A COMPARISON OF VARIABLE SELECTION METHODS FOR BLOCKWISE DIAGONAL DESIGNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Lasso is a celebrated method for variable selection in linear models, but it faces challenges when the variables are moderately or strongly correlated. This motivates alternative approaches such as using a non-convex penalty, adding a ridge regularization, or conducting a post-Lasso thresholding. In this paper, we compare Lasso with 5 other methods: Elastic net, SCAD, forward selection, thresholded Lasso, and forward backward selection. We measure their performances theoretically by the expected Hamming error, assuming that the regression coefficients are *iid* drawn from a two-point mixture and that the Gram matrix is block-wise diagonal. By deriving the rates of convergence of Hamming errors and the phase diagrams, we obtain useful conclusions about the pros and cons of different methods.

1 INTRODUCTION

Variable selection is one of the core problems in high-dimensional data analysis. Consider a linear regression, where the response $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n\times p}$ satisfy that

$$y = X\beta + z, \qquad \mathrm{diag}(X'X) = I_p, \qquad z \sim \mathcal{N}(0, \sigma^2 I_n). \tag{1}$$

The goal is estimating the support of β ($\mathrm{Supp}(\beta)$). Lasso (Tibshirani, 1996) is a popular method:

$$\hat{\beta}^{\mathrm{lasso}} = \mathrm{argmin}_\beta\big\{\|y - X\beta\|^2/2 + \lambda\|\beta\|_1\big\}. \tag{2}$$

Lasso has good rates of convergence on the L_q-estimation error or prediction error (Bickel et al., 2009). However, it can be unsatisfactory for variable selection, especially when the columns in the design matrix are moderately or strongly correlated. Zhao & Yu (2006) showed that an *irrepresentable condition* on X is necessary for Lasso to recover $\mathrm{Supp}(\beta)$ with high probability, and such a condition is restrictive when p is large (Fan & Lv, 2010). Ji & Jin (2012) studied the Hamming error of Lasso and revealed Lasso's non-optimality by lower-bounding its Hamming error rate.

Many alternative approaches were proposed for variable selection. The first is non-convex penalization methods, which replace the L^1-penalty in (2) by a non-convex penalty. Examples include SCAD (Fan & Li, 2001), MCP (Zhang et al., 2010), and truncated L^1-penalty (Shen et al., 2012). Compared with Lasso, the non-convex penalization methods require weaker conditions on the design matrix for recovering $\mathrm{Supp}(\beta)$ (Fan & Lv, 2011). The second approach adds an additional L^2-penalty term in (2) and is called Elastic net (Zou & Hastie, 2005). It was observed in Zou & Hastie (2005) that Elastic net encourages a grouping effect, that is, strongly correlated variables tend to be selected or eliminated together, and numerically improves Lasso in some settings. The third approach conducts post-processing on the Lasso estimator. For example, the adaptive Lasso (Zou, 2006) uses $\hat{\beta}^{\mathrm{lasso}}$ to further modify the L^1-penalty, and the thresholded Lasso (Zhou, 2009) applies an entry-wise thresholding to $\hat{\beta}^{\mathrm{lasso}}$. The last approach is using an iterative algorithm, where the estimate of $\mathrm{Supp}(\beta)$ is updated sequentially, and each update depends on the output of the previous steps. Examples include forward regression, forward-backward regression (Zhang, 2011), and orthogonal matching pursuit (Donoho et al., 2012).

Given these existing variable selection methods, an interesting question is how they compare with each other. Existing theoretical studies were primarily focused on 'model selection consistency' (e.g., (Fan & Li, 2001; Zhao & Yu, 2006; Zou, 2006; Meinshausen & Bühlmann, 2010; Loh & Wainwright, 2017)), which targets to derive conditions on (X, β) such that $\mathbb{P}(\mathrm{Supp}(\hat{\beta}) = \mathrm{Supp}(\beta)) \to 1$. Given

these celebrated works, we would like to provide a different perspective of understanding: in the case that model selection consistency is impossible to achieve, how to compare different methods? We tackle this problem by studying the rate of convergence of *Hamming error* – the total number of false positives and false negatives. Model selection consistency is equivalent to a convergence of the Hamming error to 0 as $n, p \to \infty$. We are interested in a parameter regime where the Hamming error grows to infinity but is still much smaller than the total number of true relevant variables. We aim to compare the speed of growth of the Hamming error for different methods.

We adopt a theoretical framework that is commonly used in large-scale multiple testing (Donoho & Jin, 2004; Arias-Castro et al., 2011; Jin & Ke, 2016). Given two constants $\vartheta \in (0, 1)$ and $r > 0$, this framework assumes that the number of nonzero β_j's is $\sim p^{1-\vartheta}$ and the magnitude of each nonzero β_j is $\sim n^{-1/2}\sqrt{2r\log(p)}$. For multiple testing, there is no dependence across test statistics, which is equivalent to letting $X'X$ be a diagonal matrix. For regression, we cannot assume that $X'X$ is diagonal. We introduce a class of blockwise diagonal designs, where the level of correlations in the design is controlled by a parameter $\rho \in (-1, 1)$ and $\rho = 0$ corresponds to the special case of a diagonal $X'X$. Under this model, our main results are the explicit convergence rates of Hamming error, as a function of (ϑ, r, ρ), for different methods. We will study six methods: Lasso (Tibshirani, 1996), Elastic net (Zou & Hastie, 2005), SCAD (Fan & Li, 2001), thresholded Lasso (Zhou, 2009), forward selection, and forward backward selection (Huang et al., 2016); see Section 3 for a precise description of each method. These methods cover several strategies for variable selection, including non-convex penalty, L^2-regularization, post-Lasso processing, and iterative algorithms. To our best knowledge, our results are the first that directly compare Hamming errors of these methods.

