Journal of Machine Learning Research 13 (2012) 1189-1219 Submitted 3/11; revised 2/12; Published 4/12

# A Multi-Stage Framework for Dantzig Selector and LASSO

**Ji Liu** JI.LIU@ASU.EDU
**Peter Wonka** PETER.WONKA@ASU.EDU
**Jieping Ye** JIEPING.YE@ASU.EDU
*Arizona State University*
*699 South Mill Avenue*
*Tempe, AZ 85287-8809, USA*

**Editor:** Tong Zhang

## Abstract

We consider the following sparse signal recovery (or feature selection) problem: given a design matrix $X \in \mathbb{R}^{n\times m}$ $(m \gg n)$ and a noisy observation vector $y \in \mathbb{R}^n$ satisfying $y = X\beta^* + \varepsilon$ where $\varepsilon$ is the noise vector following a Gaussian distribution $N(0, \sigma^2 I)$, how to recover the signal (or parameter vector) $\beta^*$ when the signal is sparse?

The Dantzig selector has been proposed for sparse signal recovery with strong theoretical guarantees. In this paper, we propose a multi-stage Dantzig selector method, which iteratively refines the target signal $\beta^*$. We show that if $X$ obeys a certain condition, then with a large probability the difference between the solution $\hat{\beta}$ estimated by the proposed method and the true solution $\beta^*$ measured in terms of the $\ell_p$ norm ($p \geq 1$) is bounded as

$$\|\hat{\beta} - \beta^*\|_p \leq \left(C(s-N)^{1/p}\sqrt{\log m} + \Delta\right)\sigma,$$

where $C$ is a constant, $s$ is the number of nonzero entries in $\beta^*$, the risk of the oracle estimator $\Delta$ is independent of $m$ and is much smaller than the first term, and $N$ is the number of entries of $\beta^*$ larger than a certain value in the order of $O(\sigma\sqrt{\log m})$. The proposed method improves the estimation bound of the standard Dantzig selector approximately from $Cs^{1/p}\sqrt{\log m}\sigma$ to $C(s-N)^{1/p}\sqrt{\log m}\sigma$ where the value $N$ depends on the number of large entries in $\beta^*$. When $N = s$, the proposed algorithm achieves the oracle solution with a high probability, where the oracle solution is the projection of the observation vector $y$ onto true features. In addition, with a large probability, the proposed method can select the same number of correct features under a milder condition than the Dantzig selector. Finally, we extend this multi-stage procedure to the LASSO case.

**Keywords:** multi-stage, Dantzig selector, LASSO, sparse signal recovery

## 1. Introduction

The sparse signal recovery problem has been studied in many areas including machine learning (Zhang, 2009b; Zhao and Yu, 2006), signal processing (Donoho et al., 2006; Romberg, 2008; Wainwright, 2009), and mathematics/statistics (Bunea et al., 2007; Candès and Plan, 2009; Candès and Tao, 2007; Koltchinskii and Yuan, 2008; Lounici, 2008; Meinshausen et al., 2006; Ravikumar et al., 2008; Zhang, 2009a). In the sparse signal recovery problem, one is mainly interested in the signal recovery accuracy, that is, the distance between the estimation $\hat{\beta}$ and the original signal or the true solution $\beta^*$. If the design matrix $X$ is considered as a feature matrix, that is, each column is a feature vector, and the observation $y$ as a target object vector, then the sparse signal recovery problem is

equivalent to feature selection (or model selection). In feature selection, one concerns the feature selection accuracy. Typically, a group of features corresponding to the coefficient values in $\hat{\beta}$ larger than a threshold form the supporting feature set. The difference between this set and the true supporting set (i.e., the set of features corresponding to nonzero coefficients in the original signal) measures the feature selection accuracy.

Two well-known algorithms for learning sparse signals include LASSO (Tibshirani, 1996) and Dantzig selector (Candès and Tao, 2007):

$$\textbf{LASSO} \qquad \min_{\beta} : \ \frac{1}{2}\|X\beta - y\|_2^2 + \lambda' ||\beta||_1,$$

$$\textbf{Dantzig Selector} \qquad \min_{\beta} : ||\beta||_1$$
$$s.t. : \|X^T(X\beta - y)\|_\infty \leq \lambda.$$

Strong theoretical results concerning LASSO and Dantzig selector have been established in the literature (Cai et al., 2009; Candès and Plan, 2009; Candès and Tao, 2007; Wainwright, 2009; Zhang, 2009a; Zhao and Yu, 2006).

### 1.1 Contributions

In this paper, we propose a multi-stage procedure based on the Dantzig selector, which estimates the supporting feature set $F_0$ and the signal $\hat{\beta}$ iteratively. The intuition behind the proposed multi-stage method is that feature selection and signal recovery are tightly correlated and they can benefit from each other: a more accurate estimation of the supporting features can lead to a better signal recovery and a more accurate signal recovery can help identify a better set of supporting features. In the proposed method, the supporting set $F_0$ starts from an empty set and its size increases by one after each iteration. At each iteration, we employ the basic framework of Dantzig selector and the information about the current supporting feature set $F_0$ to estimate the new signal $\hat{\beta}$. In addition, we select the supporting feature candidates in $F_0$ among all features in the data at each iteration, thus allowing to remove incorrect features from the previous supporting feature set.

The main contributions of this paper lie in the theoretical analysis of the proposed method. Specifically, we show: 1) the proposed method can improve the estimation bound of the standard Dantzig selector approximately from $Cs^{1/p}\sqrt{\log m}\sigma$ to $C(s-N)^{1/p}\sqrt{\log m}\sigma$ where the value $N$ depends on the number of large entries in $\beta^*$; 2) when $N = s$, the proposed algorithm can achieve the oracle solution $\bar{\beta}$ with a high probability, where the oracle solution is the projection of the observation vector $y$ onto true features (see Equation (1) for the explicit description of $\bar{\beta}$); 3) with a high probability, the proposed method can select the same number of correct features under a milder condition than the standard Dantzig selector method; 4) this multi-stage procedure can be easily extended to the LASSO case. The numerical experiments validate these theoretical results.

### 1.2 Related Work

Sparse signal recovery without observation noise was studied by Candès and Tao (2005), which showed under the restricted isometry property (RIP) sparse signals can be perfectly recovered by solving an $\ell_1$ norm minimization problem. LASSO and Dantzig selector can be considered as its noisy versions. Zhao and Yu (2006) proved the feature selection consistency of LASSO under the irrepresentable condition. It was also shown by Candès and Plan (2009) that if the true signal

is strong enough together with some additional assumptions on its supporting set and signs, the mutual incoherence property (MIP) (or incoherence condition) can guarantee the feature selection consistency and the sign consistency with a high probability. A comprehensive analysis for LASSO, including the recovery accuracy in an arbitrary $\ell_p$ norm ($p \geq 1$) and the feature selection consistency, was presented in Zhang (2009a). Candès and Tao (2007) proposed the Dantzig selector (which is a linear programming problem) for sparse signal recovery and presented a bound of recovery accuracy with the same order as LASSO under the uniform uncertainty principle (UUP). An approximate equivalence between the LASSO estimator and the Dantzig selector was given by Bickel et al. (2009). Lounici (2008) studied the $\ell_\infty$ convergence rate for LASSO and Dantzig estimators in a high-dimensional linear regression model under MIP. James et al. (2009) provided conditions on the design matrix $X$ under which the LASSO and Dantzig selector coefficient estimates are identical for certain tuning parameters. Please refer to recent papers (Zhang, 2009a; Fan and Lv, 2010) for a more comprehensive overview of LASSO and Dantzig selector.

Since convex regularization methods like LASSO and Dantzig selector give biased estimation due to convex regularization, many heuristic methods have been proposed to correct the bias of convex relaxation recently, including orthogonal matching pursuit (OMP) (Tropp, 2004; Donoho et al., 2006; Zhang, 2009b, 2011a; Cai and Wang, 2011), two stage LASSO (Zhang, 2009a), multiple thresholding LASSO (Zhou, 2009), adaptive LASSO (Zou, 2006), adaptive forward-backward greedy method (FoBa) (Zhang, 2011b), and nonconvex regularization methods (Zhang, 2010b; Fan and Lv, 2011; Lv and Fan, 2009; Zhang, 2011b). They have been shown to outperform the standard convex methods in many practical applications. It was shown that under exact recovery condition (ERC) (similar to MIP) the solution of OMP guarantees the feature selection consistency in the noiseless case (Tropp, 2004). The results of Tropp (2004) were extended to the noisy case by Zhang (2009b). Very recently, Zhang (2011a) showed that under RIP (weaker than MIP and ERC), OMP can stably recover a sparse signal in 2-norm under measurement noise. A multiple thresholding procedure was proposed to refine the solution of LASSO or Dantzig selector (Zhou, 2009). The FoBa algorithm was proposed by Zhang (2011b), and it was shown that under RIP the feature selection consistency is achieved if the minimal nonzero entry in the true solution is larger than $O(\sigma\sqrt{\log m})$. The adaptive LASSO was proposed to adaptively tune the weight value for the $\ell_1$ norm penalty, and it was shown to enjoy the oracle properties (Zou, 2006). Zhang (2010b) proposed a general multi-stage convex regularization method (MSCR) to solve a nonconvex sparse regularization problem. It was also shown that a specific case "least square loss + nonconvex sparse regularization" can eliminate the bias in signal recovery (Zhang, 2010b) and achieve the feature selection consistency (Zhang, 2011c) under the sparse eigenvalue condition (SEC) if the true signal is strong enough. More related work about nonconvex regularization methods can be found in a recent paper by Zhang and Zhang (2012).

Conditions mentioned above can be classified into two classes: 1) the $\ell_2$ conditions including RIP, UUP, and SEC; 2) the $\ell_\infty$ conditions including ERC and MIP. Overall, the $\ell_2$ conditions are considered to be weaker than the $\ell_\infty$ conditions, since the $\ell_\infty$ conditions require about $O(s^2 \log m)$ random projections while the $\ell_2$ conditions only need $O(s \log m)$ random projections.

### 1.3 Definitions, Notations, and Basic Assumptions

We use $X \in \mathbb{R}^{n \times m}$ to denote the design matrix and focus on the case $m \gg n$, that is, the signal dimension is much larger than the observation dimension. The correlation matrix $A$ is defined as

$A = X^T X$ with respect to the design matrix. The noise vector $\varepsilon$ follows the multivariate normal distribution $\varepsilon \sim N(0, \sigma^2 I)$. The observation vector $y \in \mathbb{R}^n$ satisfies $y = X\beta^* + \varepsilon$, where $\beta^*$ denotes the original signal (or true solution). $\hat{\beta}$ is used to denote the solution of the proposed algorithm. The $\alpha$-supporting set ($\alpha \geq 0$) for a vector $\beta$ is defined as

$$supp_\alpha(\beta) = \{j : |\beta_j| > \alpha\}.$$

The "supporting" set of a vector refers to the 0-supporting set. $F$ denotes the supporting set of the original signal $\beta^*$. For any index set $S$, $|S|$ denotes the size of the set and $\bar{S}$ denotes the complement of $S$ in $\{1,2,3,...,m\}$. In this paper, $s$ is used to denote the size of the supporting set $F$, that is, $s = |F|$. We use $\beta_S$ to denote the subvector of $\beta$ consisting of the entries of $\beta$ in the index set $S$. The $\ell_p$ norm of a vector $v$ is computed by $\|v\|_p = (\sum_i |v_i|^p)^{1/p}$, where $v_i$ denotes the $i$th entry of $v$. The oracle solution $\bar{\beta}$ is defined as

$$\bar{\beta}_F = (X_F^T X_F)^{-1} X_F^T y \text{ and } \bar{\beta}_{\bar{F}} = 0. \tag{1}$$

We employ the following notation to measure some properties of a PSD matrix $M \in \mathbb{R}^{K\times K}$ (Zhang, 2009a):

$$\begin{aligned} \mu_{M,k}^{(p)} &= \inf_{u\in\mathbb{R}^k,|I|=k} \frac{\|M_{I,I}u\|_p}{\|u\|_p}, \quad \rho_{M,k}^{(p)} = \sup_{u\in\mathbb{R}^k,|I|=k} \frac{\|M_{I,I}u\|_p}{\|u\|_p}, \\ \theta_{M,k,l}^{(p)} &= \sup_{u\in\mathbb{R}^l,|I|=k,|J|=l,I\cap J=\varnothing} \frac{\|M_{I,J}u\|_p}{\|u\|_p}, \quad \gamma_M = \max_{i\neq j} |M_{ij}|, \end{aligned}$$

where $p \in [1,\infty]$, $I$ and $J$ are disjoint subsets of $\{1,2,...,K\}$, and $M_{I,J} \in \mathbb{R}^{|I|\times|J|}$ is a submatrix of $M$ with rows from the index set $I$ and columns from the index set $J$. One can easily verify that $\mu_{A,k}^{(\infty)} \geq 1-\gamma_A(k-1)$, $\rho_{A,k}^{(\infty)} \leq 1+\gamma_A(k-1)$, and $\theta_{A,k,l}^{(\infty)} \leq l\gamma_A$, if all columns of $X$ are normalized to have a unit length.

