FALSE DISCOVERY RATE CONTROL WITH CONCAVE PENALTIES USING STABILITY SELECTION

Bhanukiran Vinzamuri and Kush R. Varshney

IBM Research
Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York, USA

ABSTRACT

False discovery rate (FDR) control is highly desirable in several high-dimensional estimation problems. While solving such problems, it is observed that traditional approaches such as the Lasso select a high number of false positives, which increase with higher noise and correlation levels in the dataset. Stability selection is a procedure which uses randomization with the Lasso to reduce the number of false positives. It is known that concave regularizers such as the minimax concave penalty (MCP) have a higher resistance to false positives than the Lasso in the presence of such noise and correlation. The benefits with respect to false positive control for developing an approach integrating stability selection with concave regularizers has not been studied in the literature so far. This motivates us to develop a novel upper bound on false discovery rate control obtained through this stability selection with minimax concave penalty approach.

Index Terms— sparse regression, concave penalties, false discovery rate control, stability selection, interpretability.

1. INTRODUCTION

The standard linear regression problem has the following form:

$$y = X\beta + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^n$ is a response variable, $X \in \mathbb{R}^{n\times p}$ is a feature matrix, $\beta \in \mathbb{R}^p$ is a coefficient vector, and $\varepsilon \in \mathbb{R}^n$ is a noise vector which has zero mean and sub-Gaussian noise such that $\varepsilon \sim N(0, \sigma^2 I_{n\times n})$. The sub-Gaussian property is more explicitly defined as follows.

Definition 1. *A vector ε is sub-Gaussian with noise level $\tilde{\sigma}$ if for all $t \geq 0$*

$$\mathbb{P}(|u^T\varepsilon| > \tilde{\sigma}t) \leq \exp(-t^2/2) \tag{2}$$

for all vector u with $\|u\|_2 = 1$.

This work was conducted under the auspices of the IBM Science for Social Good initiative.

Table 1. Notation

Name	Description
n	number of instances
p	number of features
X	$\mathbb{R}^{n\times p}$ feature vector matrix
y	$\mathbb{R}^n$ response variable
β	$\mathbb{R}^p$ regression coefficient vector
λ	scalar regularization parameter
Λ	vector of regularization parameters
$\hat{\Pi}^\Lambda$	$[0,1]^{p\times\lvert\Lambda\rvert}$ selection probability matrix
π_{thr}	cut-off parameter
γ	scalar concavity parameter
$h(\beta_j;\lambda;\gamma)$	family of concave penalty functions

The following class of regularized linear regression problems is studied in this paper:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} L(\beta;\lambda;\gamma), \tag{3}$$

where

$$L(\beta;\lambda;\gamma) = \frac{1}{2n}\|y - X\beta\|_2^2 + \sum_{j=1}^{p} h(\beta_j;\lambda;\gamma), \tag{4}$$

$\beta = (\beta_1, \ldots, \beta_p)$, and $h(\beta_j;\lambda;\gamma)$ is a concave penalty function consisting of parameters λ and γ. A complete description of the properties of this family of penalties can be found in Section 3.

In Table 1, we present some of the main notations used in this paper. In addition, the following terms are used frequently throughout this paper. $S = \text{supp}(\beta)$=$\{j : \beta_j \neq 0\}$. The complement of S is referred to as S^c or N which refers to the noise variables. $\hat{S}^\Lambda$ refers to the final set of variables selected after varying the regularization parameter from λ_{max} to λ_{min}.

