and $\hat{\mathbf{Y}}$ is the vector of predicted values from the given model. Suppose that the model being evaluated involves r variables with coefficients $\hat{\boldsymbol{\beta}}$, and denote the corresponding partial design matrix by $\mathbf{X_r}$. Then $\hat{\mathbf{Y}} = \mathbf{X_r}\hat{\boldsymbol{\beta}}$ is a prediction of $\mathbf{Y}_{new}$ based on the model.

For the OLS fit with coefficients $\hat{\boldsymbol{\beta}}_{\text{OLS}}$,

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X_r}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X_r}(\mathbf{X_r}^T\mathbf{X_r})^{-1}\mathbf{X_r}^T\mathbf{Y} \\ &= \mathbf{X_r}(\mathbf{X_r}^T\mathbf{X_r})^{-1}\mathbf{X_r}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \equiv \mathbf{H_r}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}).\end{aligned} \tag{4.2}$$

Therefore,

$$\begin{aligned}\text{MSPE} &= \frac{1}{n}\,\text{E}[((\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}) - \mathbf{H_r}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}))^T((\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}) - \mathbf{H_r}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}))] \\ &= \frac{1}{n}\,\text{E}[((\mathbf{I} - \mathbf{H_r})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new} - \mathbf{H_r}\boldsymbol{\epsilon})^T((\mathbf{I} - \mathbf{H_r})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new} - \mathbf{H_r}\boldsymbol{\epsilon})] \\ &= \frac{1}{n}(((\mathbf{I} - \mathbf{H_r})\mathbf{X}\boldsymbol{\beta})^T((\mathbf{I} - \mathbf{H_r})\mathbf{X}\boldsymbol{\beta}) + \text{E}[(\mathbf{H_r}\boldsymbol{\epsilon})^T\mathbf{H_r}\boldsymbol{\epsilon}] + \text{E}[\boldsymbol{\epsilon}_{new}^T\boldsymbol{\epsilon}_{new}]) \\ &= \frac{1}{n}((\mathbf{X}\boldsymbol{\beta})^T(\mathbf{I} - \mathbf{H_r})\mathbf{X}\boldsymbol{\beta} + r\sigma^2 + n\sigma^2) \\ &= \frac{1}{n}\boldsymbol{\beta}_{(1)}^T\mathbf{X}_{(1)}^T(\mathbf{I} - \mathbf{H_r})\mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \frac{r}{n}\sigma^2 + \sigma^2.\end{aligned}$$

The first term is due to the bias in estimation of the mean regression function, and it degenerates to 0 if and only if the r variables contain the first q relevant variables. The second term is the estimation variance, and it implies that MSPE linearly increases as irrelevant variables are added to the model. The third term is the irreducible prediction error due to ε.

Such analytical expression of MSPE is infeasible for LASSO since its explicit solution is unavailable in general. Instead, we approximate it by a Monte Carlo estimate described as follows.

1. For fixed $\mathbf{X}$, generate $\boldsymbol{\epsilon}$ from $N(0, \sigma^2)$ and let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Similarly, generate $\boldsymbol{\epsilon}_{new}$ from $N(0, \sigma^2)$ and let $\mathbf{Y}_{new} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}$.
2. Given λ, find the LASSO estimator $\hat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}}$ based on $\mathbf{X}$ and $\mathbf{Y}$. Compute $\hat{\mathbf{Y}} = \mathbf{X}_{\mathbf{r}}\hat{\boldsymbol{\beta}}^{\lambda}_{\text{LASSO}}$ and $\widehat{\text{MSPE}} = \frac{1}{n}(\mathbf{Y}_{new} - \hat{\mathbf{Y}})^T(\mathbf{Y}_{new} - \hat{\mathbf{Y}})$.
3. Repeat the above two steps m times and estimate MSPE by $\frac{1}{m}\sum_{i=1}^{m}\widehat{\text{MSPE}}_i$. $m = 100$ was used in our study.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics* **37**(4): 373–384.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics* **35**: 2313–2351.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces, *Biometrika* **95**(3): 759–771.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society. Series B* **70**: 849–911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics* **32**(3): 928–961.

Genovese, C. R., Jin, J. and Wasserman, L. (2009). Revisiting marginal regression. arXiv:0911.4080.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, New York, NY, USA.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models, *Statistica Sinica* **18**(4): 1603–1618.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection, *Statistica Sinica* **16**: 1273–1284.

Levina, E. and Zhu, J. (2008). Discussion of "sure independence screening for ultrahigh dimensional feature space", *Journal of the Royal Statistical Society. Series B* **70**: 897–898.