Additionally, we use the following notation to denote two probabilities:

$$\eta_1' = \eta_1(\pi \log((m-s)/\eta_1))^{-1/2}, \quad \eta_2' = \eta_2(\pi \log(s/\eta_2))^{-1/2},$$

where $\eta_1$ and $\eta_2$ are two factors between 0 and 1. In this paper, if we say "large", "larger" or "the largest", it means that the absolute value is large, larger or the largest. For simpler notation in the computation of sets, we sometimes use "$S_1 + S_2$" to indicate the union of two sets $S_1$ and $S_2$, and use "$S_1 - S_2$" to indicate the removal of the intersection of $S_1$ and $S_2$ from the first set $S_1$. In this paper, the following assumption is always admitted.

**Assumption 1** *We assume that $s = |supp_0(\beta^*)| < n$, the variable number is much larger than the feature dimension (i.e., $m \gg n$), each column vector is normalized as $X_i^T X_i = 1$ where $X_i$ indicates the ith column (or feature) of $X$, and the noise vector $\varepsilon$ follows the Gaussian distribution $N(0,\sigma^2 I)$.*

In the literature, it is often assumed that $X_i^T X_i = n$, which is essentially identical to our assumption. However, this may lead to a slight difference of a factor $\sqrt{n}$ in some conclusions. We have automatically transformed conclusions from related work according to our assumption when citing them in our paper.

### 1.4 Organization

The rest of the paper is organized as follows. We present our multi-stage algorithm in Section 2. The main theoretical results are summarized in Section 3 with detailed proofs given in Appendix A (for Dantzig selector) and Appendix B (for LASSO). The numerical simulation is reported in Section 4. Finally, we conclude the paper in Section 5.

## 2. The Multi-Stage Dantzig Selector Algorithm

In this section, we introduce the multi-stage Dantzig selector algorithm. In the proposed method, we update the support set $F_0$ and the estimation $\hat{\beta}$ iteratively; the supporting set $F_0$ starts from an empty set and its size increases by one after each iteration. At each iteration, we employ the basic framework of Dantzig selector and the information about the current supporting set $F_0$ to estimate the new signal $\hat{\beta}$ by solving the following linear program:

$$\begin{aligned} &\min \|\beta_{\bar{F}_0}\|_1 \\ &s.t.\ \|X_{F_0}^T(X\beta - y)\|_\infty \le \lambda \\ &\quad \|X_{F_0}^T(X\beta - y)\|_\infty = 0. \end{aligned} \tag{2}$$

Since the features in $F_0$ are considered as the supporting candidates, it is natural to enforce them to be orthogonal to the residual vector $X\beta - y$, that is, one should make use of them for reconstructing the overestimation $y$. This is the rationale behind the constraint: $\|X_{F_0}^T(X\beta - y)\|_\infty = 0$. The other advantage is when all correct features (i.e., the true feature set $F$) are chosen, the proposed algorithm can be shown to converge to the oracle solution. In other words, the oracle solution satisfies this constraint with $F$. The detailed procedure is formally described in **Algorithm 1** below. Apparently, when $F_0^{(0)} = \varnothing$ and $N = 0$, the proposed method is identical to the standard Dantzig selector.

**Algorithm 1** Multi-Stage Dantzig Selector

**Require:** $F_0^{(0)}$, $\lambda$, $N$, $X$, $y$
**Ensure:** $\hat{\beta}^{(N)}$, $F_0^{(N)}$
1: **while** i=0; i≤N; i++ **do**
2: Obtain $\hat{\beta}^{(i)}$ by solving the problem (2) with $F_0 = F_0^{(i)}$;
3: Form $F_0^{(i+1)}$ as the index set of the $i+1$ largest elements of $\hat{\beta}^{(i)}$;
4: **end while**

## 3. Main Results

This section introduces the main results of this paper and discusses some of their implications. The proofs are provided in the Appendix.

### 3.1 Motivation

To motivate the proposed multi-stage algorithm, we first consider a simple case where some knowledge about the supporting features is known in advance. In standard Dantzig selector, we assume

$F_0 = \varnothing$. If we assume that the features belonging to a set $F_0$ are known as supporting features, that is, $F_0 \subset F$, we have the following result:

**Theorem 1** *Assume that Assumption 1 holds. Take* $F_0 \subset F$ *and* $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ *in the optimization problem* (2)*. If there exists some* $l$ *such that*

$$\mu^{(p)}_{A,s+l} - \theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0 - \bar{F}|}{l}\right)^{1-1/p} > 0$$

*holds, then with a probability larger than* $1-\eta_1'$*, the* $\ell_p$ *norm* ($1 \le p \le \infty$) *of the difference between* $\hat{\bar{\beta}}$*, the solution of the problem* (2)*, and the oracle solution* $\bar{\beta}$ *is bounded as*

$$\|\hat{\bar{\beta}} - \bar{\beta}\|_p \le \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}(|\bar{F}_0 - \bar{F}| + l2^p)^{1/p}}{\mu^{(p)}_{A,s+l} - \theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda \tag{3}$$

*and with a probability larger than* $1-\eta_1'-\eta_2'$*, the* $\ell_p$ *norm* ($1 \le p \le \infty$) *of the difference between* $\hat{\bar{\beta}}$*, the solution of the problem* (2) *and the true solution* $\beta^*$ *is bounded as*

$$\begin{aligned}\|\hat{\bar{\beta}} - \beta^*\|_p \le & \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}(|\bar{F}_0 - \bar{F}| + l2^p)^{1/p}}{\mu^{(p)}_{A,s+l} - \theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda + \\ & \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.\end{aligned} \tag{4}$$

It is clear that both bounds (for any $1 \le p \le \infty$) are monotonically increasing with respect to the value of $|\bar{F}_0 - \bar{F}|$. In other words, the larger $F_0$ is, the lower these bounds are. This coincides with our motivation that more knowledge about the supporting features can lead to a better signal estimation. Most related literatures directly estimate the bound of $\|\hat{\bar{\beta}} - \beta^*\|_p$. Since $\beta^*$ may not be a feasible solution of problem (2), it is not easy to directly estimate the distance between $\hat{\bar{\beta}}$ and $\beta^*$.

The bound given in the inequality (4) consists of two terms. Since $m \gg n > s$, we have $\sqrt{2\log((m-s)/\eta_1)} \gg \sqrt{2\log(s/\eta_2)}$ if $\eta_1 \approx \eta_2$. When $p = 2$, the following holds:

$$\mu^{(2)}_{A,s+l} - \theta^{(2)}_{A,s+l,l}\left(\frac{|\bar{F}_0 - \bar{F}|}{l}\right)^{1-1/2} \le \mu^{(2)}_{(X_F^T X_F)^{1/2},s}$$

due to the following relationships:

$$\mu^{(2)}_{A,s+l} \le \mu^{(2)}_{A,s} \le \mu^{(2)}_{X_F^T X_F,s} \le \mu^{(2)}_{(X_F^T X_F)^{1/2},s}.$$

From the analysis in the next section, we can see that the first term is the upper bound of the distance from the optimizer to the oracle solution, that is, $\|\hat{\bar{\beta}} - \bar{\beta}\|_p$ and the second term is the upper bound of the distance from the oracle solution to the true solution, that is, $\|\bar{\beta} - \beta^*\|_p$.[1] Thus, the first term may be much larger than the second term under the assumption $m \gg n > s$.

1. The presented bound for $\|\bar{\beta} - \hat{\bar{\beta}}\|_p$ can be sharper for a particular value of $p$, for example, $\|\bar{\beta} - \beta^*\|_2 \le O(\sigma\sqrt{s})$, $\|\bar{\beta} - \beta^*\|_\infty \le O(\sigma\sqrt{\log s})$ (Zhang, 2009b). For simplicity, a general bound $\|\bar{\beta} - \beta^*\|_p \le O(\sigma s^{1/p}\sqrt{\log s})$ is used in this paper.

### 3.2 Comparison with Dantzig Selector

We first compare our estimation bound with the one derived by Candès and Tao (2007) for $p = 2$. For convenience of comparison, we rewrite their theorem (Candès and Tao, 2007) equivalently as:

**Theorem 2** *Suppose $\beta \in \mathbb{R}^m$ is any s-sparse vector of parameters obeying $\delta_{2s} + \theta^{(2)}_{A,s,2s} < 1$. Setting $\lambda_p = \sigma\sqrt{2\log(m/\eta)}$ ($0 < \eta \leq 1$), with a probability at least $1 - \eta(\pi \log m)^{-1/2}$, the solution of the standard Dantzig selector $\hat{\beta}_D$ obeys*

$$\|\hat{\beta}_D - \beta^*\|_2 \leq \frac{4}{1 - \delta_{2s} - \theta^{(2)}_{A,s,2s}} s^{1/2} \sigma \sqrt{2\log(m/\eta)}, \tag{5}$$

*where $\delta_{2s} = \max(\rho^{(2)}_{A,2s} - 1, 1 - \mu^{(2)}_{A,2s})$.*

In order to compare Theorem 1 with the result above, taking $l = |\bar{F}_0 - \bar{F}| \leq s$, $p = 2$, $\eta_1 = \frac{m-s}{m}\eta$, and $\eta_2 = \frac{s}{m}\eta$ in Theorem 1, we obtain that

$$\|\hat{\beta} - \beta^*\|_2 \leq \left( \frac{\sqrt{10l}}{\mu^{(2)}_{A,s+l} - \theta^{(2)}_{A,s+l,l}} + \frac{\sqrt{s}}{\mu^{(2)}_{(X_F^T X_F)^{1/2},s}} \right) \sigma \sqrt{2\log(m/\eta)} \tag{6}$$

holds with probability larger than $1 - \eta(\pi \log m)^{-1/2}$. It is easy to verify that

$$1 - \delta_{2s} - \theta^{(2)}_{A,s,2s} \leq \mu^{(2)}_{A,s+l} - \theta^{(2)}_{A,s+l,s} \leq \mu^{(2)}_{A,2s} \leq \mu^{(2)}_{(X_F^T X_F),s} = \left(\mu^{(2)}_{(X_F^T X_F)^{1/2},s}\right)^2 \leq \mu^{(2)}_{(X_F^T X_F)^{1/2},s} \leq 1.$$

When $F_0 = \varnothing$, the bound in (6) is comparable to the one in (5). Since $\mu^{(2)}_{A,s+l} - \theta^{(2)}_{A,s+l,l}$ in Equation (6) is a decreasing function in terms of $l$, if $F_0$ is nonempty, particularly if $F_0$ is close to $F$ (i.e., $l$ is close to 0), the condition $\mu^{(2)}_{A,s+l} - \theta^{(2)}_{A,s+l,l} > 0$ required in Equation (6) is much easier to satisfy than the condition $1 - \delta_{2s} - \theta^{(2)}_{A,s,2s} > 0$ required in Equation (5).

### 3.3 Feature Selection

The estimation bounds in Theorem 1 assume that a set $F_0$ is given. In this section, we show how the supporting set can be estimated. Similar to previous work (Candès and Plan, 2009; Zhang, 2009b), $|\beta^*_j|$ for $j \in F$ is required to be larger than a threshold value. As is clear from the proof in **Appendix A**, the threshold value $\alpha_0$ is actually proportional to the value of $\|\hat{\beta} - \beta^*\|_\infty$. We essentially employ the result with $p = \infty$ in Theorem 1 to estimate the threshold value. It shows that the value of $\|\hat{\beta} - \beta^*\|_\infty$ is bounded by $O(\lambda)$, which is consistent with the result of Lounici (2008). In the following, we first consider the simple case when $N = 0$. We have shown in the last section that the estimation bound in this case is similar to the one for Dantzig selector.