2 A PREVIEW OF MAIN RESULTS AND SOME DISCUSSIONS

For notational simplicity, we re-parametrize (X, β) to $(n^{-1/2}X, \sqrt{n}\beta)$ and assume that $X'X$ has unit diagonals after re-parametrization. As a consequence, the magnitude of nonzero β_j in this paper should be multiplied by a factor of $n^{-1/2}$ when translated back to the original parametrization. We use p as the driving asymptotic parameter and assume $n = p^\omega$ for a constant $\omega > 1$ (see the end of this section for a discussion of $\omega < 1$). Given constants $\vartheta \in (0, 1)$, $r > 0$, and $\rho \in (-1, 1)$, we consider a sequence of linear models indexed by p, where the regression coefficient vector β contains iid entries drawn from a two-point mixture:

$$\beta_j = \begin{cases} \tau_p, & \text{with probability } \epsilon_p, \\ 0, & \text{with probability } 1 - \epsilon_p, \end{cases} \qquad \text{where} \qquad \epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r\log(p)}. \tag{3}$$

Model (3) is known as the *rare and weak signal model* (see Jin & Ke (2016) for a survey of the use of this model in statistical inference). As $p \to \infty$, the cardinality of $\mathrm{Supp}(\beta)$ is approximately $p^{1-\vartheta}$, and each nonzero coordinate of β is at the minimax order of $\sqrt{\log(p)}$. Next, we assume that the Gram matrix $G = X'X$ is block-wise diagonal consisting of 2×2 blocks:

$$G = \mathrm{diag}(B, B, \ldots, B, B_0), \qquad \text{where} \quad B = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \text{ and } B_0 = \begin{cases} B, & \text{if } p \text{ is even}, \\ 1, & \text{if } p \text{ is odd}. \end{cases} \tag{4}$$

Model (4) is an idealization of the blockwise correlation structures in real applications (e.g., the correlation matrix of genetic markers are often approximately blockwise diagonal). The three parameters (ϑ, r, ρ) capture the sparsity level, signal strength and design correlations, respectively.

For any $\hat{\beta}$, its Hamming error is $H(\hat{\beta}, \beta) = \sum_{j=1}^p 1\{\hat{\beta}_j \neq 0, \beta_j = 0\} + \sum_{j=1}^p 1\{\hat{\beta}_j = 0, \beta_j \neq 0\}$. In Section 3, we shall show that, for each of the six methods studied here, there exists a function $h(\vartheta, r, \rho) \in [0, 1]$ such that $\mathbb{E}[H(\hat{\beta}, \beta)] = L_p p^{1-h(\vartheta,r,\rho)}$, where L_p is a *multi*-$\log(p)$ term as $p \to \infty$. Noting that the expected number of true relevant variables is $p^{1-\vartheta}$, we are interested in three cases:

- *Exact recovery:* $h(\vartheta, r, \rho) > 1$. In this case, the expected Hamming error is $o(1)$ as $p \to \infty$. It follows that model selection consistency holds.
- *Almost full recovery:* $\vartheta < h(\vartheta, r, \rho) < 1$. In this case, the expected Hamming error does not vanish as $p \to \infty$, but it is much smaller than the total number of true relevant variables. Variable selection is still satisfactory (although model selection consistency does not hold).
- *No recovery:* $h(\vartheta, r, \rho) \leq \vartheta$. In this case, the expected Hamming error is comparable with or much larger than the total number of true relevant variables. Variable selection fails.

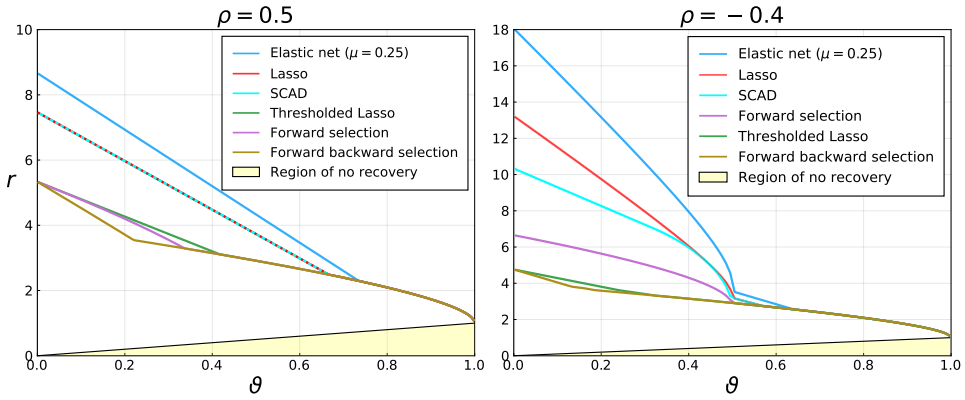


Figure 1: Phase diagrams of six variable selection methods for a block-wise diagonal design. The parameters (ϑ, r, ρ) characterize the sparsity, signal strength, and design correlations, respectively. For each method, we plot the curve $r = U(\vartheta)$ which separates Region of Almost Full Recovery and Region of Exact Recovery (the lower this curve, the better). Explicit forms of $U(\vartheta)$ are in Section 3. On the left panel, the curves for Lasso and SCAD overlap and are displayed as a dashed line. These phase curves should be interpreted with caution. Please see the discussions in Section 2.

We call the two-dimensional space (ϑ, r) the *phase space*. For each fixed ρ, the phase space is divided into three regions: *Region of Exact Recovery (ER)*, which is the subset $\{(\vartheta, r) : h(\vartheta, r, \rho) > 1\}$, and *Region of Almost Full Recovery (AFR)* and *Region of No Recovery (NR)* defined similarly. This gives rise to a *phase diagram* for each method. We denote the curve separating ER region and AFR region by $r = U(\vartheta)$ and the curve separating AFR region and NR region by $r = L(\vartheta)$; they are called the *upper and lower phase curves*, respectively. The phase diagram and phase curves are convenient ways to visualize the rates of convergence of the Hamming error.

Our main results are in Section 3. We give an overview here. Figure 1 shows the phase curves for six methods: Lasso, Elastic net, SCAD, thresholded Lasso, forward selection, and forward backward selection. These phase curves depend on the correlation parameter ρ. Under our model, for each diagonal block $(j, j+1)$, it holds that $\mathbb{E}[x_j' y|\beta] = \beta_j + \rho\beta_{j+1}$, where $\beta_j, \beta_{j+1} \in \{0, \tau_p\}$. Therefore, a *positive* ρ boosts the signal at each individual site (i.e., $\mathbb{E}[x_j' y|\beta] \geq \beta_j$), while a *negative* ρ leads to potential 'signal cancellation' (i.e., $\mathbb{E}[x_j' y|\beta] \leq \beta_j$). This is why the phase curves have different shapes for positive and negative ρ. In Figure 1, we plot the phase curves for $\rho = 0.5$ and $\rho = -0.4$. For other positive/negative value of ρ, the patterns are similar.