Meinshausen, N. (2007). Relaxed Lasso, *Computational Statistics & Data Analysis* **52**(1): 374–393.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics* **34**: 1436–1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data, *The Annals of Statistics* **37**(1): 246–270.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B* **58**: 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening, *Journal of the American Statistical Association* **104**(488): 1512–1524.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator, *Journal of the Royal Statistical Society. Series B* **69**: 143–161.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**: 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal Of The Royal Statistical Society Series B* **67**(2): 301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters, *The Annals of Statistics* **37**(4): 1733–1751.

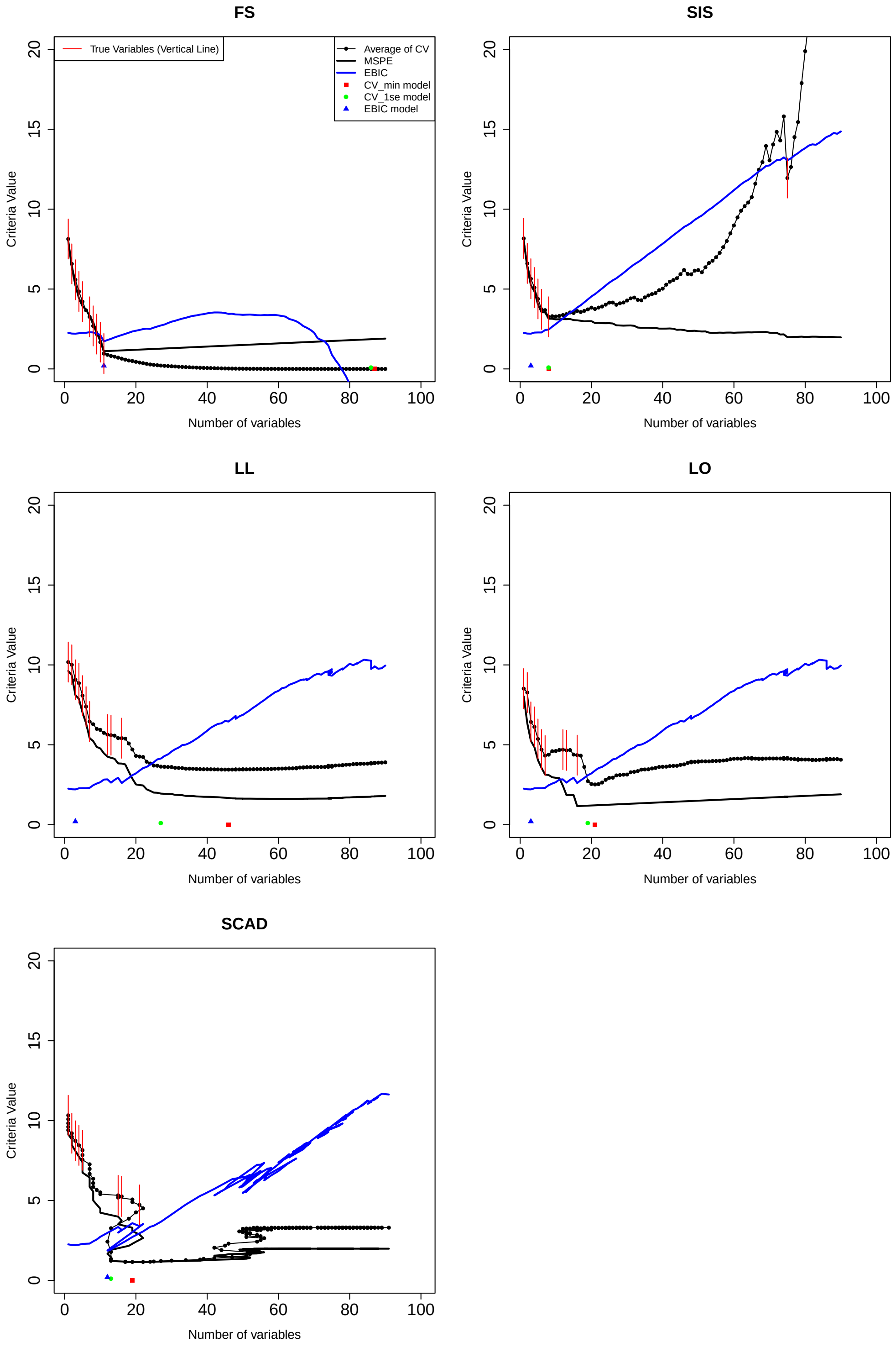
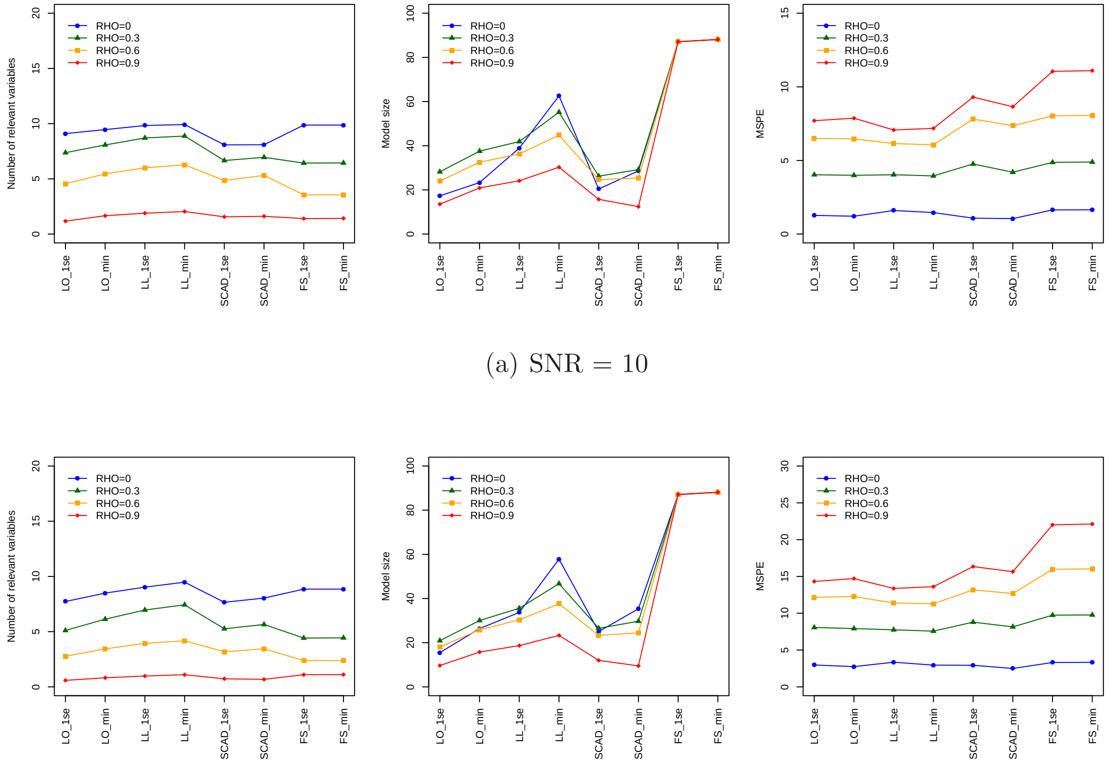