**Theorem 3** *Under the Assumption 1, if there exist a nonempty set*

$$\Omega = \{l \mid \mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right) > 0\}$$

*and an index set $J$ such that $|\beta_j^*| > \alpha_0$ for any $j \in J$, where*

$$\begin{aligned}\alpha_0 =& \|\hat{\beta}^{(0)} - \beta^*\|_\infty + \|\hat{\beta}^{(0)} - \bar{\beta}\|_\infty \\ \leq& 4 \min_{l \in \Omega} \frac{\max\left(1, \frac{s}{l}\right)}{\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right)} \lambda + \frac{1}{\mu^{(\infty)}_{(X_F^T X_F)^{1/2},s}} \sigma \sqrt{2\log(s/\eta_2)},\end{aligned}$$

*then taking $F_0 = \varnothing$, $N = 0$, $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into the problem* (2) *(equivalent to Dantzig selector), the largest $|J|$ elements of $\hat{\beta}_{std}$ (or $\hat{\beta}^{(0)}$) belong to $F$ with probability larger than $1 - \eta_1' - \eta_2'$.*

The theorem above indicates that under the given condition, if $\min_{j \in J} |\beta_j^*| > O(\sigma\sqrt{\log m})$ (assuming that there exists $l \geq s$ such that $\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right) > 0$), then with high probability the selected $|J|$ features by Dantzig selector belong to the true supporting set. In particular, if $|J| = s$, then the consistency of feature selection is achieved. In order to build up a link to the previous work, we let $l = s$. Note that $\mu^{(\infty)}_{A,2s} - \theta^{(\infty)}_{A,2s,s} \geq 1 - \gamma_A(3s-1)$. If the MIP holds like $\gamma_A s \leq 1/6$ (see Corollary 8.1 in Zhang, 2009a), then the condition required in Theorem 3 is satisfied as well. It means that the condition we require is not stronger than MIP. However, it still belongs to the $\ell_\infty$ condition like MIP. The result above is comparable to the ones for other feature selection algorithms, including LASSO/two stage LASSO (Candès and Plan, 2009; Zhao and Yu, 2006), OMP (Tropp, 2004; Donoho et al., 2006; Zhang, 2009b), and two stage LASSO (Zhang, 2009a). In all these algorithms, the conditions $\min_{j \in F} |\beta_j^*| \geq C\sigma\sqrt{\log m}$ and an $\ell_\infty$ condition are required. As pointed out by Zhang and Zhang (2012) and Zhang (2011a), these conditions required by OMP, Dantzig selector, and LASSO in feature selection cannot be improved. If one wants to use the $\ell_2$ conditions in feature selection, the minimal nonzero entry of the true solution must be in the order of $O(\sigma\sqrt{s\log m})$, which can be obtained by simply using $\|\hat{\beta}^{(0)} - \beta^*\|_\infty + \|\hat{\beta}^{(0)} - \bar{\beta}\|_\infty \leq \|\hat{\beta}^{(0)} - \beta^*\|_2 + \|\hat{\beta}^{(0)} - \bar{\beta}\|_2$. A similar requirement under the $\ell_2$ condition for LASSO (or two stage LASSO) is also implied by Zhang (2009a, Theorem 8.1).

Next, we show that the condition $|\beta_j^*| > \alpha_0$ in Theorem 3 can be relaxed by the proposed multi-stage procedure with $N > 0$, as summarized in the following theorem:

**Theorem 4** *Under the Assumption 1, if there exist a nonempty set*

$$\Omega = \{l \mid \mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right) > 0\}$$

*and a set $J$ such that $|supp_{\alpha_i}(\beta_J^*)| > i$ holds for all $i \in \{0, 1, ..., |J| - 1\}$, where*

$$\begin{aligned}\alpha_i =& \|\hat{\beta}^{(i)} - \beta^*\|_\infty + \|\hat{\beta}^{(i)} - \bar{\beta}\|_\infty \\ \leq& 4 \min_{l \in \Omega} \frac{\max\left(1, \frac{s-i}{l}\right)}{\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s-i}{l}\right)} \lambda + \frac{1}{\mu^{(\infty)}_{(X_F^T X_F)^{1/2},s}} \sigma \sqrt{2\log(s/\eta_2)},\end{aligned}$$

*then taking $F_0^{(0)} = \varnothing$, $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ and $N = |J| - 1$ into* **Algorithm** *1, the solution after $N$ iterations satisfies $F_0^{(N)} \subset F$ (i.e., $|J|$ correct features are selected) with probability larger than $1 - \eta_1' - \eta_2'$.*

Assume that one aims to select $N$ correct features by the standard Dantzig selector and the multi-stage method. These two theorems show that the standard Dantzig selector requires that at least $N$ of $|\beta^*_j|$'s with $j \in F$ are larger than the threshold value $\alpha_0$, while the proposed multi-stage method requires that at least $i$ of the $|\beta^*_j|$'s are larger than the threshold value $\alpha_{i-1}$, for $i = 1, \cdots, N$. Since the upper bounds of $\{\alpha_j\}$'s strictly decrease and the difference of two neighbors is greater than

$$\frac{4\theta^{(\infty)}_{A,s+l,l}}{l\left(\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s-i}{l}\right)\right)^2}\lambda$$

for some $l \in \Omega$, the proposed multi-stage method requires a strictly weaker condition for selecting $N$ correct features than the standard Dantzig selector. If we consider the $\ell_2$ conditions, using $\|\hat{\beta}^{(i)} - \beta^*\|_\infty + \|\hat{\beta}^{(i)} - \bar{\beta}\|_\infty \leq \|\hat{\beta}^{(i)} - \beta^*\|_2 + \|\hat{\beta}^{(i)} - \bar{\beta}\|_2$ to bound $\alpha_i$, we obtain that $\alpha_i \leq O(\sqrt{(s-i)\log m} + \Delta)\sigma$ where $\Delta$ is a small number relying on $s$. When $i$ is close to $s$, the order of $\alpha_i$ approaches $O(\sigma\sqrt{\log m})$. Recall that the FoBa algorithm (Zhang, 2011b), MSCR (Zhang, 2011c), and MC+ (Zhang, 2010a) require an $\ell_2$ condition and the threshold value is in the order of $O(\sigma\sqrt{\log m})$ for the feature selection consistency while the standard LASSO or Dantzig selector requires the threshold value in the order of $O(\sigma\sqrt{s\log m})$. Therefore, our condition lies between them.

### 3.4 Signal Recovery

In this section, we derive the estimation bound of the proposed multi-stage method by combining results from Theorems 1, 3, and 4.

**Theorem 5** *Under the Assumption 1, if there exist $l$ such that*

$$\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right) > 0 \;\; \textit{and} \;\; \mu^{(p)}_{A,2s} - \theta^{(p)}_{A,2s,s} > 0,$$

*and a set $J$ such that $|supp_{\alpha_i}(\beta^*_J)| > i$ holds for all $i \in \{0,1,...,|J|-1\}$, where the $\alpha_i$'s are defined in Theorem 4, then*

*(1) taking $F_0 = \varnothing$, $N = 0$ and $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into* **Algorithm** *1, with probability larger than $1 - \eta'_1 - \eta'_2$, the solution of the Dantzig selector $\hat{\beta}_D$ (i.e., $\hat{\beta}^{(0)}$) obeys:*

$$\|\hat{\beta}_D - \beta^*\|_p \leq \frac{(2^{p+1}+2)^{1/p}s^{1/p}}{\mu^{(p)}_{A,2s} - \theta^{(p)}_{A,2s,s}}\lambda + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)};$$

*(2) taking $F_0 = \varnothing$, $N = |J|$ and $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into* **Algorithm** *1, with probability larger than $1 - \eta'_1 - \eta'_2$, the solution of the multi-stage method $\hat{\beta}_{mul}$ (i.e., $\hat{\beta}^{(N)}$) obeys:*

$$\|\hat{\beta}_{mul} - \beta^*\|_p \leq \frac{(2^{p+1}+2)^{1/p}(s-N)^{1/p}}{\mu^{(p)}_{A,2s-N} - \theta^{(p)}_{A,2s-N,s-N}}\lambda + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.$$

Similar to the analysis in Theorem 1, the first term (i.e., the distance from $\hat{\beta}$ to the oracle solution $\bar{\beta}$) dominates in the estimated bounds. Thus, the performance of the multi-stage method approximately improves the standard Dantzig selector from $Cs^{1/p}\sqrt{\log m}\sigma$ to $C(s-N)^{1/p}\sqrt{\log m}\sigma$. When $p = 2$, our estimation has the same order as FoBa (Zhang, 2011b) and MCSR (Zhang, 2010b), but the conditions involved in our estimation belong to the $\ell_\infty$ class while they use the $\ell_2$ condition.

### 3.5 The Oracle Solution

The oracle solution $\hat{\beta}$ defined in Equation (1) is the minimum-variance unbiased estimator of the true solution given the noisy observation. We show in the following theorem that the proposed method can obtain the oracle solution with high probability under certain conditions:

**Theorem 6** *Under the assumption 1, if there exists $l$ such that $\mu^{(\infty)}_{A,s+l} - \theta^{(\infty)}_{A,s+l,l}\left(\frac{s-i}{l}\right) > 0$, and the supporting set $F$ of $\beta^*$ satisfies $|supp_{\alpha_i}(\beta^*_F)| > i$ for all $i \in \{0,1,...,s-1\}$, where the $\alpha_i$'s are defined in Theorem 4, then taking $F_0 = \varnothing$, $N = s$ and $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into* **Algorithm** *1, the oracle solution can be achieved, that is, $F_0^{(N)} = F$ and $\hat{\beta}^{(N)} = \bar{\beta}$, with probability larger than $1-\eta_1'-\eta_2'$.*

The theorem above shows that when the nonzero elements of the true coefficients vector $\beta^*$ are large enough, the oracle solution can be achieved with high probability.

### 3.6 The Multi-Stage LASSO Algorithm

Next we extend the multi-stage procedure to the LASSO case; we expect to achieve similar improvements over the standard LASSO. The multi-stage LASSO algorithm can be obtained by substituting the basic optimization problem, that is, Equation (2) in **Algorithm** 1, by the following problem:

$$\begin{aligned} \min_{\beta}:\ & \frac{1}{2}\|X\beta - y\|_2^2 + \lambda'||\beta_{\bar{F}_0}||_1 \\ s.t.:\ & \|X^T_{F_0}(X\beta - y)\|_\infty = 0. \end{aligned} \tag{7}$$

Note that the constraint in Equation (7) is satisfied automatically at the optimal solution by observing the subdifferential of its objective function. Thus, the constraint can be removed from Equation (7) in practice.

We apply the same framework in Dantzig selector to analyze the multi-stage LASSO to obtain a bound estimation for any $p \in [1,\infty]$ and show that similar improvements can be achieved over the standard LASSO. For completeness, we include all proofs and results for multi-stage LASSO in **Appendix B**.

It is worth mentioning that Zhang (2010b, 2011b) recently developed a similar method called MSCR. The main difference is that it uses a threshold value to update the candidate set $F_0^{(i+1)}$ at each iteration and may need to solve LASSO more than $s$ times to converge, while our algorithm needs to solve LASSO less than $s$ times. An advantage of MSCR is that it requires a weaker condition, that is, $\min_{i\in F}|\beta^*| > O(\sigma\sqrt{\log m})$ and an $\ell_2$ condition, to achieve the consistency on feature selection and signal recovery.

## 4. Simulation Study

We have performed simulation studies to verify our theoretical analysis. Our comparison includes two aspects: signal recovery accuracy and feature selection accuracy. The signal recovery accuracy is measured by the relative signal error: $SRA = -20\log_{10}(\|\hat{\beta}-\beta^*\|_2/\|\beta^*\|_2)$, where $\hat{\beta}$ is the solution of a specific algorithm. The feature selection accuracy is measured by the percentage of correct features selected: $FSA = |\hat{F}\cap F|/|F|$, where $\hat{F}$ is the estimated feature candidate set.

We generate an $n \times m$ random matrix $X$. Each element of $X$ follows an independent standard Gaussian distribution $N(0,1)$. We then normalize the length of the columns of $X$ to be 1.

The $s-$sparse original signal $\beta^*$ is generated with $s$ nonzero elements independently uniformly distributed from $[-10,10]$. The locations of $s$ nonzero elements are uniformly distributed in $\{1,2,\cdots,m\}$. We form the observation by $y = X\beta^* + \varepsilon$, where the noise vector $\varepsilon$ is generated by the Gaussian distribution $N(0,\sigma^2 I)$. All experiments are repeated 100 times and we use their average performance for comparison.

First we compare the standard Dantzig selector and the multi-stage version. For a fair comparison, we choose the same $\lambda = \sigma\sqrt{2\log m}$ in both algorithms. We run the proposed algorithm with $F_0^{(0)} = \varnothing$ with different values of $N$ and let the estimation $\hat{\beta}$ be the output $\hat{\beta}^{(N)}$ in **Algorithm 1**. The feature candidate set $\hat{F}$ is predicted by the index set of the $s$ largest elements in $\hat{\beta}$. Note that $\hat{F}$ identified by $\hat{\beta} = \hat{\beta}^{(N)}$ is different from the output $F_0^{(N)}$ by **Algorithm 1**. The size of $\hat{F}$ is always $s$ while the size of $F_0^{(N)}$ is $N$. Note that the solution of the standard Dantzig selector algorithm is equivalent to $\hat{\beta}^{(N)}$ with $N = 0$. We report the *SRA* curve of $\hat{\beta}^{(N)}$ with respect to $N$ in the left column of Figure 1. The right column of Figure 1 shows the *FSA* curve with respect to $N$. We allow $N > s$ in our simulation although this case is beyond our theoretical analysis, since in practice the sparsity number $s$ is usually unknown in advance. We can observe from Figure 1 that 1) the multi-stage method obtains a solution with a smaller distance to the original signal than the standard Dantzig selector method; 2) the multi-stage method selects a larger percentage of correct features than the standard Dantzig selector method; 3) the multi-stage method can achieve the oracle solution with a large probability; and 4) even when $N > s$, the multi-stage algorithm still outperforms the standard Dantzig selector and achieves high accuracy in signal recovery and feature selection. Overall, the recovery accuracy curve increases with an increasing value of $N$ before reaching the sparsity level $s$ and decreases slowly after that, and the feature selection accuracy curve increases while $N \leq s$ and becomes flat after $N$ goes beyond $s$.