Discussion of SCAD. SCAD is a representative of the non-convex penalization methods. There were inspiring works that demonstrated the advantage of using a non-convex penalty (e.g., Fan & Peng (2004); Loh & Wainwright (2017)). Our results support their insights from a different angle: The phase curve of SCAD is strictly better than that of Lasso, when $\vartheta < 0.5$ and $\rho < 0$. Furthermore, our results illustrate where the advantage of SCAD comes from — compared with Lasso, it handles 'signal cancellation' better. To see this, we recall that 'signal cancellation' can only happen for $\rho < 0$. Moreover, under our model (3), for a diagonal block $\{j, j+1\}$, we get a pair of signals (i.e., β_j and β_{j+1} are both nonzero) with probability $\epsilon_p^2 = p^{1-2\vartheta}$ approximately. Therefore, 'signal cancellation' becomes problematic only when $\vartheta < 0.5$ and $\rho < 0$ both hold. This explains why the phase curves of SCAD and Lasso are the same for the other values of ϑ and ρ. We note that in previous studies (e.g., Loh & Wainwright (2017)), the advantage of a non-convex penalty in handling 'signal cancellation' is reflected in the conditions of (X, β) required for model selection consistency. Our results support this viewpoint from a different angle.

The performance of SCAD can be further improved by adding an entry-wise thresholding on $\hat{\beta}$. We are sure that the phase diagrams of *thresholded SCAD* will be better than those of SCAD, although we do not have the results in hand (the analysis is too tedious). We are also cautious in comparing SCAD and thresholded Lasso. In our settings, Lasso has no model selection consistency because the signals are too weak (i.e., r is not sufficiently large). In such a case, thresholded Lasso outperforms SCAD in terms of the Hamming error. Alternatively, there are cases where Lasso has no model selection consistency no matter how large the signal strength is (Zhao & Yu, 2006). For those cases, it is possible that SCAD is better than thresholded Lasso (see Wainwright (2009) for a related study).

Discussion of Elastic net. The phase curve of Elastic net is worse than that of Lasso. As we will explain in Section 3.1, Elastic net is a 'bridge' between Lasso and marginal regression. Since the phase curve of marginal regression is always worse than that of Lasso for the blockwise diagonal design, we do not benefit from using Elastic net in the current setting. We must note that Elastic net was motivated by genetic applications where several correlated variables are competing as predictors, and where it is implicitly assumed that groups of correlated variables tend to be all relevant or all irrelevant (Zou & Hastie, 2005). This is not captured by our model (3). Therefore, our result has no contradiction to the benefit of Elastic net known in the literature. Our result actually supports that the advantage of Elastic net comes from 'group' appearance of signal variables. We conjecture that for a different signal model (e.g., a group prior on β), Elastic net has a better phase diagram than Lasso.

Discussion of thresholded Lasso. Thresholded Lasso is a representative of improving Lasso by post-processing. There have been inspiring works that demonstrate the advantage of such a post-processing (van de Geer et al., 2011; Wang et al., 2017; Weinstein et al., 2020). Our results support these insights from a different angle. It is surprising (and very encouraging) that the improvement by post-Lasso thresholding is so significant. We note that Lasso is a 1-stage method, which solves a single optimization to obtain $\hat{\beta}$. In comparison, thresholded Lasso is a 2-stage method. Lasso has only one algorithm parameter λ, while thresholded Lasso has two algorithm parameters λ and t (the threshold). In Lasso, we control false positives and false negatives via the same algorithm parameter λ, and it is sometimes hard to find a value of λ that simultaneously controls the two types of errors well. In thresholded Lasso, the two types of errors are controlled separately by two algorithm parameters. This explains why thresholded Lasso has such a big improvement over Lasso. It inspires us to modify other 1-stage methods, such as SCAD, by adding a post-processing step of thresholding. For example, we conjecture that thresholded SCAD also has a strictly better phase diagram than that of Lasso, even for a positive ρ. On the other hand, thresholding is not a free lunch. It yields one more tuning parameter to decide in practice. Our theoretical results are based on ideal tuning parameters. How to properly choose these tuning parameters in a data-driven way is an interesting question. Weinstein et al. (2020) proposed a promising approach, where they used cross-validation to select λ and FDR control by knock-off to select t. We leave it to future work to study the phase diagram with data-driven tuning parameters.

Discussion of the two iterative algorithms. We consider two iterative algorithms, forward selection ('Forward') and forward backward selection (FB). The FB algorithm we analyze is a simplified version in Huang et al. (2016), which has only one backward step (after all the forward steps are completed) by thresholding the refitted least-squares solution. Our results show that both methods outperform Lasso, and between these two methods, FB is strictly better than Forward. In the literature, there were very interesting theoretical works showing the advantage of iterative algorithms for variable selection (Donoho et al., 2012; Zhang, 2011). Our results support their insights from a different angle. We discover that, for a wide range of ρ, FB has the best phase diagram among all six methods. This is a very encouraging result. At the same time, we keep in mind that the performance of an iterative algorithm tends to be more sensitive to the form of the design, due to its sequential nature.

Technical novelty. The phase diagram has been used before for studying variable selection (Genovese et al., 2012; Ji & Jin, 2012; Jin et al., 2014; Ke et al., 2014), but those results focused on a special method, Screen-and-Clean. Unfortunately, we cannot use their techniques to study the methods of interest in this paper. The derivation of phase diagrams requires *case-by-case* analysis, where the key is a precise characterization of the geometry of the solution path of a method. Whenever the method changes or the design class changes, the geometry of the solution can change significantly, making the analysis very different. For this reason, we need a significant amount of work (71 pages of proofs) to obtain the phase diagrams of the six methods considered here.

Extension to $n < p$. We have assumed $n = p^\omega$ for some $\omega > 1$. This ensures that the block-wise structure on $X'X$ is eligible (if $n < p$, this matrix is singular). However, this does not mean that our study is limited to $n > p$. For SCAD, Elastic net, Lasso and thresholded Lasso, we can easily extend all the theoretical results to the case of $n = p^\omega$ for $\omega \in (0, 1]$, by assuming that the rows of X are iid generated from $N(0, n^{-1}\Sigma)$, with Σ having the same blockwise structure as in (4) (how to extend a phase diagram for fixed X to iid Gaussian random X has been given in Ji & Jin (2012) and Jin et al. (2014)). It is yet unclear how to extend the results for Forward and FB to iid Gaussian random X, because in these iterative algorithms, the outcome of each iteration crucially depends on the outcome of the previous iteration. We leave these to future work.