Figure 6: Plots of model evaluation criteria for a sample with SNR $= 10$, $\rho = 0$ and BETA =“pure”.



(a) SNR = 10

(b) SNR = 5

Figure 7: The average number of relevant variables, final model size and MSPE of LO, LL, FS and SCAD when SNR = 5 and 10.

Table 4: A summary of the average model size $|\mathcal{M}|$, the number of selected relevant variables T and MSPE of the models selected by LO_{min}, LL_{min}, FS_{min}, $SCAD_{min}$, LO_{1se}, LL_{1se}, FS_{1se}, and $SCAD_{1se}$, respectively, for SNR = 5 and SNR = 10 scenarios. The standard errors are attached in parentheses.

		SNR = 10			SNR = 5		
ρ	Method	$\|\mathcal{M}\|$	T	MSPE	$\|\mathcal{M}\|$	T	MSPE
0	LO_{min}	23.31 (0.73)	9.46 (0.06)	1.21 (0.02)	26.38 (1.1)	8.49 (0.11)	2.75 (0.04)
	LL_{min}	62.58 (0.79)	9.92 (0.02)	1.46 (0.01)	57.83 (0.91)	9.48 (0.05)	2.95 (0.02)
	FS_{min}	88.18 (0.03)	9.87 (0.04)	1.65 (< 0.01)	88.16 (0.03)	8.85 (0.1)	3.33 (< 0.01)
	$SCAD_{min}$	28.6 (0.62)	8.09 (0.16)	1.05 (< 0.01)	35.36 (0.55)	8.04 (0.16)	2.5 (0.05)
	LO_{1se}	17.3 (0.44)	9.1 (0.07)	1.27 (0.03)	15.5 (0.48)	7.73 (0.11)	2.97 (0.06)
	LL_{1se}	38.91 (0.59)	9.83 (0.03)	1.6 (0.01)	33.72 (0.64)	9.03 (0.08)	3.35 (0.05)
	FS_{1se}	87.09 (0.05)	9.87 (0.04)	1.64 (< 0.01)	87.16 (0.05)	8.85 (0.11)	3.31 (< 0.01)
	$SCAD_{1se}$	20.4 (0.49)	8.08 (0.16)	1.08 (0.01)	25.04 (0.6)	7.66 (0.19)	2.93 (0.1)
0.3	LO_{min}	37.53 (1.03)	8.07 (0.09)	3.99 (0.03)	30.04 (0.97)	6.13 (0.08)	7.92 (0.04)
	LL_{min}	55.18 (0.8)	8.88 (0.06)	3.95 (0.02)	46.72 (0.79)	7.43 (0.09)	7.57 (0.03)
	FS_{min}	88.14 (0.03)	6.44 (0.12)	4.89 (< 0.01)	88.07 (0.04)	4.44 (0.1)	9.77 (0.01)
	$SCAD_{min}$	29.2 (0.56)	6.95 (0.19)	4.2 (0.05)	29.72 (0.55)	5.65 (0.17)	8.15 (0.04)
	LO_{1se}	28.18 (0.69)	7.37 (0.11)	4.03 (0.03)	20.92 (0.66)	5.11 (0.06)	8.07 (0.05)