1.1. Motivation for Concave Penalties

Supervised feature selection [1, 2, 3] from high-dimensional data involves using sparsity inducing norms in a classification or regression framework to extract relevant variables. Generally, most popularly used sparsity inducing norms are convex functions which create an optimization problem that can

978-1-5386-4410-2/18/$31.00 ©2018 IEEE

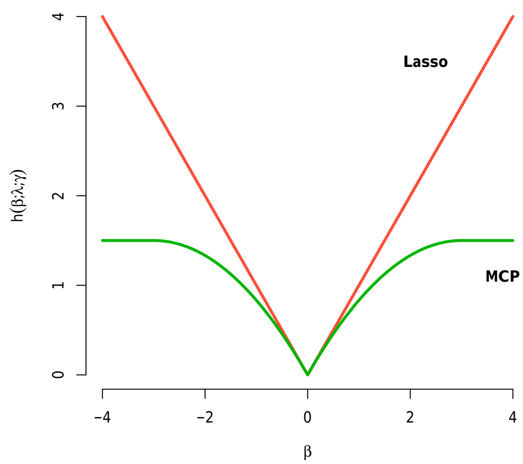

Fig. 1. Penalty plots for Lasso and MCP to show the *stabilizing effect* which results in lower bias for MCP.

be solved efficiently. However, recent advances in the literature [4, 5, 6] have indicated that non-convex penalties are more effective at extracting the true set of relevant features. However, solving these problems is non-trivial, for example, the ℓ_0-norm regularized least squares which is also called the *best subset selection* problem is NP-hard [7]. Due to the computational difficulty associated with solving it, a convex surrogate such as the ℓ_1-norm regularized least squares problem (Lasso) is solved. However, Lasso is known to *overpenalize* coefficients which results in biased estimates for true variables with significant non-zero coefficient values.

Concave penalties are proven to handle such bias effectively and are also easier to solve than ℓ_0 problems [8]. In particular, folded concave penalties such as the minimax concave penalty (MCP) have a *stabilizing effect* where the value of the penalty becomes constant after a certain range of β values which controls the penalization effectively [9]. This is illustrated in Figure 1.

Theoretical results indicate that for a noise level with standard deviation σ and universal amount of penalization $\lambda_{univ} \equiv \sigma\sqrt{\frac{2logp}{n}}$, MCP is said to have a *selection consistency* property [10], which implies that the set of selected variables is identical to the set of true nonzero regression coefficients with high probability. However, estimating noise level precisely from real-world data is a non-trivial task which makes it difficult to set λ_{univ}.

Our proposed stability selection with concave penalties approach handles this problem by defining a range of permissible regularization parameters. This is easier to define and makes the framework less parameter dependent.

2. FALSE DISCOVERY RATE CONTROL

Effective FDR control [11] helps in improving model consistency and interpretation. In this section, we derive an upper bound on the expected number of false positives with our approach. This analysis is very important from a practitioner's perspective, as he or she can tune the number of features to be selected at a specified false positive error rate or vice versa. We now discuss the details of our theoretical analysis.

For any regularization parameter $\lambda \in \Lambda$, the selected set $\hat{S}^{\lambda}$ represents the set of features active in the model at parameter λ. For every set $K \subseteq \{1, 2, \ldots, p\}$, the probability of being in the selected set $\hat{S}^{\lambda}$ is:

$$\hat{\Pi}_K^{\lambda} = \mathbb{P}^*\{K \subseteq \hat{S}^{\lambda}(I)\}, \tag{5}$$

where $\mathbb{P}^*$ represents the probability estimate.

For every variable $k = \{1, 2, \ldots, p\}$, the selection probabilities are given by $\hat{\Pi}_k^{\lambda}, \lambda \in \Lambda$. Let $\hat{S}^{\Lambda}$=$\cup_{\lambda\in\Lambda}\hat{S}^{\lambda}$, be the set of selected variables if varying the regularization parameter λ in the set Λ. Let V be the number of falsely selected variables where V=$|N \cap \hat{S}^{\Lambda}| = |N_{\Lambda}|$. We now state the false positive control theorem. The proof builds upon results presented in Sec. 3.

Theorem 1. *Assume that the distribution of* $\{1_{\{k\in\hat{S}^{\lambda}\}}, k \in N\}$ *is exchangeable for all* $\lambda \in \Lambda$. *The expected number* V *of false positives for our approach is then bounded for* $\pi_{thr} \in (\frac{1}{2}, 1)$ *by*

$$\mathbb{E}(V) < \frac{1}{2\pi_{thr} - 1} \frac{\Big(\alpha + 9/4\Big)^2 |S|^2}{|N|} \tag{6}$$

where $\alpha > 0$.