Next we apply the multi-stage procedure to the LASSO case and compare the multi-stage LASSO to the standard LASSO and the two-stage LASSO (Zhang, 2009a). The two-stage LASSO algorithm first estimates a support set $F_0 = supp_\alpha(\beta')$ from the solution $\beta'$ of the standard LASSO where $\alpha > 0$ is the threshold parameter; the second stage estimates the signal by solving the following problem

$$\min_{\beta}: \ \frac{1}{2}\|X\beta - y\|_2^2 + \lambda'\|\beta_{\bar{F}_0}\|_1, \tag{8}$$

which is indeed identical to Equation (7). In order to make it comparable to the proposed multi-stage LASSO algorithm with the parameter $N$, we properly choose $\alpha$ such that $|F_0| = N$ and use the output $\hat{\beta}'$ from Equation (8) and the feature candidate set by $\hat{\beta}'$ for comparison. Similarly, we use the same $\lambda' = 2\lambda$ in the three algorithms. The comparison reported in Figure 2 also indicates the advantage of the proposed multi-stage procedure.

## 5. Conclusion

In this paper, we propose a multi-stage procedure to improve the performance of the Dantzig selector and the LASSO by iteratively selecting the supporting features and recovering the original signal. The proposed method makes use of the information of supporting features to estimate the signal and simultaneously makes use of the information of the estimated signal to select the supporting features. Our theoretical analysis shows that the proposed method improves upon the standard

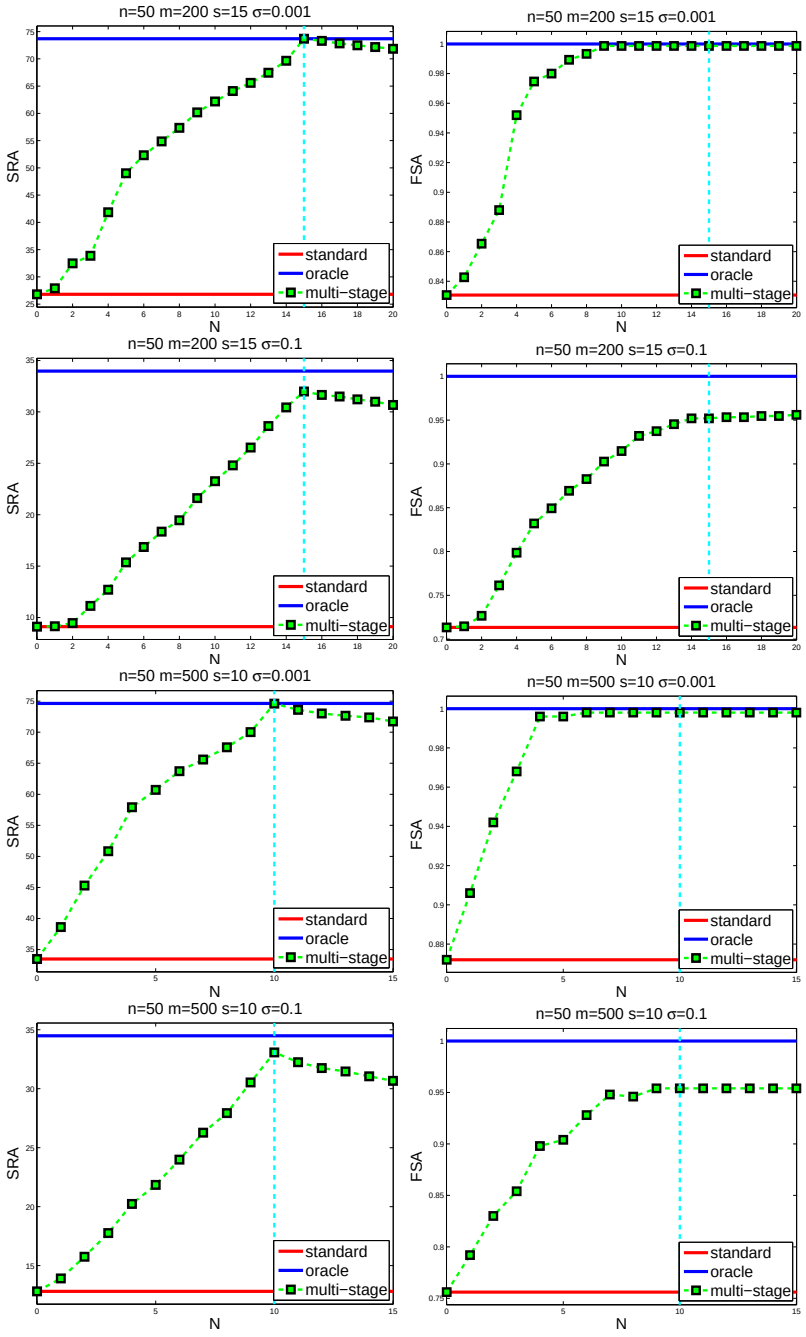


Figure 1: Numerical simulation. We compare the solutions of the standard Dantzig selector method ($N = 0$), the proposed method for different values of $N = 0, 1, \cdots, s, \cdots, s+5$, and the oracle solution. The *SRA* and *FSA* comparisons are reported on the left column and the right column, respectively. The red line indicates the *SRA* (or *FSA*) value of the standard Dantzig selector method; the blue line indicates the value of the oracle solution; the green curve with black boxes records the results by the proposed method for different values of $N$; the vertical cyan line distinguishes two cases $N <= s$ and $N > s$.

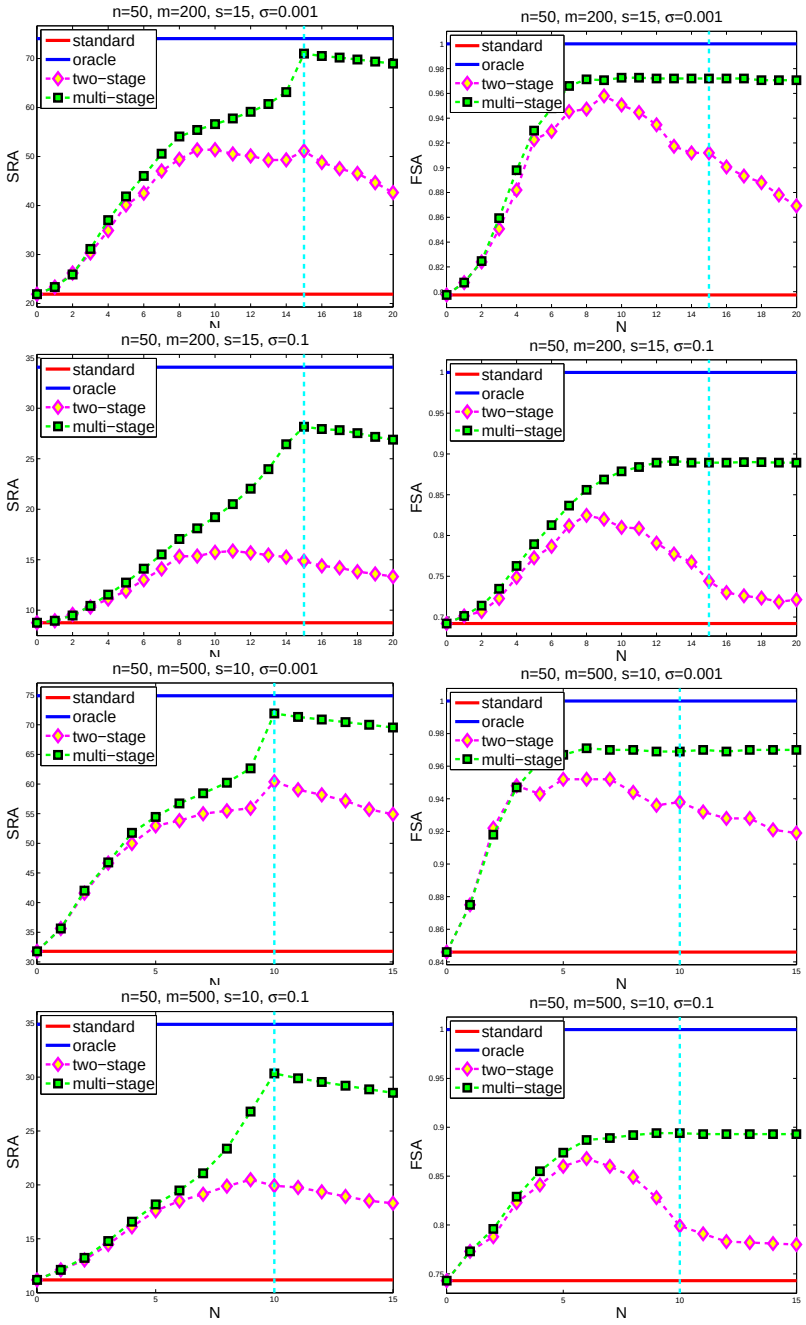


Figure 2: Numerical simulation. We compare the solutions of the standard Dantzig selector method ($N = 0$), the two-stage LASSO algorithm, the proposed method for different values of $N = 0, 1, \cdots, s, \cdots, s+5$, and the oracle solution. The *SRA* and *FSA* comparisons are reported on the left column and the right column, respectively. The red line indicates the *SRA* (or *FSA*) value of the standard Dantzig selector method; the blue line indicates the value of the oracle solution; the green curve with black boxes records the results of the proposed method for different values of $N$; the magenta curve with yellow diamonds indicates the results of the two-stage LASSO algorithm; the vertical cyan line distinguishes two cases $N <= s$ and $N > s$.

Dantzig selector and the LASSO in both signal recovery and supporting feature selection. The final numerical simulation confirms our theoretical analysis.

## Acknowledgments

This work is supported by NSF CCF-0811790, IIS-0953662, and CCF-1025177. We appreciate the constructive comments from the editor and three reviewers.

## Appendix A.

Theorem 1 is fundamental for the rest of the theorems. We first highlight a brief architecture for its proof. Theorem 1 estimates $\|\hat{\beta}-\beta^*\|_p$, which is bounded by the sum of two parts: $\|\hat{\beta}-\beta^*\|_p \le \|\hat{\beta}-\bar{\beta}\|_p + \|\bar{\beta}-\beta^*\|_p$. We use the upper bounds of these two parts to estimate the bound of $\|\hat{\beta}-\beta^*\|_p$. The analysis in Section 3.2 shows that the first term $\|\hat{\beta}-\bar{\beta}\|_p$ may be much larger than the second term $\|\bar{\beta}-\beta^*\|_p$. In Lemma 7, we estimate the bound of $\|\beta-\beta^*\|_p$ and its holding probability. The remaining part of the proof focuses on the estimation of the bound of $\|\hat{\beta}-\bar{\beta}\|_p$. For convenience, we use $h$ to denote $\hat{\beta}-\bar{\beta}$. $h$ can be divided into $h_{\bar{F}_1-T_1}$ and $h_{F_1+T_1}$, where $F_0 \subset F_1 \subset F$. Lemma 9 studies the relationship between $h_{\bar{F}_1-T_1}$ and $h_{F_1+T_1}$, if $\beta$ is feasible (Lemma 8 computes its holding probability). Then, Lemma 11 shows that $\|h\|_p$ can be bounded in terms of $\|h_{F_1+T_1}\|_p$. In Theorem 12, we estimate the bound of $\|h_{F_1+T_1}\|_p$. Finally, letting $F_1 = F$, we prove Theorem 1.