	LL_{1se}	41.89 (0.5)	8.7 (0.06)	4.03 (0.01)	35.58 (0.44)	6.96 (0.06)	7.75 (0.02)
	FS_{1se}	87.04 (0.05)	6.43 (0.12)	4.87 (< 0.01)	87.05 (0.05)	4.42 (0.05)	9.74 (0.01)
	$SCAD_{1se}$	26.22 (0.61)	6.66 (0.19)	4.76 (0.07)	26.52 (0.55)	5.26 (0.16)	8.79 (0.06)
0.6	LO_{min}	32.51 (0.91)	5.45 (0.1)	6.47 (0.03)	25.86 (0.84)	3.44 (0.11)	12.27 (0.05)
	LL_{min}	44.78 (0.75)	6.27 (0.09)	6.06 (0.02)	37.64 (0.67)	4.17 (0.08)	11.29 (0.04)
	FS_{min}	88.15 (0.03)	3.56 (0.1)	8.06 (< 0.01)	88.12 (0.04)	2.38 (0.11)	16.03 (0.01)
	$SCAD_{min}$	25.38 (0.9)	5.31 (0.2)	7.36 (0.04)	24.44 (0.72)	3.45 (0.17)	12.68 (0.05)
	LO_{1se}	24.01 (0.69)	4.55 (0.11)	6.5 (0.03)	18.03 (0.59)	2.77 (0.12)	12.17 (0.04)
	LL_{1se}	36.3 (0.42)	6 (0.09)	6.15 (0.01)	30.31 (0.37)	3.94 (0.08)	11.4 (0.02)
	FS_{1se}	87.14 (0.05)	3.55 (0.1)	8.03 (0.01)	87.12 (0.05)	2.38 (0.11)	15.97 (0.01)
	$SCAD_{1se}$	24.64 (0.84)	4.85 (0.18)	7.81 (0.06)	23.34 (0.61)	3.18 (0.16)	13.18 (0.07)
0.9	LO_{min}	20.88 (0.71)	1.66 (0.07)	7.87 (0.03)	15.76 (0.68)	0.82 (0.09)	14.71 (0.07)
	LL_{min}	30.28 (0.73)	2.04 (0.08)	7.18 (0.04)	23.32 (0.67)	1.1 (0.09)	13.61 (0.09)
	FS_{min}	88.07 (0.03)	1.42 (0.06)	11.1 (< 0.01)	88.12 (0.03)	1.11 (0.08)	22.12 (< 0.01)
	$SCAD_{min}$	12.41 (0.98)	1.61 (0.16)	8.65 (0.04)	9.5 (0.76)	0.68 (0.1)	15.65 (0.05)
	LO_{1se}	13.54 (0.43)	1.17 (0.06)	7.7 (0.02)	9.66 (0.33)	0.59 (0.04)	14.32 (0.03)
	LL_{1se}	24.13 (0.29)	1.89 (0.07)	7.07 (0.01)	18.67 (0.29)	0.98 (0.05)	13.37 (0.02)
	FS_{1se}	87.11 (0.05)	1.4 (0.06)	11.05 (< 0.01)	87.11 (0.05)	1.1 (0.06)	22.01 (0.01)
	$SCAD_{1se}$	15.72 (0.88)	1.56 (0.14)	9.3 (0.05)	11.98 (0.65)	0.73 (0.09)	16.34 (0.06)

Table 5: A summary of the average model size $|\mathcal{M}|$ and the number of selected relevant variables T.