Conclusion. Most theoretical works for variable selection focus on one method. In contrast, we provide a platform to directly compare the main-stream approaches for variable selection. It yields useful insights for understanding the pros and cons of each method. Our results about the convergence rates of Hamming errors and phase diagrams require very delicate analysis. We have to start from a manageable setting by using a model where sparsity, signal strength and design correlations are characterized by three simple parameters (ϑ, r, ρ). We expect that similar analysis can be conducted for more general models (e.g., block-wise correlation structure with larger blocks, factor models, etc.; see our simulations). It will require a significant amount of work, which we leave for future.

3 MAIN RESULTS

Consider the linear regression model (1), (3), and (4), where we set $\sigma^2 = 1$ without loss of generality. Let $\mathbb{E}[H(\hat{\beta}, \beta)]$ be the expected Hamming error, where the expectation is with respect to the randomness of β and z. Let L_p denote a generic multi-$\log(p)$ term such that $L_p p^{\epsilon} \to \infty$ and $L_p p^{-\epsilon} \to 0$ for any $\epsilon > 0$. For $\hat{\beta}$ from each method, we will show that there is a function $h(\vartheta, r, \rho)$ such that $\mathbb{E}[H(\hat{\beta}, \beta)] = L_p p^{1-h(\vartheta,r,\rho)}$. Define

$$U(\vartheta; \rho) = \inf\{r > 0 : h(\vartheta, r, \rho) > 1\}, \qquad L(\vartheta; \rho) = \inf\{r > 0 : h(\vartheta, r, \rho) > \vartheta\}. \tag{5}$$

We write them as $U(\vartheta)$ and $L(\vartheta)$ for short and call them the upper/lower phase curves, respectively. The upper phase curve $U(\vartheta)$ separates the ER region and AFR region, and the lower-phase-curve $L(\vartheta)$ separates the AFR region and NR region. Note that $h(\vartheta, r, \rho)$ captures the **rate of convergence of the expected Hamming error**. In the supplement, for each method, we present a theorem about $h(\vartheta, r, \rho)$. In the main article, to save space, we only present the phase curves $L(\vartheta)$ and $U(\vartheta)$.

3.1 ELASTIC NET AND LASSO

The Elastic net (Zou & Hastie, 2005) is a method that estimates β by

$$\hat{\beta}^{\mathrm{EN}} = \mathrm{argmin}_{\beta}\big\{\|y - X\beta\|^2/2 + \lambda\|\beta\|_1 + (\mu/2)\|\beta\|^2\big\}. \tag{6}$$

Compared with Lasso, it adds an additional L^2-penalty to the objective function. Below, we fix $\mu > 0$ and re-parametrize $\lambda = \sqrt{2q\log(p)}$, for some constant $q > 0$. The choice of q affects the exponent, $1 - h(\vartheta, r, \rho)$, in the expression of $\mathbb{E}[H(\hat{\beta}, \beta)]$. We choose the ideal q that minimizes $1 - h(\vartheta, r, \rho)$. The next theorem is proved in the supplement.

Theorem 1 (Elastic Net)**.** *Under Models* (1)*,* (3)*, and* (4)*, let* $\hat{\beta}^{\mathrm{EN}}$ *be the Elastic net estimator in* (6)*. Fix* μ *and write* $\eta = \rho/(1+\mu)$*. Let* $\lambda = \sqrt{2q\log(p)}$ *with an ideal choice of* q *that minimizes the exponent of* $\mathbb{E}[H(\hat{\beta}, \beta)]$*. The phase curves are given by* $L(\vartheta) = \vartheta$*, and*

$$U(\vartheta) = \begin{cases} \max\{h_1(\vartheta), h_2(\vartheta)\}, & \textit{when } \rho \geq 0, \\ \max\{h_1(\vartheta), h_2(\vartheta), h_3(\vartheta), h_4(\vartheta)\}, & \textit{when } \rho < 0, \end{cases}$$

where $h_1(\vartheta) = (1 + \sqrt{1-\vartheta})^2$*,* $h_2(\vartheta) = \big(\frac{1-|\eta|}{1-|\rho|} + \frac{\sqrt{1+\eta^2-2\rho\eta}}{1-|\rho|}\big)^2(1-\vartheta)$*,* $h_3(\vartheta) = \frac{1}{(1-|\rho|)^2}\big(1 + \frac{\sqrt{1+\eta^2-2\rho\eta}}{1+|\eta|}\sqrt{1-2\vartheta}\big)^2$*, and* $h_4(\vartheta) = \frac{1+\eta^2-2\rho\eta}{(1-2|\rho|+\rho\eta)_+^2}\big(\sqrt{1-\vartheta} + \frac{1-|\eta|}{1+|\eta|}\sqrt{1-2\vartheta}\big)^2$*.*

Lasso is a special case with $\mu = 0$. By setting $\mu = 0$ (equivalently, $\eta = \rho$) in Theorem 1, we obtain the phase curves for Lasso. They agree with the results in Ji & Jin (2012) (but Ji & Jin (2012) does not cover Elastic net).

To see the effect of the L^2-penalty, we consider an extreme case where $\mu \to \infty$. Some elementary algebra shows that $(1+\mu)\hat{\beta}^{\mathrm{EN}}$ converges to the soft-thresholding of $X'y$ at the threshold λ. In other words, as $\mu \to \infty$, Elastic net converges to marginal screening (selecting variables by thresholding the marginal regression coefficients). At the same time, when $\mu = 0$, $(1+\mu)\hat{\beta}^{\mathrm{EN}}$ equals the Lasso estimate. Hence, Elastic net serves as a bridge between Lasso and marginal regression. In the setting here, the phase diagram of marginal regression is inferior to that of Lasso, and so the phase diagram of Elastic net is also inferior to that of Lasso. See the proposition below and Figure 2:

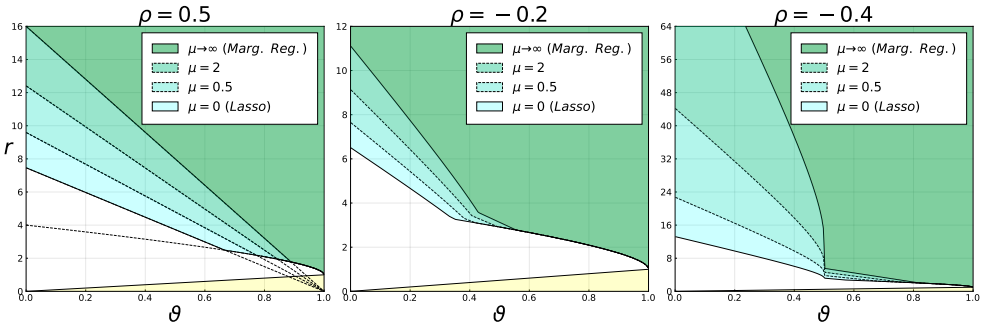


Figure 2: The phase diagrams of Elastic net and its comparison with Lasso (notation: $\eta = \rho/(1+\mu)$).