**Lemma 7** *With probability larger than* $1-\eta(\pi\log(s/\eta))^{-1/2}$, *the following holds:*

$$\|\bar{\beta}-\beta^*\|_p \le \frac{s^{1/p}\sigma\sqrt{2\log(s/\eta)}}{\mu^{(p)}_{(X_F^TX_F)^{1/2},s}}. \tag{9}$$

**Proof** According to the definition of $\bar{\beta}$, we have

$$\begin{aligned}\bar{\beta}_F &= (X_F^TX_F)^{-1}X_F^Ty = (X_F^TX_F)^{-1}X_F^T(X\beta^*+\varepsilon) = (X_F^TX_F)^{-1}X_F^T(X_F\beta_F^*+\varepsilon)\\ &= \beta_F^* + (X_F^TX_F)^{-1}X_F^T\varepsilon.\end{aligned}$$

It follows that

$$\bar{\beta}_F - \beta_F^* = (X_F^TX_F)^{-1}X_F^T\varepsilon \sim N(0,(X_F^TX_F)^{-1}\sigma^2).$$

Since $\|\bar{\beta}-\beta^*\|_p = \|\bar{\beta}_F-\beta^*_F\|_p$, we only need to consider the bound for $\|\bar{\beta}_F-\beta^*_F\|_p$. Let $Z = (X_F^T X_F)^{1/2}(\beta^*_F-\beta_F)/\sigma \sim N(0,I)$. We have

$$\begin{aligned}
P(\|Z\|_p \geq t) &= (2\pi)^{-s/2}\int_{\|Z\|_p\geq t} e^{-Z^TZ/2}dZ \\
&\leq (2\pi)^{-s/2}\int_{s^{1/p}\|Z\|_\infty\geq t} e^{-Z^TZ/2}dZ \quad (\text{due to } \|Z\|_p \leq s^{1/p}\|Z\|_\infty) \\
&= 1-(2\pi)^{-s/2}\int_{\|Z\|_\infty\leq s^{-1/p}t} e^{-Z^TZ/2}dZ \\
&= 1-\left[(2\pi)^{-1/2}\int_{|Z_i|\leq s^{-1/p}t} e^{-Z_i^2/2}dZ_i\right]^s \\
&= 1-\left[1-2(2\pi)^{-1/2}\int_{s^{-1/p}t}^{\infty} e^{-Z_i^2/2}dZ_i\right]^s \\
&\leq s\left[2(2\pi)^{-1/2}\int_{s^{-1/p}t}^{\infty} e^{-Z_i^2/2}dZ_i\right] \\
&\leq \frac{2s^{1+1/p}}{t(2\pi)^{1/2}}\exp\left[\frac{-t^2}{2s^{2/p}}\right].
\end{aligned}$$

Thus the following bound holds with probability larger than $1-\frac{2s^{1+1/p}}{t(2\pi)^{1/2}}\exp\left[\frac{-t^2}{2s^{2/p}}\right]$:

$$\begin{aligned}
P(\|Z\|_p \leq t) &= P(\|(X_F^TX_F)^{1/2}(\beta^*_F-\bar{\beta}_F)\|_p \leq t\sigma) \\
&\leq P(\mu^{(p)}_{(X_F^TX_F)^{1/2},s}\|\beta^*_F-\bar{\beta}_F\|_p \leq t\sigma) = P(\|\beta^*_F-\bar{\beta}_F\|_p \leq t\sigma/\mu^{(p)}_{(X_F^TX_F)^{1/2},s}).
\end{aligned}$$

Taking $t=\sqrt{2\log(s/\eta)}s^{1/p}$, we prove the claim. Note that the presented bound holds for any $p\geq 1$. ■

**Lemma 8** *With probability larger than* $1-\eta(\pi\log\frac{m-s}{\eta})^{-1/2}$, *the following bound holds:*

$$\|X_F^T(X\bar{\beta}-y)\|_\infty \leq \lambda,$$

*where* $\lambda=\sigma\sqrt{2\log(m-s)/\eta}$.

**Proof** Let us first consider the probability of $\|X_F^T(X\bar{\beta}-y)\|_\infty\leq\lambda$. For any $j\in\bar{F}$, define $v_j$ as

$$\begin{aligned}
v_j &= X_j^T(X\bar{\beta}-y) \\
&= X_j^T\left(X_F(X_F^TX_F)^{-1}X_F^T(X_F\beta^*_F+\varepsilon)-X_F\beta^*_F-\varepsilon\right) \\
&= X_j^T\left(X_F(X_F^TX_F)^{-1}X_F^T-I\right)\varepsilon \\
&\sim N(0,X_j^T(I-X_F(X_F^TX_F)^{-1}X_F^T)X_j\sigma^2).
\end{aligned}$$

Since $(I-X_F(X_F^TX_F)^{-1}X_F^T)$ is a projection matrix, we have $X_j^T(I-X_F(X_F^TX_F)^{-1}X_F^T)X_j\sigma^2\leq\sigma^2$. Thus,

$$P(\|X_F^T(X\bar{\beta}-y)\|_\infty\geq\lambda) = P(\sup_{j\in\bar{F}}|v_j|\geq\lambda) \leq \frac{2(m-s)\sigma}{\lambda(2\pi)^{1/2}}\exp\{-\lambda^2/2\sigma^2\}.$$

Taking $\lambda = \sigma\sqrt{2\log(m-s)/\eta}$ in the inequality above, we prove the claim. ■

It follows from the definition of $\bar{\beta}$ that $\|X_F^T(X\bar{\beta}-y)\|_\infty = 0$ always holds. In the following discussion, we assume that the following assumption holds:

**Assumption 2** *$\bar{\beta}$ is a feasible solution of the problem* (2)*, if $F_0 \subset F$.*

Under the assumption above, both $\|X_F^T(X\bar{\beta}-y)\|_\infty \le \lambda$ and $\|X_F^T(X\bar{\beta}-y)\|_\infty = 0$ hold.

Note that this assumption is just used to simplify the description for following proofs. Our proof for the final theorems will substitute this assumption by the probability it holds.

In the following, we introduce an additional set $F_1$ satisfying $F_0 \subset F_1$ (Zhang, 2009a).

**Lemma 9** *Let $F_0 \subset F$. Assume that Assumption 2 holds. Given any index set $F_1$ such that $F_0 \subset F_1$, we have the following conclusions:*

$$\begin{aligned}\|h_{\bar{F}_0-\bar{F}_1}\|_1 + 2\|\bar{\beta}_{\bar{F}_1}\|_1 \ge& \|h_{\bar{F}_1}\|_1\\ \|X_{F_0}^T X h\|_\infty =& 0\\ \|X_F^T X h\|_\infty \le& 2\lambda\\ \|X_{\bar{F}_0-\bar{F}}^T X h\|_\infty \le& \lambda.\end{aligned}$$

**Proof** Since $\bar{\beta}$ is a feasible solution, the following holds

$$\begin{aligned}\|\hat{\beta}_{\bar{F}_0}\|_1 &\le \|\bar{\beta}_{\bar{F}_0}\|_1\\ \|\hat{\beta}_{\bar{F}_0-\bar{F}_1}\|_1 + \|\hat{\beta}_{\bar{F}_1}\|_1 &\le \|\bar{\beta}_{\bar{F}_0-\bar{F}_1}\|_1 + \|\bar{\beta}_{\bar{F}_1}\|_1\\ \|\hat{\beta}_{\bar{F}_1}\|_1 &\le \|h_{\bar{F}_0-\bar{F}_1}\|_1 + \|\bar{\beta}_{\bar{F}_1}\|_1\\ \|h_{\bar{F}_1} + \bar{\beta}_{\bar{F}_1}\|_1 &\le \|h_{\bar{F}_0-\bar{F}_1}\|_1 + \|\bar{\beta}_{\bar{F}_1}\|_1\\ \|h_{\bar{F}_1}\|_1 &\le \|h_{\bar{F}_0-\bar{F}_1}\|_1 + 2\|\bar{\beta}_{\bar{F}_1}\|_1.\end{aligned}$$

Thus, the first inequality holds. Since

$$X_{F_0}^T X h = X_{F_0}^T X(\hat{\beta}-\bar{\beta}) = X_{F_0}^T(X\hat{\beta}-y) - X_{F_0}^T(X\bar{\beta}-y),$$

the second inequality can be obtained as follows:

$$\|X_{F_0}^T X h\|_\infty \le \|X_{F_0}^T(X\hat{\beta}-y)\|_\infty + \|X_{F_0}^T(X\bar{\beta}-y)\|_\infty = 0.$$

The third inequality holds since

$$\|X_F^T X h\|_\infty \le \|X_F^T(X\hat{\beta}-y)\|_\infty + \|X_F^T(X\bar{\beta}-y)\|_\infty \le 2\lambda.$$

Similarly, the fourth inequality can be obtained as follows:

$$\|X_{\bar{F}_0-\bar{F}}^T X h\|_\infty \le \|X_{\bar{F}_0-\bar{F}}^T(X\hat{\beta}-y)\|_\infty + \|X_{\bar{F}_0-\bar{F}}^T(X\bar{\beta}-y)\|_\infty \le \lambda.$$

■

**Lemma 10** *Given any $v \in \mathbb{R}^m$, its index set $T$ is divided into a group of subsets $T_j$'s ($j = 1,2,...$) without intersection such that $\bigcup_j T_j = T$. If $\max_j |T_j| \le l$ and $\max_{i \in T_{j+1}} |v_{T_{j+1}}[i]| \le \|v_{T_j}\|_1 / l$ hold for all $j$'s, then we have*

$$\|v_{\bar{T}_1}\|_p \le \|v\|_1 l^{1/p-1}.$$

**Proof** Since $|v_{T_{j+1}}[i]| \le \|v_{T_j}\|_1 / l$, we have

$$\begin{aligned} \|v_{T_{j+1}}\|_p^p = \sum_{i \in T_{j+1}} |v_{T_{j+1}}^p[i]| \le \|v_{T_j}\|_1^p l^{1-p}, \\ \Rightarrow \|v_{T_{j+1}}\|_p \le \|v_{T_j}\|_1 l^{1/p-1}. \end{aligned}$$

Thus,

$$\|v_{\bar{T}_1}\|_p \le \sum_{j \ge 1} \|v_{T_{j+1}}\|_p \le \sum_{j \ge 1} \|v_{T_j}\|_1 l^{1/p-1} = \|v\|_1 l^{1/p-1},$$

which proves the claim. ■

Note that similar techniques as those in Lemma 10 have been used in the literature (Candès and Tao, 2007; Zhang, 2009a).

**Lemma 11** *Assume that $F_0 \subset F$ and $F_0 \subset F_1$. We divide the index set $\bar{F}_1$ into a group of subsets $T_j$'s ($j = 1,2,...$) such that they satisfy all conditions in Lemma 10 with $v = h$. Then the following holds:*

$$\begin{aligned} \|h_{\bar{F}_1 - T_1}\|_p \le & l^{1/p-1} \left( |\bar{F}_0 - \bar{F}_1|^{1-1/p} \|h_{\bar{F}_0 - \bar{F}_1}\|_p + 2\|\bar{\beta}_{\bar{F}_1}\|_1 \right), \\ \|h\|_p \le & \left[ 1 + \left( \frac{|\bar{F}_0 - \bar{F}_1|}{l} \right)^{p-1} \right]^{1/p} \|h_{F_1 + T_1}\|_p + 2l^{1/p-1} \|\bar{\beta}_{\bar{F}_1}\|_1. \end{aligned}$$

**Proof** Using Lemma 10 with $T = \bar{F}_1$, the first inequality can be obtained using the first inequality in lemma 9 as follows:

$$\begin{aligned} \|h_{\bar{F}_1 - T_1}\|_p \le & l^{1/p-1} \|h_{\bar{F}_1}\|_1 \le l^{1/p-1} \left( \|h_{\bar{F}_0 - \bar{F}_1}\|_1 + 2\|\bar{\beta}_{\bar{F}_1}\|_1 \right) \\ \le & l^{1/p-1} \left( |\bar{F}_0 - \bar{F}_1|^{1-1/p} \|h_{\bar{F}_0 - \bar{F}_1}\|_p + 2\|\bar{\beta}_{\bar{F}_1}\|_1 \right). \end{aligned}$$

For any $x \ge 0$, $y \ge 0$, $p \ge 1$, and $a \ge 0$, it can be easily verified that

$$(x^p + (ax+y)^p)^{1/p} \le (1 + a^p)^{1/p} x + y. \tag{10}$$

It follows that

$$\begin{aligned} \|h\|_p = & \left[ \|h_{F_1 + T_1}\|_p^p + \|h_{\bar{F}_1 - T_1}\|_p^p \right]^{1/p} \\ \le & \left[ \|h_{F_1 + T_1}\|_p^p + \left[ \left( \frac{|\bar{F}_0 - \bar{F}_1|}{l} \right)^{1-1/p} \|h_{\bar{F}_0 - \bar{F}_1}\|_p + 2l^{1/p-1} \|\bar{\beta}_{\bar{F}_1}\|_1 \right]^p \right]^{1/p} \\ \le & \left[ 1 + \left( \frac{|\bar{F}_0 - \bar{F}_1|}{l} \right)^{p-1} \right]^{1/p} \|h_{F_1 + T_1}\|_p + 2l^{1/p-1} \|\bar{\beta}_{\bar{F}_1}\|_1. \end{aligned}$$

The first inequality is due to the first claim in this lemma; the second inequality is due to $\|h_{\bar{F}_0-\bar{F}_1}\|_p \leq \|h_{F_1+T_1}\|_p$ and (10). We complete the proof for the second claim. ■

**Theorem 12** *Under Assumption 1, taking $F_0 \subset F$ and $\lambda = \sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into the optimization problem* (2)*, for any given index set $F_1$ satisfying $F_0 \subset F_1 \subset F$ , if there exists some $l$ such that $\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p} > 0$ holds where $s_1 = |F_1|$, then with probability larger than $1-\eta_1'$, the $\ell_p$ norm ($1 \leq p \leq \infty$) of the difference between the optimizer of the problem* (2) *and the oracle solution is bounded as*

$$\|\hat{\beta} - \bar{\beta}\|_p \leq \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l}l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} + 2l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1$$

*and with probability larger than $1-\eta_1'-\eta_2'$, the $\ell_p$ norm ($1 \leq p \leq \infty$) of the difference between the optimizer of the problem* (2) *and the true solution is bounded as*

$$\|\hat{\beta} - \beta^*\|_p \leq \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l}l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} + 2l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.$$