Proof. Using the exchangeability assumption, we have that for all $k \in N$

$$\begin{aligned}\mathbb{P}(k \in \hat{S}^{\Lambda}) &= \mathbb{E}(|N_{\Lambda}|)/|N| \\ &= \mathbb{E}(|N \cap \hat{S}^{\Lambda}|)/|N|.\end{aligned}$$

Using Corollary 1,

$$\mathbb{P}(k \in \hat{S}^{\Lambda}) < \frac{\Big(\alpha + 9/4\Big)|S|}{|N|}.$$

This result is independent of the sample size used to construct $\hat{S}^{\lambda}$ for $\lambda \in \Lambda$. Using the second part of Lemma 4, it follows that

$$\mathbb{P}\{\max_{\lambda\in\Lambda}(\hat{\Pi}_K^{simult,\lambda}) \geq \xi\} < \frac{\left(\frac{\big(\alpha+9/4\big)|S|}{|N|}\right)^2}{\xi}$$

for all $0 < \xi < 1$ and $k \in N$. Using Lemma 3,

$$\mathbb{P}\{\max_{\lambda\in\Lambda}(\hat{\Pi}_k^{\lambda}) \geq \pi_{thr}\}$$
$$\leq \mathbb{P}[\{\max_{\lambda\in\Lambda}(\hat{\Pi}_K^{simult,\lambda}) + 1\}/2 \geq \pi_{thr}]$$
$$= \mathbb{P}[\{\max_{\lambda\in\Lambda}(\hat{\Pi}_K^{simult,\lambda}) \geq 2\pi_{thr} - 1].$$

Using the equation above by setting $\xi = 2\pi_{thr} - 1$

$$< \frac{1}{2\pi_{thr} - 1}\Bigg(\frac{\Big(\alpha + 9/4\Big)|S|}{|N|}\Bigg)^2.$$

Hence,

$$\mathbb{E}(V) = \sum_{k\in N} \mathbb{P}\{\max_{\lambda\in\Lambda}(\hat{\Pi}_k^{\lambda}) \geq \pi_{thr}\}$$
$$< \frac{1}{2\pi_{thr} - 1}\frac{\Big(\alpha + 9/4\Big)^2|S|^2}{|N|}$$

□

3. CONCAVE PENALTIES: THEORY

Let $h(t;\lambda;\gamma)$ represent an arbitrary regularizer from this family of penalties for a given scalar real valued variable t. We then assume that the following conditions hold on the penalty function:

1. $h(0;\lambda;\gamma) = 0$
2. $h(-t;\lambda;\gamma) = h(t;\lambda;\gamma)$
3. $h(t;\lambda;\gamma)$ is nondecreasing in t in $[0,\infty)$
4. $h(t;\lambda;\gamma)$ is subadditive with respect to t, $h(x+y;\lambda;\gamma) \leq h(x;\lambda;\gamma) + h(y;\lambda;\gamma)\ \forall\ x,y \geq 0$.

These have also been studied as *good penalties* among non-convex functions with Lipschitz sparsity yielding properties in [12]. We now define the threshold level of the penalty λ^* and the concavity parameter γ^* as follows:

$$\lambda^* = \inf_{t>0}\{t/2 + h(t;\lambda;\gamma)/t\} \qquad (7)$$
$$\gamma^* = \max_t h(t;\lambda;\gamma)/(\lambda^*)^2.$$

The value of this penalty evaluated for a specific regression coefficient vector $\beta \in \mathbb{R}^p$

$$\|h(\beta;\lambda;\gamma)\|_1 = \sum_{j=1}^{p} h(\beta_j;\lambda;\gamma) \qquad (8)$$
$$h_{MCP}(t;\lambda;\gamma) = \min\{\lambda t - t^2/2\gamma, \lambda^2\gamma/2\}.$$