Proposition 1. *In Theorem 1, for each fixed $\vartheta \in (0,1)$, as $\mu \to \infty$, $U(\vartheta)$ is monotone decreasing and converges to $U_{\mathrm{Lasso}}(\vartheta)$, which is the upper phase curve of Lasso; as $\mu \to \infty$, $U(\vartheta)$ is monotone increasing and converges to $U_{\mathrm{MR}}(\vartheta)$, which is the upper phase curve of marginal regression. Furthermore, when $\rho \le -\frac{1}{2}$, $U_{\mathrm{MR}}(\theta) = \infty$ for all $0 < \vartheta \le \frac{1}{2}$ (i.e. exact recovery is impossible to achieve no matter how large r is).*

3.2 SCAD

SCAD (Fan & Li, 2001) is a non-convex penalization method. For any $a > 2$, it defines a penalty function $q_\lambda(\theta)$ on $(0,\infty)$ by $q_\lambda(\theta) = \int_0^\theta q'_\lambda(t)dt$, where $q'_\lambda(\theta) = \lambda\big\{I(\theta \le \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} I(\theta > \lambda)\big\}$. The resulting penalty function $q_\lambda(\cdot)$ coincides with the L^1-penalty in $(0,\lambda]$ and becomes a constant in $[a\lambda,\infty)$. Let $Q_\lambda(\beta) = \sum_{j=1}^p q_\lambda(|\beta_j|)$. Then, SCAD estimates β by

$$\hat{\beta}^{\mathrm{SCAD}} = \mathrm{argmin}_\beta\big\{\|y - X\beta\|^2/2 + Q_\lambda(\beta)\big\}. \tag{7}$$

The following theorem is proved in the supplemental material (see Figure 3, left panel):

Theorem 2 (SCAD)**.** *Under Models* (1)*,* (3)*, and* (4)*, let $\hat{\beta}^{\mathrm{SCAD}}$ be the SCAD estimator in* (7)*. Fix $a \in (2, \frac{2}{1-|\rho|})$. Let $\lambda = \sqrt{2q\log(p)}$ with an ideal choice of q that minimizes the rates of convergence of the expected Hamming error. The phase curves are given by $L(\vartheta) = \vartheta$, and*

$$U(\vartheta) = \begin{cases} \max\{h_1(\vartheta), h_2(\vartheta), h_3(\vartheta)\}, & \text{when } \rho \ge 0, \\ \max\{h_1(\vartheta), h_2(\vartheta), h_4(\vartheta), h_5(\vartheta)\}, & \text{when } \rho < 0, \end{cases}$$

where $h_1(\vartheta) = (1+\sqrt{1-\vartheta})^2$, and $h_2(\vartheta) = \big(1+\sqrt{\frac{1+|\rho|}{1-|\rho|}}\big)^2(1-\vartheta)$, $h_4(\vartheta) = \big(\sqrt{\frac{1-2\vartheta}{1-|\rho|^2}} + \frac{1}{1-|\rho|}\big)^2$, $h_3(\vartheta) = \big(\frac{3+\rho}{2(1-\rho^2)}\sqrt{\frac{1+\rho}{1-\rho}}\sqrt{1-\vartheta} + \frac{1}{2}\sqrt{\frac{2(1-2\vartheta)}{1+\rho} - \frac{(1-\vartheta)}{(1-\rho)^2}}\big)^2$, and

$$h_5(\vartheta) = \begin{cases} \big(\frac{5+3|\rho|}{1-|\rho|}\big)(1-\vartheta), & \text{if } \sqrt{\frac{1-2\vartheta}{1-\vartheta}} \ge \frac{3-4|\rho|-3\rho^2}{(1-|\rho|)}\sqrt{\frac{1+|\rho|}{5+3|\rho|}}, \\ \frac{1}{(1-|\rho|)^2}\big(\sqrt{\frac{1+|\rho|}{1-|\rho|}}\sqrt{1-\vartheta} + \sqrt{\frac{1-|\rho|}{1+|\rho|}}\sqrt{1-2\vartheta}\big)^2, & \text{if } \sqrt{\frac{1-2\vartheta}{1-\vartheta}} \le \frac{(1+|\rho|)(1-2|\rho|)}{1-|\rho|}, \\ h_6(\vartheta) & \text{other wise}, \end{cases}$$

with

$$h_6(\vartheta) = \left\{\sqrt{\frac{1-2\vartheta}{1-\rho^2}} + \frac{\frac{1-2|\rho|}{1-|\rho|}\sqrt{\frac{1-2\vartheta}{1-\rho^2}} + \sqrt{\big[\big(\frac{1-2|\rho|}{1-|\rho|}\big)^2 + \frac{1-|\rho|}{1+|\rho|}\big](1-\vartheta) - \frac{1-2\vartheta}{(1+|\rho|)^2}}}{(1-|\rho|)\big[\big(\frac{1-2|\rho|}{1-|\rho|}\big)^2 + \frac{1-|\rho|}{1+|\rho|}\big]}\right\}^2.$$

Note that the phase curves of Lasso are given in Theorem 1 by setting $\eta = \rho$. We compare SCAD with Lasso. When $\rho < 0$, the upper phase curve in Theorem 2 is strictly lower than that of Lasso (see Figure 3, middle and right panels). When $\rho \ge 0$, the upper phase curve in Theorem 2 is sometimes higher than that of Lasso. Note that we restrict $a < \frac{2}{1-|\rho|}$ in Theorem 2. In fact, a larger a may be preferred for $\rho \ge 0$. The next proposition is about using an optimal a.