**Proof** First, we assume Assumption 2 and the inequality (9) hold. Divide $\bar{F}_1$ into a group of subsets $T_j$'s ($j = 1,2,...$) without intersection such that $\bigcup_j T_j = \bar{F}_1$, $\max_j |T_j| \leq l$ and $\max_{i\in T_{j+1}} h_{T_{j+1}}[i] \leq \|h_{T_j}\|_1/l$ hold. Note that such a partition always exists. Simply, let $T_1$ be the index set of the largest $l$ elements in $h$, $T_2$ be the index set of the largest $l$ elements among the remaining elements, and so on (the size of the last set may be less than $l$). It is easy to verify that this group of sets satisfy all

conditions above. For convenience of presentation, we denote $T_0 = \bar{F}_0 - \bar{F}_1$ and $T_{01} = T_0 + T_1$. Since

$$\begin{aligned}
&\|X^T_{T_{01}+F_0} Xh\|_p \\
=&\|X^T_{T_{01}+F_0} X_{T_{01}+F_0} h_{T_{01}+F_0} + \sum_{j\geq 2} X^T_{T_{01}+F_0} X_{T_j} h_{T_j}\|_p \\
\geq&\mu^{(p)}_{A,s_1+l}\|h_{T_{01}+F_0}\|_p - \sum_{j\geq 2}\theta^{(p)}_{A,s_1+l,l}\|h_{T_j}\|_p \\
\geq&\mu^{(p)}_{A,s_1+l}\|h_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l}\sum_{j\geq 2}\|h_{T_j}\|_p \\
\geq&\mu^{(p)}_{A,s_1+l}\|h_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|h_{\bar{F}_1}\|_1 \quad \text{(due to lemma 10)} \\
\geq&\mu^{(p)}_{A,s_1+l}\|h_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\left(\|h_{T_0}\|_1 + 2\|\bar{\beta}_{\bar{F}_1}\|_1\right) \quad \text{(due to lemma 9)} \\
\geq&\mu^{(p)}_{A,s_1+l}\|h_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l}\left(\frac{l}{|T_0|}\right)^{1/p-1}\|h_{T_0}\|_p - 2\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 \\
\geq&\left(\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{l}{|T_0|}\right)^{1/p-1}\right)\|h_{T_{01}+F_0}\|_p - 2\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1
\end{aligned}$$

and

$$\begin{aligned}
&\|X^T_{T_{01}+F_0} Xh\|^p_p \\
=&\|X^T_{F_0} Xh\|^p_p + \|X^T_{T_{01}\cap F} Xh\|^p_p + \|X^T_{T_{01}\cap \bar{F}} Xh\|^p_p \\
\leq&|T_{01}\cap F|\lambda^p + |T_{01}\cap \bar{F}|(2\lambda)^p \quad \text{(due to lemma 9)} \\
\leq&|T_0\cap F|\lambda^p + |T_1\cap F|\lambda^p + |T_0\cap \bar{F}|(2\lambda)^p + |T_1\cap \bar{F}|(2\lambda)^p \quad \text{(due to } F_1\subset F\text{)} \\
\leq&|T_0|\lambda^p + l(2\lambda)^p, \quad \text{(due to } T_0\cap \bar{F} = \varnothing\text{)}
\end{aligned}$$

we have

$$\begin{aligned}
\|h_{F_1+T_1}\|_p = \|h_{T_{01}+F_0}\|_p &\leq \frac{(|T_0|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1}{\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{l}{|T_0|}\right)^{1/p-1}} \\
&= \frac{(|\bar{F}_0-\bar{F}_1|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1}{\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}}.
\end{aligned}$$

Due to the second inequality in Lemma 11, we have

$$\begin{aligned}
\|h\|_p &\leq \left[1+\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\|h_{F_1+T_1}\|_p + 2l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 \\
&= \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - \theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} + \\
&\quad 2l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1.
\end{aligned}$$

Thus, we can bound $\|\hat{\beta}-\beta^*\|_p$ as

$$\begin{aligned}\|\hat{\beta}-\beta^*\|_p \leq & \|\hat{\beta}-\bar{\beta}\|_p + \|\bar{\beta}-\beta^*\|_p \\ \leq & \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+2^p l)^{1/p}\lambda + 2\theta^{(p)}_{A,s_1+l,l}l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l}-\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} \\ & + 2l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^TX_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.\end{aligned}$$

Finally, taking $\lambda=\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$, Lemma 8 with $\eta=\eta_1$ implies that Assumption 2 holds with probability larger than $1-\eta_1'$ and Lemma 7 with $\eta=\eta_2$ implies that (9) holds with probability larger than $1-\eta_2'$. Thus, these two bounds above hold with probabilities larger than $1-\eta_1'$ and $1-\eta_1'-\eta_2'$, respectively. ■

**Remark 13** *Candès and Tao (2007) provided a more general upper bound for the Dantzig selector solution in the order of* $O\left(k^{1/2}\sigma\sqrt{\log m}+r_k^{(2)}(\beta^*)\sqrt{\log m}\right)$, *where* $1\leq k\leq s$ *and* $r_k^{(p)}(\beta)=\left(\sum_{i\in L_k}|\beta_i|^p\right)^{1/p}$ *($L_k$ is the index set of the $k$ largest entries in $\beta$). We argue that the result in Theorem 12 potentially implies a tighter bound for Dantzig selector. Setting $F_0=\varnothing$ (equivalent to the standard Dantzig selector) and $l=k$ with $k=|\bar{F}_1|$ in Theorem 12, it is easy to verify that the order of the bound for $\|\hat{\beta}_D-\bar{\beta}\|_p$ is determined by* $O\left(k^{1/p}\sigma\sqrt{\log m}+k^{1/p-1}r_k^{(1)}(\bar{\beta})\right)$, *or* $O\left(k^{1/p}\sigma\sqrt{\log m}+k^{1/p-1}r_k^{(1)}(\beta^*)\right)$ *due to Lemma 7. This bound achieves the same order as the bound of the LASSO solution given by Zhang (2009a), which is the sharpest bound for LASSO to our knowledge.*

We are now ready to prove Theorem 1.

**Proof of Theorem 1:** Taking $F_1=F$ in theorem 12 which indicates that $\bar{\beta}_{\bar{F}_1}=0$, we conclude that

$$\|\hat{\beta}-\bar{\beta}\|_p \leq \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}(|\bar{F}_0-\bar{F}|+l2^p)^{1/p}}{\mu^{(p)}_{A,s+l}-\theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda$$

holds with probability larger than $1-\eta_1'$ and

$$\begin{aligned}\|\hat{\beta}-\beta^*\|_p \leq & \|\hat{\beta}-\bar{\beta}\|_p + \|\bar{\beta}-\beta^*\|_p \\ \leq & \frac{\left[1+\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}(|\bar{F}_0-\bar{F}|+l2^p)^{1/p}}{\mu^{(p)}_{A,s+l}-\theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^TX_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}\end{aligned}$$

holds with probability larger than $1-\eta_1'-\eta_2'$. ■

**Proof of Theorem 3:** From the proof in Theorem 12, the bounds (3) and (4) in Theorem 1 hold with probability 1 if Assumption 2 and the inequality (9) hold. It is easy to verify by Theorem 1 that for any $j \in J$, the following holds: $|\beta_j^*| > \alpha_0 \geq \|\hat{\beta}-\beta\|_\infty + \|\hat{\beta}-\beta^*\|_\infty$. For any $j \in J$, we have

$$|\hat{\beta}_j| \geq |\beta_j^*| - |\hat{\beta}_j - \beta_j^*| > \|\hat{\beta}-\bar{\beta}\|_\infty + \|\hat{\beta}-\beta^*\|_\infty - |\hat{\beta}_j-\beta_j^*| \geq \|\hat{\beta}-\bar{\beta}\|_\infty \geq \|\hat{\beta}_{\bar{F}}\|_\infty.$$

Thus, there exist at least $|J|$ elements of $\hat{\beta}_F$ larger than $\|\hat{\beta}_{\bar{F}}\|_\infty$. If we pick up the largest $|J|$ elements in $\hat{\beta}$, then all of them correspond to the location of nonzero entries in the true solution $\beta^*$. Since Assumption 2 and the inequality (9) hold, the bounds (3) and (4) in Theorem 1 hold with probability larger than $1-\eta_1'-\eta_2'$. Thus the claim above holds with probability larger than $1-\eta_1'-\eta_2'$. Note that the probability will not accumulate, as we only need the holding probability of Assumption 2 and the inequality (9). The proofs below follow the same principle. ■

**Proof of Theorem 4:** From the proof in Theorem 12, the bounds (3) and (4) in Theorem 1 hold with probability 1 if assumption 2 and the inequality (9) hold. In the multi-stage algorithm, the problem in (2) is solved $N$ times. It is easy to verify that the following holds:

$$\alpha_0 \geq \|\hat{\beta}^{(0)}-\bar{\beta}\|_\infty + \|\hat{\beta}^{(0)}-\beta^*\|_\infty.$$

Since $|supp_{\alpha_0}(\beta_J^*)| > 0$, there exists at least 1 element in $\hat{\beta}_J^{(0)}$ larger than $\|\hat{\beta}_{\bar{F}}^{(0)}\|_\infty$. Thus, $F_0^{(1)}$ must be a subset of $F$. Then, we can verify that

$$\alpha_1 \geq \|\hat{\beta}^{(1)}-\bar{\beta}\|_\infty + \|\hat{\beta}^{(1)}-\beta^*\|_\infty,$$

and $|supp_{\alpha_1}(\beta_J^*)| > 1$ guarantee that there exist at least 2 elements in $\hat{\beta}_J^{(1)}$ larger than $\|\hat{\beta}_{\bar{F}}^{(1)}\|_\infty$. Thus, $F_0^{(2)}$ must be a subset of $F$. Similarly, we can show that $F_0^{(N)}$ is guaranteed to be a subset of $F$. Since the bounds (3) and (4) in Theorem 1 hold with probability larger than $1-\eta_1'-\eta_2'$, the claim $F_0^{(N)} \subset F$ holds with probability larger than $1-\eta_1'-\eta_2'$. ■

**Proof of Theorem 5:** From Theorem 1, the first conclusion holds with probability larger than $1-\eta_1'-\eta_2'$ by choosing $F_0 = \varnothing$ and $l = s$.

Assuming Assumption 2 and the inequality (9) hold, the bounds (3) and (4) in Theorem 1 hold with probability 1. Since the conditions in Theorem 4 are satisfied, the $|J|$ correct features can be selected from the feature set, that is, $F_0^{(|J|)} \subset F$. Using the conclusion in (4) of Theorem 1, the bound of the multi-stage method can be estimated by taking $l = |\bar{F_0} - \bar{F}|$ as follows:

$$\|\hat{\beta}_{mul}-\beta^*\|_p \leq \frac{(2^{p+1}+2)^{1/p}(s-N)^{1/p}}{\mu_{A,2s-N}^{(p)} - \theta_{A,2s-N,s-N}^{(p)}}\lambda + \frac{s^{1/p}}{\mu_{(X_F^T X_F)^{1/2},s}^{(p)}}\sigma\sqrt{2\log(s/\eta_2)}.$$

Note that since

$$\mu_{A,2s-N}^{(p)} - \theta_{A,2s-N,s-N}^{(p)} \geq \mu_{A,2s}^{(p)} - \theta_{A,2s,s}^{(p)},$$

the following always holds: $\mu_{A,2s-N}^{(p)} - \theta_{A,2s-N,s-N}^{(p)} > 0$. Since Assumption 2 and the inequality (9) hold, the bounds (3) and (4) in Theorem 1 hold with probability larger than $1-\eta_1'-\eta_2'$. Thus the

claim above holds with probability larger than $1-\eta_1'-\eta_2'$. ■

**Proof of Theorem 6:** First, we assume that Assumption 2 and the inequality (9) hold. In this case, the claim in Theorem 4 holds with probability 1. Since all conditions in Theorem 4 are satisfied, after $s$ iterations, $s$ correct features will be selected (i.e., $F_0^{(N)} = F$) with probability 1. Since all correct features are obtained, the optimization problem in the last iteration can be formulated as:

$$\begin{aligned} &\min : \|\beta_{\bar{F}}\|_1 \\ &s.t. : \|X_F^T(X\beta - y)\|_\infty \le \lambda \\ &\quad \|X_F^T(X\beta - y)\|_\infty = 0. \end{aligned} \tag{11}$$

The oracle solution minimizes the objective function to 0. Since Assumption 2 indeed implies that the oracle is a feasible solution, the oracle solution is one optimizer. We can also show that it is the unique optimizer. If there is another optimizer $\beta \neq \bar{\beta}$, then $\beta_{\bar{F}} = 0$ and $\beta_F = (X_F^T X_F)^{-1} X_F^T y$, which is identical to the definition of the oracle solution. Thus, we conclude that the oracle is the unique optimizer for the optimization problem (11) with probability 1. Since the holding probability of Assumption 2 and the inequality (9) is larger than $1-\eta_1'-\eta_2'$, the oracle solution can be achieved with the same probability. ■

## Appendix B.

In this section, we expound the properties of the multi-stage LASSO which are very similar to the multi-stage Dantzig selector. The complete proof is given below.