Definition 2. *We now define an upper bound that bounds a general penalty from this family of penalties as given in* (8) *via ℓ_1 penalty for sparse vectors*

$$\Delta(a,k;\lambda;\gamma) = \sup\{\|h(\beta;\lambda;\gamma)\|_1 : \|\beta\|_1 \leq ak, \|\beta\|_0 = k\}. \qquad (9)$$

Proposition 1. *Let λ^* and γ^* be as defined in* (7)*. Then, the following inequality holds*

$$\Delta(a,k;\lambda;\gamma) \leq k\gamma^*(\lambda^*)^2. \qquad (10)$$

Proof. As $h(t;\lambda;\gamma)$ is concave in $[0,\infty)$ then by the Jensen's inequality $\Delta(a,k;\lambda;\gamma) \leq kh(a;\lambda;\gamma)$. Here we use the fact that k represents the number of non-zero regression coefficients as per the definition of the ℓ_0 norm. Now using the definition of γ^* from (7), we know that the maximum value of $h(a;\lambda;\gamma)$ is $(\gamma^*)(\lambda^*)^2$. Hence, $\Delta(a,k;\lambda;\gamma) \leq k\gamma^*(\lambda^*)^2$. Here, we can use the fact that $\gamma*_{MCP} = \gamma/2$ by setting $\lambda^* = \lambda$ and refine this bound for MCP as $\Delta_{MCP}(a,k;\lambda;\gamma) \leq k(\gamma/2)(\lambda)^2$. The γ value for the MCP is generally set to 3 in practice which expresses the bound as function of λ alone if needed. □

Assumption 1. *Let $\eta \in (0,1]$. We say that the regularization method in* (3) *satisfies the η-null-consistency condition (η-NC) if the following equality holds:*

$$\underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}(\|\varepsilon/\eta - X\beta\|_2^2/(2n) + \|h(\beta;\lambda;\gamma)\|_1) \qquad (11)$$
$$= \|\varepsilon/\eta\|_2^2/(2n).$$

Lemma 1. *If $\hat{\beta}$ is the global solution of* (3)*, then $\|X^T(y - X\hat{\beta})/n\|_\infty \leq \lambda^*$. In particular $\|X^T\varepsilon/n\|_\infty \leq \eta\lambda^*$ under the η-NC condition given in Assumption 1.*

Proof. Please refer to Proof of Lemma 1 from [4]. □

Lemma 2. *Assume the η-NC condition given in Assumption 1 holds for $\eta \in (0,1)$. Suppose $\hat{\beta} \in \mathbb{R}^p$ satisfy*

$$\|y - X\hat{\beta}\|_2^2/(2n) + \|h(\hat{\beta};\lambda;\gamma)\|_1 \leq \|y - X\beta\|_2^2/(2n) + \|h(\beta;\lambda;\gamma)\|_1 + \nu$$

with a certain $\nu > 0$. Let $\phi = \hat{\beta} - \beta$, $\varepsilon = \frac{1+\eta}{1-\eta}$ Then,

$$\|X\phi\|_2^2/(2n) + \|h(\phi_{S^c};\lambda;\gamma)\|_1 \leq \xi\|h(\phi_S;\lambda;\gamma)\|_1 + \nu/(1-\eta).$$

Proof. Please refer to Proof of Lemma 2 from [4].

□

Theorem 2. *Let $\{S, \hat{\beta}, \lambda^*, \gamma^*, \eta, \xi\}$ and $\Delta(a,k;\lambda;\gamma)$ be as defined in* (9) *and $\hat{S}_\Lambda$ represents the set of variables selected by the model while varying λ from λ_{max} to λ_{min}. Suppose that the η-NC condition given in Assumption 1 holds. Consider $t_0 \geq 0$, $m_0 \geq 0$ and the upper sparse eigenvalue for* *a matrix X defined as* $\kappa_+(m_0) := \max_{\|u\|_0 \le m_0 : \|u\|_2 = 1} \|Xu\|_2^2/n$ *for which the following holds:*

$$\sqrt{(2\xi\kappa_+(m_0)\Delta(a_1\lambda^*, |S|; \lambda; \gamma)/m_0)} + \|X^T\varepsilon/n\|_\infty \le \inf_{0<s<t_0} \dot{h}(s;\lambda;\gamma) \quad (12)$$