SNR = 1

BETA	Method	ρ=0		ρ=0.3		ρ=0.6		ρ=0.9	
		$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T
Pure	FS_{EBIC}	1.08	0.72	2.09	0.37	1.56	0.11	1	0.05
	FS_{min}	88.14	3.28	9.15	0.83	5.16	0.28	2.22	0.07
	FS_{1se}	87.21	3.27	87.09	1.49	86.96	1	87.05	0.99
	LO_{EBIC}	1.06	0.69	1.64	0.29	1.33	0.1	1	0.05
	LL_{min}	24.17	4.02	88.17	1.49	88.12	1.01	88.08	0.99
	LL_{1se}	10.03	2.74	22.94	1.89	17.58	0.85	9.63	0.26
	LO_{min}	17.18	3.19	31.29	2.21	24.79	0.95	15.41	0.34
	LO_{1se}	5.92	1.99	12.66	1.21	9.97	0.49	4.64	0.11
	SIS_{EBIC}	1.06	0.7	1.56	0.28	1.3	0.1	1	0.05
	SIS_{min}	29.2	4.91	20.05	1.57	15	0.73	8.78	0.24
	SIS_{1se}	23.65	4.55	6.74	0.7	3.99	0.22	1.81	0.07
	$SCAD_{EBIC}$	1.32	0.68	1.76	0.28	1.67	0.14	1.24	0.05
	$SCAD_{min}$	17.11	2.94	25.53	1.93	16.21	0.77	5.91	0.19
	$SCAD_{1se}$	6.58	1.96	20.92	1.65	14.35	0.74	6.81	0.22
Two-levels	FS_{EBIC}	1.09	0.78	2.09	0.44	1.65	0.11	1	0.01
	FS_{min}	88.11	3.65	88.17	1.59	88.1	1.14	88.11	0.89
	FS_{1se}	86.99	3.63	87.08	1.57	87.2	1.14	87.1	0.88
	LO_{EBIC}	1.08	0.82	1.73	0.39	1.44	0.13	1	0.01
	LL_{min}	24.79	4.24	32.3	2.41	25.47	1.09	14.67	0.17
	LL_{1se}	11.13	3.25	22.63	2.13	17.01	0.9	9.69	0.13
	LO_{min}	13.7	2.8	20.55	1.79	15.47	0.68	8.71	0.13
	LO_{1se}	6.18	2.05	11.21	1.27	7.96	0.46	4.07	0.11
	SIS_{EBIC}	1.07	0.77	1.64	0.37	1.39	0.11	1	0.01
	SIS_{min}	27.64	4.72	8.36	1.12	4.81	0.33	2.2	0.08
	SIS_{1se}	22.14	4.47	6.53	0.96	3.84	0.27	1.8	0.05
	$SCAD_{EBIC}$	1.27	0.77	1.83	0.38	1.74	0.13	1.19	0.01
	$SCAD_{min}$	19.03	3.61	24.35	2.12	15.73	0.82	5.08	0.1
	$SCAD_{1se}$	7.93	2.54	19.59	1.96	13.89	0.75	6.4	0.11
Mixed	FS_{EBIC}	1.04	0.68	1.57	0.5	1.38	0.14	1	0.05
	FS_{min}	88.13	3.33	88.12	2.12	88.02	1.33	88.18	0.89
	FS_{1se}	87.22	3.32	87.14	2.1	86.98	1.32	87.06	0.88
	LO_{EBIC}	1.02	0.74	1.39	0.39	1.23	0.11	1	0.07
	LL_{min}	23.08	4.11	29.21	2.83	22.29	1.39	15.89	0.45
	LL_{1se}	9.29	2.79	19.1	2.39	16.19	1.16	9.62	0.31
	LO_{min}	17.68	3.05	20.3	2.12	14.01	1.01	8.7	0.28
	LO_{1se}	6.46	1.98	8.79	1.34	8.03	0.64	4.48	0.2
	SIS_{EBIC}	1.02	0.66	1.34	0.4	1.19	0.12	1	0.05
	SIS_{min}	28.4	5.01	6.71	1.18	4.71	0.39	2.22	0.14
	SIS_{1se}	21.9	4.47	5.53	1.05	3.68	0.35	1.79	0.12
	$SCAD_{EBIC}$	1.3	0.7	1.69	0.42	1.57	0.11	1.22	0.07
	$SCAD_{min}$	15.79	3.24	21.18	2.32	14.63	1.05	5.98	0.24
	$SCAD_{1se}$	6.41	1.97	15.92	1.96	13.21	0.94	6.46	0.24
Half-half	FS_{EBIC}	1.01	0.69	1.26	0.83	1.47	0.83	1.84	1
	FS_{min}	88.06	3.55	88.12	3.71	88.12	3.3	88.12	3.66
	FS_{1se}	87.09	3.55	86.8	3.7	87.13	3.3	87.15	3.65
	LO_{EBIC}	1.01	0.7	1.15	0.77	1.24	0.74	1.55	0.97
	LL_{min}	29.14	4.54	32.16	4.71	27.12	4.19	24.43	4.03
	LL_{1se}	11.37	3.04	13.19	3.31	10.68	2.88	9.46	2.71
	LO_{min}	18.54	3.32	23.11	3.68	20.63	3.2	16.38	2.76
	LO_{1se}	7.93	2.17	8.74	2.36	7.6	2.12	4.5	1.58
	SIS_{EBIC}	1.01	0.69	1.08	0.68	1.05	0.6	1.07	0.62
	SIS_{min}	29.77	5.05	17.32	3.64	10.18	2.17	6.05	1.21
	SIS_{1se}	22.74	4.5	13.55	3.2	8.24	1.92	4.46	1.17
	$SCAD_{EBIC}$	1.22	0.71	1.34	0.75	1.55	0.73	1.54	0.91
	$SCAD_{min}$	16.63	3.18	14.53	3.29	5.86	1.9	2.05	1.05
	$SCAD_{1se}$	6.76	2.07	6.52	2.2	2.73	1.06	1.39	0.79