In the following discussion, we use $\hat{\beta}'$ to denote the solution in Equation (7) and let $h' = \hat{\beta}' - \bar{\beta}$. We first consider the simple case $F_0 \subset F$ as in Section 3.1; we have the following theorem.

**Theorem 14** *Assume Assumption 1 holds. Take $F_0 \subset F$ and*

$$\lambda' = 2\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$$

*into the optimization problem* (7)*. If there exists some $l$ such that*

$$\mu_{A,s+l}^{(p)} - 3\theta_{A,s+l,l}^{(p)}\left(\frac{|\bar{F}_0 - \bar{F}|}{l}\right)^{1-1/p} > 0$$

*holds, then with probability larger than $1-\eta_1'$, the $\ell_p$ norm ($1 \le p \le \infty$) of the difference between $\hat{\beta}'$, the optimizer of the problem* (7) *and the oracle solution $\bar{\beta}$ is bounded as*

$$\|\hat{\beta}' - \bar{\beta}\|_p \le \frac{\left[1+3\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}\left(|\bar{F}_0 - \bar{F}| + (3/2)^p l\right)^{1/p}}{\mu_{A,s+l}^{(p)} - 3\theta_{A,s+l,l}^{(p)}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda'$$

*and with probability larger than* $1-\eta_1'-\eta_2'$, *the* $\ell_p$ *norm* ($1 \le p \le \infty$) *of the difference between* $\hat{\beta}'$, *the optimizer of the problem* (7) *and the true solution* $\beta^*$ *is bounded as*

$$\|\hat{\beta}'-\beta^*\|_p \le \frac{\left[1+3\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{p-1}\right]^{1/p}(|\bar{F}_0-\bar{F}|+(3/2)^p l)^{1/p}}{\mu^{(p)}_{A,s+l}-3\theta^{(p)}_{A,s+l,l}\left(\frac{|\bar{F}_0-\bar{F}|}{l}\right)^{1-1/p}}\lambda' + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^TX_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.$$

This theorem is similar to Theorem 1 for the multi-stage Dantzig selector. Like Equations (3) and (4), the two bounds in the above theorem are strictly decreasing in terms of $|\bar{F}_0-\bar{F}|$. Thus, feature selection and signal recovery can benefit from each other. For this reason, the multi-stage LASSO has similar properties as the multi-stage Dantzig selector. We expound them as follows.
*(a)* First, like Theorem 4, in the LASSO case the multi-stage procedure can lead to a weaker requirement to choose $|J|$ correct features than the standard LASSO as shown in the following theorem.

**Theorem 15** *Under Assumption 1, if there exist a nonempty set*

$$\Omega = \{l|\mu^{(p)}_{A,s+l}-3\theta^{(p)}_{A,s+l,l}\left(\frac{s}{l}\right)^{1-1/p}>0\}$$

*and a set J such that* $|supp_{\alpha_i}(\beta^*_J)|>i$ *holds for all* $i\in\{0,1,...,|J|-1\}$, *where*

$$\alpha_i = \frac{3}{2}\min_{l\in\Omega}\frac{\max(1,\frac{3(s-i)}{l})}{\mu^{(\infty)}_{A,s+l}-3\theta^{(\infty)}_{A,s+l,l}\left(\frac{s-i}{l}\right)}\lambda' + \frac{1}{\mu^{(\infty)}_{(X_F^TX_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)},$$

*then taking* $F_0^{(0)}=\varnothing$, $\lambda=\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ *and* $N=|J|-1$ *into the multi-stage algorithm 1, the result after N iterations satisfies* $F_0^{(N)}\subset F$ *(i.e.,* $|J|$ *correct features are chosen) with probability larger than* $1-\eta_1'-\eta_2'$.

It is easy to see that $\alpha_0>\alpha_1>...>\alpha_{|J|-1}$ holds strictly. Referring to the analysis for Theorem 4, we know that the multi-stage method for LASSO requires weaker conditions to obtain $|J|$ correct features than the standard LASSO.
*(b)* Second, like Theorem 5 the following theorem shows that with a high probability the multi-stage procedure can improve the upper bound of the standard LASSO from $Cs^{1/p}\sqrt{\log m}+\Delta$ to $C(s-N)^{1/p}\sqrt{\log m}+\Delta$, where $C$ is a constant and $\Delta$ is a small number independent from $m$.

**Theorem 16** *Under Assumption 1, if there exist l such that* $\mu^{(\infty)}_{A,s+l}-3\theta^{(\infty)}_{A,s+l,l}\left(\frac{s}{l}\right)>0$, $\mu^{(p)}_{A,2s}-3\theta^{(p)}_{A,2s,s}>0$, *and a set J such that* $|supp_{\alpha_i}(\beta^*_J)|>i$ *holds for all* $i\in\{0,1,...,|J|-1\}$, *where* $\alpha_i$*'s follow the definition in Theorem 15, then taking* $F_0=\varnothing$, $N=|J|$ *and* $\lambda'=2\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ *into the multi-stage LASSO algorithm, the solution* $\hat{\beta}'_{mul}$ *of the multi-stage LASSO obeys*

$$\|\hat{\beta}'_{mul}-\beta^*\|_p \le \frac{4\left(\left(\frac{3}{2}\right)^p+1\right)^{1/p}(s-N)^{1/p}}{\mu^{(p)}_{A,2s-N}-3\theta^{(p)}_{A,2s-N,s-N}}\lambda' + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^TX_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}$$

*with probability larger than* $1-\eta_1'-\eta_2'$.

*(c)* Finally, the proposed method can obtain the oracle solution with high probability under certain conditions:

**Theorem 17** *Under Assumption 1, if there exists $l$ such that $\mu^{(\infty)}_{A,s+l} - 3\theta^{(\infty)}_{A,s+l,l}\left(\frac{s-i}{l}\right) > 0$, and the supporting set $F$ of $\beta^*$ satisfies $|supp_{\alpha_i}(\beta^*_F)| > i$ for all $i \in \{0,1,...,s-1\}$, where $\alpha_i$ follows the definition in theorem 15, then taking $F_0 = \varnothing$, $N = s$ and $\lambda' = 2\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ into the multi-stage LASSO algorithm, the oracle solution can be achieved, that is, $F_0^{(N)} = F$ and $\hat{\beta}'^{(N)} = \bar{\beta}$ with probability larger than $1-\eta_1'-\eta_2'$.*

In the following, we provide the complete proof for the theorems above.

**Lemma 18** *Let $\hat{\beta}'$ be defined above. We have*

$$\|X_{F_0}^T(X\hat{\beta}' - y)\|_\infty \leq \lambda'.$$

**Proof** The subdifferential of the objective function in Equation (7) at the optimal solution $\hat{\beta}'$ is given by:

$$X_i^T(X\hat{\beta}' - y) + \lambda' sgn(\hat{\beta}'_i)$$

where $i \in \bar{F_0}$ and

$$sgn(x) = \begin{cases} 1, & x > 0; \\ -1, & x < 0; \\ \text{[-1,1]}, & x = 0. \end{cases}$$

Since 0 must belong to the subdifferential at the optimal solution, we have

$$|X_i^T(X\hat{\beta}' - y)| \leq \lambda',$$

which implies the claim. ■

Let us assume that the oracle solution satisfies the following assumption.

**Assumption 3**

$$\|X_{\bar{F}}^T(X\bar{\beta} - y)\|_\infty \leq \lambda'/2.$$

This assumption actually plays the same role as Assumption 2 in the Dantzig selector.

In the following, we introduce an additional set $F_1$ satisfying $F_0 \subset F_1$ (Zhang, 2009a).

Similar to Lemma 9, we have the following results for the LASSO case:

**Lemma 19** *Let $F_0 \subset F$. Assume that Assumption 3 holds. Given any index set $F_1$ such that $F_0 \subset F_1$, we have the following conclusions:*

$$3\|h'_{\bar{F}_0 - \bar{F}_1}\|_1 + 4\|\bar{\beta}_{\bar{F}_1}\|_1 \geq \|h'_{\bar{F}_1}\|_1$$
$$\|X_{F_0}^T X h'\|_\infty = 0$$
$$\|X_{\bar{F}}^T X h'\|_\infty \leq \frac{3}{2}\lambda'$$
$$\|X_{\bar{F}_0 - \bar{F}}^T X h'\|_\infty \leq \lambda'.$$

**Proof** We only show the proof for the first inequality and the rest can be easily proven by following the proof in Lemma 9.

Let $\varepsilon = X\bar{\beta} - y$ and $f(.)$ be the objective function in Equation (7) with respect to $\beta$. One can verify that $\varepsilon^T X_F = 0$. Since $\hat{\beta}'$ is the optimal solution of Equation (7), we have

$$\begin{aligned}
0 \geq & f(\hat{\beta}') - f(\bar{\beta}) \\
= & \frac{1}{2}(\|X\hat{\beta}' - y\|_2^2 - \|X\bar{\beta} - y\|_2^2) + \lambda'(\|\hat{\beta}'_{\bar{F}_0}\|_1 - \|\bar{\beta}_{\bar{F}_0}\|_1) \\
= & \frac{1}{2}(Xh')^T(X\hat{\beta}' - y + \varepsilon) + \lambda'(\|\hat{\beta}'_{\bar{F}_0}\|_1 - \|\bar{\beta}_{\bar{F}_0}\|_1) \\
\geq & \varepsilon^T X h' + \lambda'(\|\hat{\beta}'_{\bar{F}_0}\|_1 - \|\bar{\beta}_{\bar{F}_0}\|_1) \\
\geq & \varepsilon^T (X_F h'_F + X_{\bar{F}} h'_{\bar{F}}) + \lambda'(\|\hat{\beta}'_{\bar{F}_0}\|_1 - \|\bar{\beta}_{\bar{F}_0}\|_1) \\
\geq & -\lambda'/2\|h'_{\bar{F}}\|_1 + \lambda'(\|\hat{\beta}'_{\bar{F}_0 - \bar{F}_1}\|_1 + \|\hat{\beta}'_{\bar{F}_1}\|_1 - \|\bar{\beta}_{\bar{F}_0 - \bar{F}_1}\|_1 - \|\bar{\beta}_{\bar{F}_1}\|_1) \quad \text{(due to Assumption 3)} \\
\geq & -\lambda'/2\|h'_{\bar{F}_0}\|_1 + \lambda'(-\|h'_{\bar{F}_0 - \bar{F}_1}\|_1 + \|h'_{\bar{F}_1}\|_1 - 2\|\bar{\beta}_{\bar{F}_1}\|_1) \\
= & \lambda'/2\|h'_{\bar{F}_1}\|_1 - \frac{3}{2}\lambda'\|h'_{\bar{F}_0 - \bar{F}_1}\|_1 - 2\lambda'\|\bar{\beta}_{\bar{F}_1}\|_1,
\end{aligned}$$

which implies the first inequality. ■

Similar to Lemma 11, the following result holds in the LASSO case:

**Lemma 20** *Assume $F_0 \subset F$ and $F_0 \subset F_1$ and the index set $\bar{F}_1$ is divided into a group of subsets $T_j$'s such that they satisfy all conditions in Lemma 10 with $v = h'$. Then the following holds:*

$$\begin{aligned}
\|h'_{\bar{F}_1 - T_1}\|_p \leq & l^{1/p-1}\left(3|\bar{F}_0 - \bar{F}_1|^{1-1/p}\|h'_{\bar{F}_0 - \bar{F}_1}\|_p + 4\|\bar{\beta}_{\bar{F}_1}\|_1\right) \\
\|h'\|_p \leq & \left[1 + 3^p(|\bar{F}_0 - \bar{F}_1|/l)^{p-1}\right]^{1/p}\|h'_{F_1 + T_1}\|_p + 4l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1,
\end{aligned}$$

*where $s_1 = |F_1|$.*

**Proof** Using the claim in Lemma 10 with $v = h'$, we have

$$\begin{aligned}
& \|h'_{\bar{F}_1 - T_1}\|_p \\
\leq & l^{1/p-1}\|h'_{\bar{F}_1}\|_1 \\
\leq & l^{1/p-1}\left(3\|h'_{\bar{F}_0 - \bar{F}_1}\|_1 + 4\|\bar{\beta}_{\bar{F}_1}\|_1\right) \quad \text{(due to the first inequality in Lemma 19)} \\
\leq & l^{1/p-1}\left(3|\bar{F}_0 - \bar{F}_1|^{1-1/p}\|h'_{\bar{F}_0 - \bar{F}_1}\|_p + 4\|\bar{\beta}_{\bar{F}_1}\|_1\right).
\end{aligned}$$