Then for $t_0 > 0$,

$$\mathbb{E}(|N \cap \hat{S}_\Lambda|) < m_0 + \lfloor \xi\Delta(a, |S|; \lambda; \gamma)/h(t_0;\lambda;\gamma) \rfloor \quad (13)$$

Proof. Let $\hat{S}_1 = \{j \in N \cap \hat{S}_\Lambda : |\hat{\beta}_j| \ge t_0\}$ and $\hat{S}_2 = \{j \in N \cap \hat{S}_\Lambda : |\hat{\beta}_j| < t_0\}$. It follows from (9), that $\|h(\phi_S; \lambda; \gamma)\|_1 \le \Delta(a_1, |S|; \lambda; \gamma)$ with the given a_1. Thus,

$$|\hat{S}_1| \le \|h(\phi_{S^c}; \lambda; \gamma)\|_1 / h(t_0; \lambda; \gamma)$$

Using Lemma 2

$$\le \xi \|h(\phi_S; \lambda; \gamma)\|_1 / h(t_0; \lambda; \gamma)$$
$$\le \xi \Delta(a_1, |S|; \lambda; \gamma)/h(t_0; \lambda; \gamma)$$

We now derive an upper bound on $|\hat{S}_2|$. Let $\lambda_2 > \sqrt{(2\xi\kappa_+(m_0)\Delta(a_1\lambda^*, |S|; \lambda; \gamma)/m_0)}$ satisfying

$$\lambda_2 + \|X^T\varepsilon/n\|_\infty \le \inf_{0<s<t_0} \dot{h}(s;\lambda;\gamma)$$

Here $\dot{h}$ represents the derivative of function h. The first order KKT optimality condition implies that for all $j \in \hat{S}_\Lambda, x_j^T(y - X\hat{\beta})/n = \dot{h}(t;\lambda;\gamma)|_{t=\hat{\beta}_j}$.

We construct this proof by proving that for $j \in \hat{S}_2, |\hat{\beta}_j| \in (0, t_0)$, where $|x_j^T(y - X\hat{\beta})/n| \ge \lambda_2 + \|X^T\varepsilon/\eta\|_\infty$ any set $A \subseteq \hat{S}_2$ satisfies $|A| \le m_0$.

$$(\lambda_2 + ||X^T\varepsilon/n||_\infty)|A| \le ||X_A^T(y - X\hat{\beta})/n||_1$$

Using Holder's inequality, the R.H.S can be expanded as

$$\le ||X^T\varepsilon/n||_\infty |A|$$
$$+ |A|^{1/2} ||X_A/\sqrt{n}||_2 ||X\phi||_2/\sqrt{n}.$$

Since, $||X_A/\sqrt{n}||_2^2 \le \kappa_+(m_0)$ from its definition.

$$\lambda_2|A| \le |A|^{1/2}\sqrt{(\kappa_+(m_0)||X\phi||_2^2/n)}$$
$$|A| \le \kappa_+(m_0)||X\phi||_2^2/n\lambda_2^2$$

Using Lemma 2

$$\le 2\xi\kappa_+(m_0)\Delta(a_1, |S|; \lambda; \gamma)/\lambda_2^2$$

Using our definition of λ_2 we obtain,

$$|A| < m_0$$

Thus, It follows that $\max_{A \subseteq \hat{S_2}, |A| \le m_0} |A| < m_0$ which implies that $|\hat{S_2}| < m_0$. Combining this with the bound for $|\hat{S_1}|$ completes the proof. □