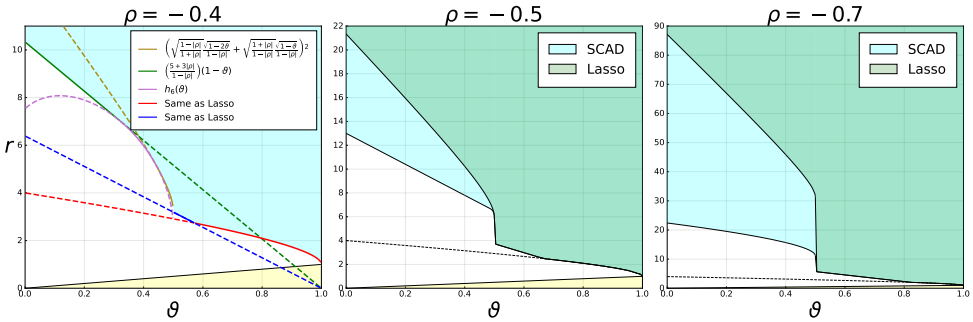


Figure 3: Left: Phase curves of SCAD. Middle and Right: Comparison of SCAD and Lasso.

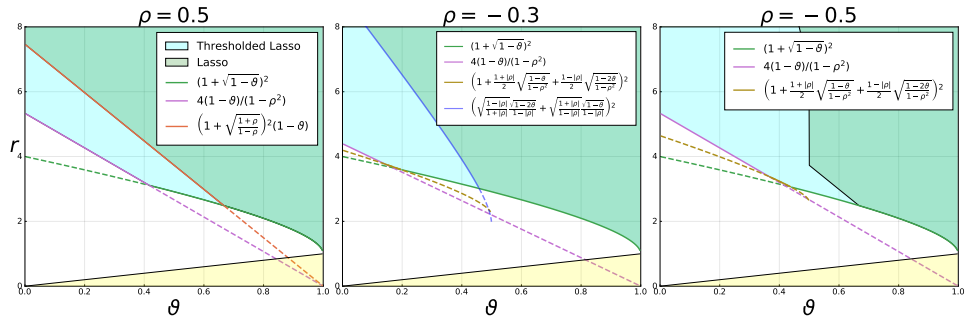


Figure 4: Comparison of the phase diagrams of thresholded Lasso and Lasso.

Proposition 2. *In the SCAD estimator, we choose $a = a^*$ and $\lambda = \sqrt{2q^* \log(p)}$ such that $(a^*, q^*) = (a^*(\vartheta, r, \rho), q^*(\vartheta, r, \rho))$ minimize the rates of convergence of the expected Hamming error among all choices of (a, q). Let $U^*(\vartheta)$ be the resulting upper phase curve for SCAD. Then, $U(\vartheta) = U_{\text{Lasso}}(\vartheta)$ when $\rho \geq 0$, and $U(\vartheta) < U_{\text{Lasso}}(\vartheta)$ when $\rho < 0$.*

The phase curves of SCAD are insensitive to the choice of a. When $a < 0$, the optimal a^* can be any value in $(2, \frac{2}{1-|\rho|})$. When $\rho \geq 0$, there exists a constant $c = c(\vartheta, \rho)$ such that the optimal a^* is any value in (c, ∞). As $a \to \infty$, the SCAD penalty reduces to the L^1-penalty. This explains why the phase curve of SCAD is the same as that of Lasso when $\rho \geq 0$.

3.3 Thresholded Lasso

Let $\hat{\beta}^{\text{Lasso}}$ be the Lasso estimator in (2). The thresholded Lasso estimator $\hat{\beta}^{\text{TL}}$ is obtained by applying coordinate-wise hard-thresholding to the Lasso estimator:

$$\hat{\beta}_j^{\text{TL}} = \hat{\beta}^{\text{Lasso}} \cdot 1\{|\hat{\beta}^{\text{Lasso}}| > t\}, \qquad 1 \leq j \leq p. \tag{8}$$

Theorem 3 (Thresholded Lasso)**.** *Under Models* (1), (3), *and* (4), *let $\hat{\beta}^{\text{TL}}$ be the thresholded Lasso estimator in* (8). *Let $\lambda = \sqrt{2q \log(p)}$ and $t = \sqrt{2w \log(p)}$ with the ideal (q, w) that minimize the exponent of the expected Hamming error. The phase curves are given by $L(\vartheta) = \vartheta$, and*

$$U(\vartheta) = \begin{cases} \max\{h_1(\vartheta), h_2(\vartheta)\}, & \text{when } \rho \geq 0, \\ \max\{h_1(\vartheta), h_2(\vartheta), h_3(\vartheta)\}, & \text{when } \rho < 0, \end{cases}$$

where $h_1(\vartheta) = (1+\sqrt{1-\vartheta})^2$, $h_2(\vartheta) = \frac{4(1-\vartheta)}{1-\rho^2}$, and $h_3(\vartheta) = \big(1 + \frac{1+|\rho|}{2}\sqrt{\frac{1-\vartheta}{1-\rho^2}} + \frac{1-|\rho|}{2}\sqrt{\frac{1-2\vartheta}{1-\rho^2}}\big)^2$.

See Figure 4 for a comparison with Lasso (a special case of $t = 0$). With the flexibility of using an optimal t, the phase diagram of thresholded Lasso is always better than that of Lasso.

Theorem 3 also gives other interesting facts about thresholded Lasso. First, the shape of phase curves is much less affected by the sign of ρ. This differs from Lasso, Elastic net, and SCAD, for which the shape of phase curves is significantly different for positive and negative ρ. Second, the optimal λ in thresholded Lasso is considerably smaller than the optimal λ in Lasso (it can be seen from the proofs

of Theorem 3 and Theorem 1). This is because the λ in thresholded Lasso only serves to control false negatives, but the λ in Lasso is used to simultaneously control false positives and false negatives, hence, cannot be too small. We observe the same phenomenon in simulations; see Section 4.