SNR = 5

BETA	Method	ρ=0		ρ=0.3		ρ=0.6		ρ=0.9	
		$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T	$\|\mathcal{M}\|$	T
Pure	FS_{EBIC}	5.82	5.63	3.93	1.5	3.12	0.55	1.84	0.07
	FS_{min}	88.23	9.21	88.1	3.68	88.19	2.05	88.12	1.06
	FS_{1se}	87.08	9.21	87	3.68	87.02	2.04	87.11	1.04
	LO_{EBIC}	2.12	2.02	2.99	1.16	2.71	0.48	1.59	0.1
	LL_{min}	59.17	9.74	48.55	7.46	38.23	3.71	24.58	1.01
	LL_{1se}	35.01	9.38	40.04	7.14	32.12	3.64	18.88	0.87
	LO_{min}	30.68	9.16	34.35	6.41	26.59	3.1	16.98	0.8
	LO_{1se}	18.7	8.48	25.22	5.22	19.95	2.53	9.97	0.55
	SIS_{EBIC}	1.45	1.4	2.76	0.98	2.62	0.42	1.55	0.08
	SIS_{min}	31.73	7.82	15.52	2.85	10.43	1.25	4.56	0.26
	SIS_{1se}	24.68	7.41	12.21	2.51	8.15	1.04	3.58	0.2
	$SCAD_{EBIC}$	2.37	1.73	3.1	0.93	2.77	0.4	1.79	0.1
	$SCAD_{min}$	34.92	8.13	33.63	5.62	26.88	3.27	10.09	0.64
	$SCAD_{1se}$	23.57	7.7	30.61	5.39	26.41	3.02	12.55	0.66
Two-levels	FS_{EBIC}	5.47	5.39	4.01	1.9	3.06	0.63	1.86	0.12
	FS_{min}	88.15	8.54	88.09	3.87	88.15	2.03	88.09	0.97
	FS_{1se}	87.25	8.54	87.16	3.84	87.24	2.03	87.11	0.96
	LO_{EBIC}	3.88	3.79	3.17	1.52	2.58	0.54	1.6	0.12
	LL_{min}	55.52	8.99	45.02	6.74	37.42	3.81	22.9	0.89
	LL_{1se}	32.34	8.53	36.03	6.41	31.35	3.76	18.93	0.86
	LO_{min}	21.78	7.46	28.04	5.37	25.75	3.24	15.98	0.69
	LO_{1se}	11.48	6.53	19.31	4.44	18.21	2.51	10.01	0.49
	SIS_{EBIC}	3.14	2.96	2.95	1.44	2.42	0.52	1.55	0.11
	SIS_{min}	26.77	6.51	13.41	2.83	10.27	1.29	4.78	0.21
	SIS_{1se}	21.41	6.31	11.32	2.71	8.24	1.14	3.61	0.2
	$SCAD_{EBIC}$	3.91	3.17	3.22	1.25	2.76	0.44	1.92	0.12
	$SCAD_{min}$	35.19	8	30.23	5.12	25.5	3.35	10.83	0.66
	$SCAD_{1se}$	26.63	7.72	27.79	4.91	24.33	2.99	13.34	0.72
Mixed	FS_{EBIC}	4.65	4.51	3.26	2.35	2.66	1.07	1.63	0.28
	FS_{min}	88.11	8.8	88.03	5.76	88.03	3.07	88.14	1.31
	FS_{1se}	87.15	8.8	86.99	5.75	87.09	3.06	87.11	1.29
	LO_{EBIC}	2.11	2	2.66	1.77	2.37	0.89	1.38	0.23
	LL_{min}	58.79	9.72	46.58	8.09	37.26	5	22.49	1.39
	LL_{1se}	33.82	9.19	30.66	7.33	27.47	4.43	18.2	1.21
	LO_{min}	26.69	8.85	27.72	6.62	25.25	3.99	14.31	0.97
	LO_{1se}	16.31	8.17	18.23	5.66	15.94	3.27	8.99	0.72
	SIS_{EBIC}	1.55	1.48	2.45	1.63	2.3	0.77	1.36	0.25
	SIS_{min}	26.78	7.72	10.11	3.59	8.68	1.93	4.33	0.47
	SIS_{1se}	22.3	7.53	8.43	3.33	7.06	1.71	3.51	0.43
	$SCAD_{EBIC}$	2.28	1.66	2.67	1.43	2.52	0.73	1.55	0.24
	$SCAD_{min}$	35.96	7.98	25.31	6.22	20.94	3.72	7.57	0.75
	$SCAD_{1se}$	24.93	7.55	21.16	5.47	19.27	3.53	10.04	0.82
Half-half	FS_{EBIC}	5.51	5.28	4.45	4.29	5.05	4.75	5.58	5.32
	FS_{min}	88.16	9.14	88.1	8.76	88.11	8.88	88.1	9.27
	FS_{1se}	87.1	9.14	87.03	8.76	87.27	8.88	87	9.27
	LO_{EBIC}	2.03	1.9	2.16	2.01	2.27	2.02	2.63	2.38
	LL_{min}	62.13	9.71	59.61	9.65	61.81	9.56	60.29	9.61
	LL_{1se}	37.23	9.36	35.75	9.23	35.57	9.05	38.01	9.33
	LO_{min}	30.35	9.11	30.93	8.98	30.3	8.85	27.86	8.98
	LO_{1se}	18.79	8.51	17.45	7.85	18.38	8.1	18.22	8.39
	SIS_{EBIC}	1.46	1.36	1.43	1.33	1.31	1.16	1.17	1.05
	SIS_{min}	31.76	7.87	19.78	5.91	15.73	4.21	9.15	2.55
	SIS_{1se}	25.22	7.57	16.08	5.53	12.88	3.84	7.82	2.43
	$SCAD_{EBIC}$	2.35	1.61	3.82	2.79	5.93	4.74	4.66	2.97
	$SCAD_{min}$	35.93	7.93	24.63	7.79	13.86	6.94	8.93	4.67
	$SCAD_{1se}$	24.94	7.5	16.23	6.9	8.87	5.93	5.69	3.39