This proves the first inequality. Using this equality, we can obtain the second inequality as follows:

$$\begin{aligned}
& \|h'\|_p \\
= & (\|h'_{F_1 + T_1}\|_p^p + \|h'_{\bar{F}_1 - T_1}\|_p^p)^{1/p} \\
\leq & \left[\|h'_{F_1 + T_1}\|_p^p + \left(3\left(\frac{|\bar{F}_0 - \bar{F}_1|}{l}\right)^{1-1/p}\|h'_{\bar{F}_0 - \bar{F}_1}\|_p + \frac{4}{l^{1-1/p}}\|\bar{\beta}_{\bar{F}_1}\|_1\right)^p\right]^{1/p} \\
\leq & \left[1 + 3^p(|\bar{F}_0 - \bar{F}_1|/l)^{p-1}\right]^{1/p}\|h'_{F_1 + T_1}\|_p + 4l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1.
\end{aligned}$$

The last inequality is due to Equation (10). ■

Similar to the Lemma 12, the following result holds in the LASSO case:

**Theorem 21** *Under Assumption 1, taking* $F_0 \subset F$ *and* $\lambda' = 2\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)}$ *into the optimization problem* (7)*, if for any index set* $F_1$ *satisfying* $F_0 \subset F_1 \subset F$ *there exists some* $l$ *such that* $\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p} > 0$ *holds where* $s_1 = |F_1|$*, then with probability larger than* $1-\eta_1'$*, the* $\ell_p$ *norm* ($1 \le p \le \infty$) *of the difference between the optimizer of the problem* (7) *and the oracle solution is bounded as*

$$\begin{aligned}
&\|\hat{\beta}' - \bar{\beta}\|_p \\
\le &\frac{\left[1+3^p\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+(3/2)^p l)^{1/p}\lambda' + \frac{4\theta^{(p)}_{A,s_1+l,l}}{l^{1-1/p}}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} \\
&+ 4l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1
\end{aligned}$$

*and with probability larger than* $1-\eta_1'-\eta_2'$*, the* $\ell_p$ *norm* ($1 \le p \le \infty$) *of the difference between the optimizer of the problem* (7) *and the true solution is bounded as*

$$\begin{aligned}
&\|\hat{\beta}' - \beta^*\|_p \\
\le &\frac{\left[1+3^p\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+(3/2)^p l)^{1/p}\lambda' + \frac{4\theta^{(p)}_{A,s_1+l,l}}{l^{1-1/p}}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} \\
&+ 4l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.
\end{aligned}$$

**Proof** The proof follows the same strategy as in Theorem 12. First, we assume that Assumption 3 and the inequality (9) hold. Divide $\bar{F}_1$ into a group of subsets $T_j$'s ($j = 1, 2, \ldots$) without intersection

such that $\bigcup_j T_j = \bar{F_1}$, $\max_j |T_j| \le l$ and $\max_{i\in T_{j+1}} h_{T_{j+1}}[i] \le \|h_{T_j}\|_1/l$ hold. Since

$$\begin{aligned}
&\|X^T_{T_{01}+F_0} X h'\|_p \\
=&\|X^T_{T_{01}+F_0} X_{T_{01}+F_0} h'_{T_{01}+F_0} + \sum_{j\ge 2} X^T_{T_{01}+F_0} X_{T_j} h'_{T_j}\|_p \\
\ge&\mu^{(p)}_{A,s_1+l}\|h'_{T_{01}+F_0}\|_p - \sum_{j\ge 2}\theta^{(p)}_{A,s_1+l,l}\|h'_{T_j}\|_p \\
\ge&\mu^{(p)}_{A,s_1+l}\|h'_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l}\sum_{j\ge 2}\|h'_{T_j}\|_p \\
\ge&\mu^{(p)}_{A,s_1+l}\|h'_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|h'_{\bar{F_1}}\|_1 \\
\ge&\mu^{(p)}_{A,s_1+l}\|h'_{T_{01}+F_0}\|_p - \theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\left(3\|h'_{\bar{F_0}-\bar{F_1}}\|_1 + 4\|\bar{\beta}_{\bar{F_1}}\|_1\right) \\
&\quad \text{(due to the first inequality of Lemma 19)} \\
\ge&\mu^{(p)}_{A,s_1+l}\|h'_{T_{01}+F_0}\|_p - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{l}{|T_0|}\right)^{1/p-1}\|h'_{T_0}\|_p - 4\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F_1}}\|_1 \\
\ge&\left[\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|T_0|}{l}\right)^{1-1/p}\right]\|h'_{T_{01}+F_0}\|_p - 4\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F_1}}\|_1
\end{aligned}$$

and

$$\begin{aligned}
&\|X^T_{T_{01}+F_0} X h'\|^p_p \\
=&\|X^T_{F_0} X h'\|^p_p + \|X^T_{T_{01}\cap F} X h'\|^p_p + \|X^T_{T_{01}\cap \bar{F}} X h'\|^p_p \\
\le&|T_{01}\cap F|\lambda'^p + |T_{01}\cap\bar{F}|(3\lambda'/2)^p \quad \text{(due to Lemma 19)} \\
\le&|T_0\cap F|\lambda'^p + |T_1\cap F|\lambda'^p + |T_0\cap\bar{F}|(3\lambda'/2)^p + |T_1\cap\bar{F}|(3\lambda'/2)^p \quad \text{(due to } F_1\subset F\text{)} \\
\le&|T_0|\lambda'^p + l(3\lambda'/2)^p, \quad \text{(due to } T_0\cap\bar{F}=\varnothing\text{)}
\end{aligned}$$

thus we have

$$\|h'_{F_1+T_1}\|_p = \|h'_{T_{01}+F_0}\|_p \le \frac{(|\bar{F_0}-\bar{F_1}|+(3/2)^p l)^{1/p}\lambda' + 4\theta^{(p)}_{A,s_1+l,l} l^{1/p-1}\|\bar{\beta}_{\bar{F_1}}\|_1}{\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F_0}-\bar{F_1}|}{l}\right)^{1-1/p}}.$$

It follows that

$$\begin{aligned}
\|h'\|_p \le& \left[1+3^p\left(\frac{|\bar{F_0}-\bar{F_1}|}{l}\right)^{p-1}\right]^{1/p}\|h'_{F_1+T_1}\|_p + 4l^{1/p-1}\|\bar{\beta}_{\bar{F_1}}\|_1 \\
\le& \frac{\left[1+3^p\left(\frac{|\bar{F_0}-\bar{F_1}|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F_0}-\bar{F_1}|+(3/2)^p l)^{1/p}\lambda' + \frac{4\theta^{(p)}_{A,s_1+l,l}}{l^{1-1/p}}\|\bar{\beta}_{\bar{F_1}}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F_0}-\bar{F_1}|}{l}\right)^{1-1/p}} \\
&+ 4l^{1/p-1}\|\bar{\beta}_{\bar{F_1}}\|_1,
\end{aligned}$$

and

$$\begin{aligned}
&\|\hat{\beta}' - \beta^*\|_p \\
\leq &\frac{\left[1+3\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{p-1}\right]^{1/p}\left((|\bar{F}_0-\bar{F}_1|+(3/2)^p l)^{1/p}\lambda' + \frac{4\theta^{(p)}_{A,s_1+l,l}}{l^{1-1/p}}\|\bar{\beta}_{\bar{F}_1}\|_1\right)}{\mu^{(p)}_{A,s_1+l} - 3\theta^{(p)}_{A,s_1+l,l}\left(\frac{|\bar{F}_0-\bar{F}_1|}{l}\right)^{1-1/p}} \\
&+ 4l^{1/p-1}\|\bar{\beta}_{\bar{F}_1}\|_1 + \frac{s^{1/p}}{\mu^{(p)}_{(X_F^T X_F)^{1/2},s}}\sigma\sqrt{2\log(s/\eta_2)}.
\end{aligned}$$

Finally, taking

$$\lambda' = 2\sigma\sqrt{2\log\left(\frac{m-s}{\eta_1}\right)},$$

Lemma 8 (letting $\eta = \eta_1$) implies that Assumption 3 holds with probability larger than $1-\eta_1'$ and Lemma 7 (letting $\eta = \eta_2$) implies that Equation (9) holds with probability larger than $1-\eta_2'$. Thus, these two bounds above hold with probability larger than respectively $1-\eta_1'$ and $1-\eta_1'-\eta_2'$. ■

**Proof to Theorem 14:** By taking $F_1 = F$ in Theorem 21, the claims above can be obtained immediately. ■

**Proof to Theorem 15:** Please refer to the proof for Theorem 4. ■

**Proof to Theorem 16:** Please refer to the proof for Theorem 5. ■

**Proof to Theorem 17:** First, we assume that Assumption 3 and the inequality (9) holds. Then, the claim in Theorem 15 holds with probability 1. Since all conditions in Theorem 15 are satisfied, after $s$ iterations $s$ correct features can be chosen (i.e., $F_0^{(N)} = F$) with probability 1. Since all correct features are obtained, the optimization problem in the last iteration can be formulated as

$$\min : \frac{1}{2}\|X\beta - y\|_2^2 + \lambda'\|\beta_{\bar{F}}\|_1. \tag{12}$$

A minimizer should satisfy the following conditions:

$$\begin{aligned}
0 \in &X_{\bar{F}}^T(X\beta - y) + \lambda' sgn(\beta_{\bar{F}}) \\
0 = &X_F^T(X\beta - y),
\end{aligned} \tag{13}$$

where the first formula is based on the subdifferential set. Because of Assumption 3, the oracle solution satisfies these two conditions. Since the objective function is not strictly convex, we need to show that the oracle solution is the unique minimizer.

From the second equality in Equation (13), we have $\beta_F = -(X_F^T X_F)^{-1}X_F^T(X_{\bar{F}}\beta_{\bar{F}} - y)$. It follows that the objective function in Equation (12) can be expressed as

$$f(\beta_{\bar{F}}) = \frac{1}{2}\|(I - X_F(X_F^T X_F)^{-1}X_F^T)(X_{\bar{F}}\beta_{\bar{F}} - y)\|_2^2 + \lambda'\|\beta_{\bar{F}}\|_1.$$

Because the oracle solution is a minimizer of the Equation (12), "0" should be one of the minimizers of $f(\beta_{\bar{F}})$. Next we show that "0" is the unique minimizer, which implies that the oracle solution

is the unique minimizer for Equation (12). We can compute the directional derivative along any direction $\Delta$ at the point "0" for the function $f(\beta_{\bar{F}})$ as follows:

$$\begin{aligned}\frac{df(0+t\Delta)}{dt}|_{t=0} &= -y^T(I - X_F(X_F^TX_F)^{-1}X_F^T)X_{\bar{F}}^T\Delta + \lambda'\|\Delta\|_1 \\ &\geq \lambda'\|\Delta\|_1 - \|\Delta\|_1\|y^T(I - X_F(X_F^TX_F)^{-1}X_F^T)X_{\bar{F}}^T\|_\infty \\ &= \|\Delta\|_1(\lambda' - \|X_{\bar{F}}^T(X\bar{\beta} - y)\|_\infty) \\ &> 0. \quad \text{(due to Assumption 3)}\end{aligned}$$

Thus, the directional derivative at "0" is always strictly greater than 0 at arbitrary directions, which shows that "0" should be the unique minimizer for $f(\beta_{\bar{F}})$.

Finally, because the probability of Assumption 3 and the inequality (9) holding is larger than $1-\eta_1'-\eta_2'$, the oracle solution is achieved with the same probability. ■

## References

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via $\ell_1$ minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, 2009.

E. J. Candès and Y. Plan. Near-ideal model selection by $\ell_1$ minimization. *Annals of Statistics*, 37 (5A):2145–2177, 2009.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

E. J. Candès and T. Tao. The Dantzig selector: Statistical estimation when *p* is much larger than *n*. *Annals of Statistics*, 35(6):2313–2351, 2007.

D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *(invited review article) Statistica Sinica*, 20:101–148, 2010.

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

G. M. James, P. Radchenko, and J. Lv. DASSO: connections between the Dantzig selector and Lasso. *Journal of The Royal Statistical Society Series B*, 71(1):127–142, 2009.

V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines on-line learning and bandits. In *Proceedings of the Twenty-First Annual Conference on Learning Theory (COLT)*, pages 229–238, Helsinki, Finland, 2008.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528, 2009.

N. Meinshausen, P. Bhlmann, and E. Zrich. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

P. Ravikumar, G. Raskutti, M. J. Wainwright, and B. Yu. Model selection in gaussian graphical models: High-dimensional consistency of $\ell_1$-regularized MLE. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1329–1336, Vancouver, British Columbia, Canada, 2008.

J. Romberg. The Dantzig selector and generalized thresholding. In *Proceedings of the Forty-Second Annual Conference on Information Sciences and Systems (CISS)*, pages 22–25, Princeton, New Jersey, USA, 2008.

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

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010a.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high dimensional sparse estimation problems. Technical report, Department of Statistics, Rutgers University, Piscataway, New Jersey, USA, 2012.

T. Zhang. Some sharp performance bounds for least squares regression with $\ell_1$ regularization. *Annals of Statistics*, 37(5A):2109–2114, 2009a.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009b.

T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010b.

T. Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):5215–6221, 2011a.

T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011b.

T. Zhang. Multi-stage convex relaxation for feature selection. Technical report, Department of Statistics, Rutgers University, Piscataway, New Jersey, USA, 2011c.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2304–2312, Vancouver, British Columbia, Canada, 2009.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.