3.4 FORWARD SELECTION AND FORWARD BACKWARD SELECTION

Forward selection is a classical textbook method for variable selection. Write $X = [x_1, x_2, \ldots, x_p]$, where $x_i \in \mathbb{R}^n$ for $1 \leq i \leq p$. For any subset $A \subset \{1, 2, \ldots, p\}$, let $P_A^{\perp}$ be the projection onto the orthogonal complement of the linear space spanned by $\{x_i : i \in A\}$. Given a threshold $t > 0$, the forward selection algorithm initializes with $S_0 = \emptyset$ and $\hat{r}_0 = y$. At the kth iteration, compute

$$i^* = \mathrm{argmax}_{i \notin S_{k-1}} |x_i' \hat{r}_{k-1}|, \qquad \delta = |x_{i^*}' \hat{r}_{k-1}| / \|P_{S_{k-1}}^{\perp} x_{i^*}\|.$$

If $\delta > t$, compute $S_k = S_{k-1} \cup \{i^*\}$ and $\hat{r}_k = P_{S_k}^{\perp} y$; otherwise, output $\hat{\beta}^{\mathrm{forward}}$ as the least-squares estimator restricted to S_{k-1}. The stopping rule of $\delta \leq t$ is equivalent to measuring the decrease of the residual sum of squares. The following theorem is proved in the supplemental material:

Theorem 4 (Forward Selection)**.** *Under Models* (1)*,* (3)*, and* (4)*, let $\hat{\beta}^{\mathrm{forward}}$ be the estimator from forward selection. Let $t = \sqrt{2q\log(p)}$ with the ideal q that minimizes the exponent of the expected Hamming error. The phase curves are given by $L(\vartheta) = \vartheta$, and*

$$U(\vartheta) = \begin{cases} \max\{h_1(\vartheta), h_2(\vartheta), h_3(\vartheta)\}, & \text{when } \rho \geq 0, \\ \max\{h_1(\vartheta), h_2(\vartheta), h_3(\vartheta), h_4(\vartheta)\}, & \text{when } \rho < 0, \end{cases}$$

with $h_1(\vartheta) = (1 + \sqrt{1-\vartheta})^2$, $h_2(\vartheta) = \frac{2(1-\vartheta)}{1-|\rho|}$, $h_3(\vartheta) = \frac{(1+\sqrt{1-2\vartheta})^2}{1-\rho^2}$, $h_4(\vartheta) = \big(\sqrt{\frac{1-2\vartheta}{2(1-|\rho|)}} + \frac{1}{1-|\rho|}\big)^2$.

Forward backward selection (FB) modifies forward selection by allowing to drop variables. We use the FB algorithm in Huang et al. (2016), where the backward step is conducted after all the forward steps are finished. For a threshold $v > 0$, it applies entry-wise thresholding on $\hat{\beta}^{\mathrm{forward}}$:

$$\hat{\beta}_j^{\mathrm{FB}} = \hat{\beta}_j^{\mathrm{forward}} \cdot 1\{|\hat{\beta}_j^{\mathrm{forward}}| > v\}, \qquad 1 \leq j \leq p. \tag{9}$$

Theorem 5 (Forward Backward Selection)**.** *Under Models* (1)*,* (3)*, and* (4)*, let $\hat{\beta}^{\mathrm{FB}}$ be the estimator from forward selection. Let $t = \sqrt{2q\log(p)}$ and $v = \sqrt{2u\log(p)}$ with the ideal (q, u) that minimize the exponent of the expected Hamming error. When $\rho \geq 0$, the phase curves are given by $L(\vartheta) = \vartheta$, and*

$$U(\vartheta) = \max\{h_1(\vartheta), h_2(\vartheta), h_3^*(\vartheta)\},$$

where $h_1(\vartheta)$ and $h_2(\vartheta)$ are the same as in Theorem 4 and $h_3^*(\vartheta) = \frac{(\sqrt{1-\vartheta}+\sqrt{1-2\vartheta})^2}{1-\rho^2}$. When $\rho < 0$,

$$U(\vartheta) \leq \max\{g_1(\vartheta), g_2(\vartheta), g_3(\vartheta), g_4(\vartheta)\}$$

where $g_1(\vartheta) = (v_{\min}(\vartheta) + \sqrt{1-\vartheta})^2$, $g_2(\vartheta) = \frac{2(1-\vartheta)}{1-|\rho|}$, $g_3(\vartheta) = \big(\sqrt{\frac{1-2\vartheta}{1-\rho^2}} + v_{\min}(\vartheta)\big)^2$, $g_4(\vartheta) = \big(\sqrt{\frac{1-2\vartheta}{2(1-|\rho|)}} + \frac{t_{\min}(\vartheta)}{1-|\rho|}\big)^2$, $v_{\min}(\vartheta) = \max\big\{1, \sqrt{\frac{1-\vartheta}{1-\rho^2}}\big\}$, *and* $t_{\min}(\vartheta) = \max\big\{\frac{\sqrt{2}}{2}, \frac{v_{\min}(\vartheta)}{1+|\rho|/\sqrt{1-\rho^2}}\big\}$.

Theorem 5 gives $U(\vartheta)$ for $\rho \geq 0$ and an upper bound of it for $\rho < 0$. Combining it with Theorems 1 and 4, we conclude that the upper phase curve of FB is always better than those of Lasso and forward selection (for $\rho < 0$, the upper bound here is already better than $U(\vartheta)$ for the other two methods).

We remark that we did obtain the exact phase curve for $\rho < 0$ in the proof of Theorem 5. It is just too complicated and space-consuming to present it in the main text. However, given specific values of (ϑ, ρ), we can always plot the exact phase curve using the (complicated) formulas in the supplement. In Figures 1 and 5, the phase curves of FB are indeed the exact ones.

4 SIMULATIONS

We conduct small-scale simulations. In all experiments below, we fix $(n, p) = (1000, 300)$.

Experiment 1 (block-wise diagonal designs). We generate (X, β) as in (3)-(4). For each method, we select the ideal tuning parameters that minimize the averaged Hamming error over 50 repetitions.

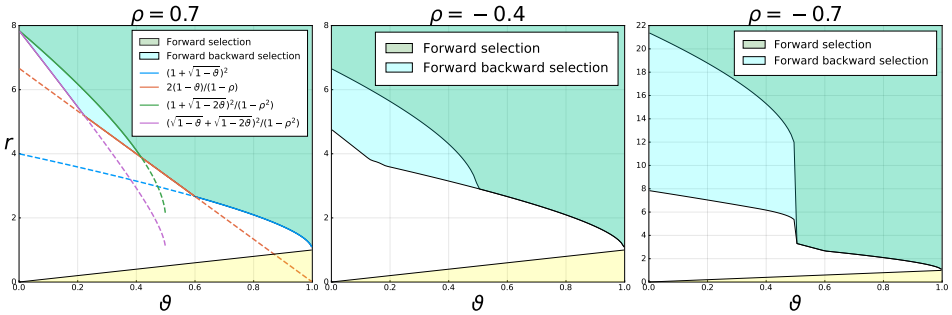


Figure 5: The phase diagrams of forward selection and forward backward selection.