Corollary 1. *Let* $h_{MCP}(t, \lambda; \gamma)$ *be the MCP penalty as defined in* (8) *and* γ^* *be as defined in* (7). *Suppose* (3), *is* $\eta - NC$ *and* $h(a, \lambda; \gamma) \ge \lambda(1 - \frac{a_1}{\gamma})$ *for some* $a > 0$ *and* $a_1 \ge 0$. *If* $m_0 = \alpha|S|$ *is an integer and* $2\gamma^*\kappa_+(\alpha|S|)/\alpha < (1 - a_1/\gamma - \eta)^2(1-\eta)/(1+\eta)$, *then*

$$\mathbb{E}(|N \cap \hat{S}_\Lambda|) < \Big(\alpha + 9/4\Big)|S| \quad (14)$$

Proof. Using Theorem 2, we can write that

$$\mathbb{E}[|N \cap \hat{S}_\Lambda|] < m_0 + \lfloor \xi\Delta(a, |S|, \lambda; \gamma)/h(t_0, \lambda; \gamma) \rfloor$$
$$< \alpha|S| + \xi\Delta(a, |S|, \lambda; \gamma)/h(t_0, \lambda; \gamma)$$

Using the fact that $h(t_0, \lambda; \gamma) \ge \lambda(1 - \frac{a_1}{\gamma})$ for $t_0 > 0$ we get

$$< \alpha|S| + \xi\Delta(a, |S|, \lambda; \gamma)/\lambda(1 - \frac{a_1}{\gamma})$$

Using Proposition 1, we know that,

$$\Delta(a, |S|, \lambda; \gamma) \le |S|\gamma^*(\lambda^*)^2$$
$$\mathbb{E}[|N \cap \hat{S}_\Lambda|] < \alpha|S| + \xi\frac{|S|\gamma^*(\lambda^*)^2)}{\lambda(1 - \frac{a_1}{\gamma})}$$

Using the fact that $\gamma^* = \frac{\gamma}{2}$ with setting $\lambda = \lambda^*$ where γ=3 for the MCP and setting a_1=1, this can be simplified as

$$= \alpha|S| + \frac{9|S|\xi\lambda}{4}.$$

As λ is non-negative and non-zero setting ξ in the range of $1/\lambda$ would make $\xi\lambda \sim 1$

$$= \Big(\alpha + 9/4\Big)|S|.$$

□

Lemma 3. *Lower bound for simultaneous selection probabilities. For any set* $K \subseteq \{1, \ldots, p\}$, *a lower bound for the simultaneous selection probabilities is given by*

$$\hat{\Pi}_K^{simult,\lambda} \ge 2\hat{\Pi}_K^\lambda - 1 \quad (15)$$

Proof. Please refer to Proof of sample splitting A.1 from [13]. □

Lemma 4. *Let* $K \subset \{1, 2, \ldots, p\}$ *and let* $\hat{S}^\lambda$ *be the set of selected variables based on a sample of size* $\lfloor \frac{n}{2} \rfloor$. *If* $\mathbb{P}(K \subseteq \hat{S}^\lambda) \le \varepsilon$, *then*

$$\mathbb{P}(\hat{\Pi}_K^{simult,\lambda} \ge \xi) \le \varepsilon^2/\xi$$

if $\mathbb{P}(K \subseteq \cup_{\lambda\in\Lambda}\hat{S}^\lambda) \le \varepsilon$ *for some* $\Lambda \subseteq \mathbb{R}^+$, *then*

$$\mathbb{P}\{\max_{\lambda\in\Lambda}(\hat{\Pi}_K^{simult,\lambda}) \ge \xi\} \le \varepsilon^2/\xi$$

Proof. Please refer to Proof of Lemma 2 from [13]. □

4. REFERENCES

[1] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," arXiv:1601.07996, 2016.

[2] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1490–1507, 2017.

[3] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[4] C. H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[5] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *Proceedings of the International Conference on Machine Learning*, Atlanta, GA, 2013, pp. 37–45.

[6] P. Wang, K. K. Padthe, B. Vinzamuri, and C. K. Reddy, "CRISP: Consensus regularized selection based prediction," in *Proceedings of the ACM Conference on Information and Knowledge Management*, Indianapolis, IN, 2016, pp. 1019–1028.

[7] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[8] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[9] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[10] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[11] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015.

[12] Z. Pan and C. Zhang, "High-dimensional inference via Lipschitz sparsity-yielding regularizers," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Scottsdale, AZ, 2013, pp. 481–488.

[13] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.