SNR = 10

BETA	Method	ρ=0		ρ=0.3		ρ=0.6		ρ=0.9	
		$\lvert\mathcal{M}\rvert$	T	$\lvert\mathcal{M}\rvert$	T	$\lvert\mathcal{M}\rvert$	T	$\lvert\mathcal{M}\rvert$	T
Pure	FS_{EBIC}	9.63	9.19	5.04	2.5	3.89	1.05	2.1	0.2
	FS_{min}	88.19	9.79	88.09	5.21	88.13	2.81	88.09	1.27
	FS_{1se}	87.06	9.79	87.15	5.2	87.05	2.8	87.11	1.26
	LO_{EBIC}	5.46	4.83	3.53	1.4	3.24	0.7	2.1	0.22
	LL_{min}	64.86	10	55.62	8.88	45.6	5.88	30.42	1.52
	LL_{1se}	40.64	9.98	44.45	8.9	38.72	5.83	24.32	1.44
	LO_{min}	25.21	9.92	42.41	8.3	33.52	5.09	20.68	1.21
	LO_{1se}	19.9	9.77	32.01	7.59	26.8	4.34	13.91	0.86
	SIS_{EBIC}	2.2	2.11	3.24	1.25	3.08	0.66	2.05	0.22
	SIS_{min}	30.97	8.23	18.9	3.55	13.58	1.81	6.54	0.46
	SIS_{1se}	27.82	8.14	15.64	3.24	10.81	1.51	5.3	0.37
	$SCAD_{EBIC}$	8.23	5.13	3.42	1.17	3.48	0.67	2.2	0.2
	$SCAD_{min}$	24.48	7.66	32.51	6.76	29.12	5.13	12.13	1.13
	$SCAD_{1se}$	16.87	7.66	31.53	6.56	28.65	4.84	16.21	1.11
Two-levels	FS_{EBIC}	9.38	9.2	5.54	3.76	3.94	1.29	2.18	0.25
	FS_{min}	88.17	9.9	88.23	5.75	88.12	3.09	88.06	1.3
	FS_{1se}	87.01	9.9	86.92	5.73	87.13	3.08	87.26	1.29
	LO_{EBIC}	5.85	5.57	4.02	2.36	3.29	0.95	2.05	0.2
	LL_{min}	58.69	9.8	51.36	8.19	44.07	5.61	30.26	1.83
	LL_{1se}	35.94	9.56	41.69	7.96	38	5.45	24.75	1.71
	LO_{min}	20.6	8.61	35.62	7.24	35.42	5.16	21.4	1.58
	LO_{1se}	13.57	7.89	26.95	6.47	25.27	4.1	14.82	1.15
	SIS_{EBIC}	4.1	3.84	3.47	1.89	3.07	0.85	2.06	0.18
	SIS_{min}	25.92	6.78	17.18	3.63	12.36	1.82	6.33	0.52
	SIS_{1se}	19.76	6.49	14.72	3.47	10.71	1.72	5.39	0.41
	$SCAD_{EBIC}$	3.91	3.17	3.22	1.25	2.76	0.44	1.92	0.12
	$SCAD_{min}$	35.19	8	30.23	5.12	25.5	3.35	10.83	0.66
	$SCAD_{1se}$	26.63	7.72	27.79	4.91	24.33	2.99	13.34	0.72
Mixed	FS_{EBIC}	9.58	9.29	4.94	4.06	3.29	1.82	2	0.44
	FS_{min}	88.19	9.92	88.1	8.36	88.19	4.79	88.07	1.68
	FS_{1se}	87.2	9.92	87.06	8.36	87.23	4.76	86.96	1.65
	LO_{EBIC}	5.91	5.1	3	2.11	2.82	1.41	1.99	0.36
	LL_{min}	64.2	9.97	58.55	9.58	44.67	7.31	30.16	2.77
	LL_{1se}	40.16	9.94	39.52	9.24	32.17	6.71	23.32	2.51
	LO_{min}	24.11	9.84	34.55	8.66	28.6	6.11	20.57	2.19
	LO_{1se}	18.43	9.65	25.59	8.06	19.96	5.21	11.89	1.51
	SIS_{EBIC}	1.97	1.92	2.64	1.8	2.65	1.23	1.94	0.36
	SIS_{min}	31.91	8.44	12.39	4.25	10.67	2.68	5.52	0.87
	SIS_{1se}	27.9	8.28	9.84	3.87	8.49	2.38	4.53	0.73
	$SCAD_{EBIC}$	9.28	6.32	2.95	1.42	2.91	1.34	2.07	0.34
	$SCAD_{min}$	24.56	8.16	26.56	7.78	19.25	5.9	12.04	2.08
	$SCAD_{1se}$	17.12	8.16	20.92	7.28	18	5.18	14.36	2.05
Half-half	FS_{EBIC}	9.57	9.17	10.13	9.65	9.83	9.2	10.09	9.64
	FS_{min}	88.11	9.71	88.15	9.96	88.14	9.66	88.17	9.93
	FS_{1se}	87.03	9.71	87.03	9.96	86.98	9.66	86.97	9.93
	LO_{EBIC}	5.68	4.96	5.95	4.98	5.73	4.8	6.24	5.19
	LL_{min}	65.05	9.96	62.97	10	66.12	9.95	62.27	9.95
	LL_{1se}	40.58	9.92	39.46	9.91	41.2	9.87	39.5	9.93
	LO_{min}	24.86	9.83	26.03	9.93	24.98	9.7	25.94	9.83
	LO_{1se}	18.92	9.59	18.97	9.66	19.57	9.48	19.04	9.56
	SIS_{EBIC}	1.81	1.79	1.95	1.81	1.3	1.21	1.24	1.16
	SIS_{min}	32.85	8.36	23.51	6.46	16.76	4.59	8.28	2.87
	SIS_{1se}	28.11	8.16	20.08	6.17	15.54	4.37	7.94	2.84
	$SCAD_{EBIC}$	9.67	6.65	9.01	6.9	9.59	7.33	7.86	5.35
	$SCAD_{min}$	23.8	8.04	15.51	8.19	11.96	7.65	10.34	6.36
	$SCAD_{1se}$	15.97	8.01	11.19	8.19	10.05	7.63	8.63	5.59