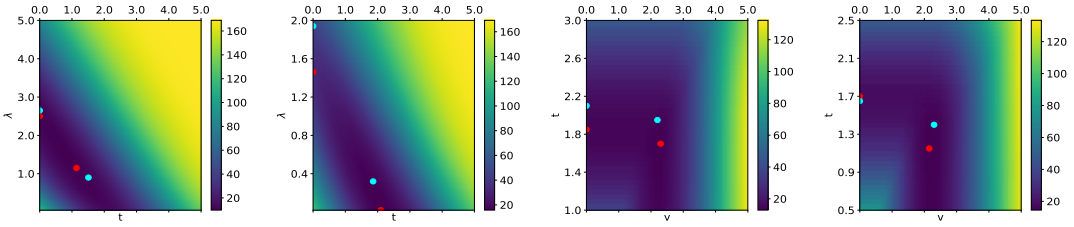
The averaged Hamming errors and its standard deviations under the ideal tuning parameters over 500 repetitions are reported below. The results are consistent with the theoretical phase diagrams (see Figure 1). E.g., thresholded Lasso and forward backward selection are the two methods that perform the best; Lasso is more unsatisfactory when $\rho < 0$; SCAD improves Lasso when $\rho < 0$.

ρ	θ	r	Lasso	ThresLasso	ElasticNet	SCAD	Forward	FoBackward
0.5	0.1	1.5	11.57 (3.59)	**10.48** (3.34)	11.57 (3.31)	11.72 (3.33)	14.88 (4.12)	13.35 (3.90)
0.5	0.1	4	1.00 (1.00)	**0.42** (0.65)	1.03 (1.00)	1.00 (0.96)	0.66 (0.84)	0.51 (0.73)
-0.5	0.1	1.5	35.62 (5.09)	15.62 (4.06)	35.48 (5.64)	25.87 (5.04)	19.48 (4.61)	**14.82** (3.82)

Experiment 2 (general designs). In the Toeplitz design, we let $(X'X)_{i,j} = 0.7^{|i-j|}$ and set $(\vartheta, r) = (0.1, 2.5)$. In the factor model design, we let $X'X = BB' - \mathrm{diag}(BB') + I_p$, where entries of $B \in \mathbb{R}^{p\times 2}$ are *iid* from $\mathrm{Unif}(0, 0.6)$, and set $(\vartheta, r) = (0.1, 1.5)$. Same as in Experiment 1, we use the ideal tuning parameters. The averaged Hamming errors and its standard deviations are reported below. The Toeplitz design is a setting where each variable is only highly correlated with a few other variables. The factor model design is a setting where a variable is (weakly) correlated with all the other variables. The results are quite similar to those in Experiment 1. This confirms that the insight gained in the study of the block-wise diagonal design continues to apply to more general designs.

design	Lasso	ThresLasso	ElasticNet	SCAD	Forward	FoBackward
Toeplitz	47.15 (6.32)	**22.02** (5.31)	47.40 (6.41)	24.61 (5.70)	30.77 (6.18)	22.93 (5.44)
Factor model	21.14 (4.52)	**15.90** (3.87)	21.20 (4.45)	19.68 (4.23)	20.04 (4.34)	16.13 (3.76)

Experiment 3 (tuning parameters). Fix $(\vartheta, r) = (0.1, 1.5)$ and $\rho \in \{\pm 0.5\}$ in the block-wise diagonal design. We study the effect of tuning parameters in Lasso, thresholded Lasso (ThreshLasso), forward selection (ForwardSelect), and forward backward selection (FB). In (a)-(b), we show the heatmap of averaged Hamming error (over 50 repetitions) of ThreshLasso for a grid of (t, λ); when $t = 0$, it reduces to Lasso. In (c)-(d), we show the Hamming error of FB for a grid of (v, t); when $v = 0$, it reduces to ForwardSelect. Cyan points are theoretically optimal tuning parameters (formulas are in proofs of theorems). Red points are empirically optimal tuning parameters that minimize the averaged Hamming error. The theoretical tuning parameter values are quite close to the empirically optimal ones. Moreover, the optimal λ in ThreshLasso is smaller than the optimal λ in Lasso.



(a) Lasso and ThreshLasso ($\rho = 0.5$)
(b) Lasso and ThreshLasso ($\rho = -0.5$)
(c) ForwardSelect and FB ($\rho = 0.5$)
(d) ForwardSelect and FB ($\rho = -0.5$)

REFERENCES

Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556,

2011.

Peter J Bickel, Ya?acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

David L Donoho, Yaakov Tsaig, Iddo Drori, and Jean-Luc Starck. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE transactions on Information Theory*, 58(2):1094–1121, 2012.

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

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The annals of statistics*, 32(3):928–961, 2004.

Christopher R Genovese, Jiashun Jin, Larry Wasserman, and Zhigang Yao. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13(Jun):2107–2143, 2012.

Shiqiong Huang, Jiashun Jin, Zhigang Yao, et al. Partial correlation screening for estimating large precision matrices, with applications to classification. *The Annals of Statistics*, 44(5):2018–2057, 2016.

Pengsheng Ji and Jiashun Jin. UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, 40(1):73–103, 2012.

Jiashun Jin and Zheng Tracy Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica*, pp. 1–34, 2016.

Jiashun Jin, Cun-Hui Zhang, and Qi Zhang. Optimality of graphlet screening in high dimensional variable selection. *The Journal of Machine Learning Research*, 15(1):2723–2772, 2014.

Tracy Ke, Jiashun Jin, and Jianqing Fan. Covariance assisted screening and estimation. *Annals of statistics*, 42(6):2202, 2014.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Sara van de Geer, Peter Bühlmann, Shuheng Zhou, et al. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is optimal for variable selection? *arXiv preprint arXiv:1705.08617*, 2017.

Asaf Weinstein, Weijie J Su, Małgorzata Bogdan, Rina F Barber, and Emmanuel J Candès. A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*, 2020.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7):4689–4708, 2011.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Shuheng Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22:2304–2312, 2009.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.