Statistica Sinica **23** (2013), 615-634
doi:http://dx.doi.org/10.5705/ss.2012.061

MODEL SELECTION CONSISTENCY OF DANTZIG SELECTOR

Yujie Gai, Lixing Zhu and Lu Lin

Central University of Finance and Economics, Hong Kong Baptist University and Shandong University

Abstract: Consistency of model selection hinges on the correlation between significant and insignificant predictors for "large p, small n" problems. Thus, Irrepresentable Conditions play an important role in consistency, that insignificant predictors are irrepresentable by significant ones. In this paper, we provide Irrepresentable Conditions when the Dantzig selector is applied; they ensure that the Dantzig selector consistently selects the true model with fixed p and diverging p (number of predictors) even at an exponential rate of n. Our conditions are sufficient for a strong sign consistency and Weak Irrepresentable Conditions are necessary for a weak sign consistency. Strong sign consistency leads to the conventional consistency of the estimation. As a by-product, the results also show the difference between the Dantzig selector and the Lasso when consistency is at issoe. Simulation studies are performed to examine the theoretical results.

Key words and phrases: Consistency, Dantzig selector, Irrepresentable Conditions.

1. Introduction

There are a large number of references referring to variable selection due to its importance in applications. The aim of variable selection is to select a subset of significant predictors of a given outcome from a large collection of candidate predictors so that the low dimensional sub-model can be regarded as a working model.

Penalized likelihood methods have been extensively studied as useful tools. Examples include the Lasso (Tibshirani (1996)), the SCAD (Fan (1997); Fan and Li (2001)) and the Elastic-Net (Zou and Hastie (2005)). For all existing methods, consistency of model selection and estimation have been investigated. It is well known that consistency hinges critically on the correlation between significant and insignificant predictors. For instance, the Lasso estimator is in general biased (Zou (2006)), especially when the number of significant entries is relatively large. However, there are few works about this, particularly for "large p, small n" paradigms. To the best of our knowledge, Zhao and Yu (2006) first proposed an almost necessary and sufficient condition to guarantee the consistency of model selection by the Lasso. They called it the Irrepresentable Condition.

Although the Dantzig selector (Candes and Tao (2007)) is in a certain sense asymptotically equivalent to the Lasso (Bickel, Ritov, and Tsybakov (2009); Asif and Romberg (2010)), their estimation consistency requires different conditions on the correlations between predictors because the Dantzig selector is related to an estimating equation, whereas the Lasso requires a specific likelihood or an objective function. We will see this very clearly when we compare the corresponding results for the Lasso, Zhao and Yu (2006), and the Dantzig selector. The two methods depend on different correlation structures of predictors for sign consistency. Dicker and Lin (2009, 2011) considered random design of predictors and suggested Irrepresentable Conditions for the Dantzig selector in the fixed p case. Their method, however, cannot be extended to handle the case of p growing with n or the $p > n$ paradigm. In this paper, we consider fixed design with both fixed p and diverging p, even $p = \exp(n^c)$ for some constant $c > 0$. Irrepresentable Conditions are provided for the sign consistency of model selection. These conditions are sufficient for a strong sign consistency and necessary for a weak sign consistency. These two consistencies are defined in the next section. Moreover, after shrinking the ultra-high dimension to a value that is smaller than the sample size, we also provide the conventional consistency of estimation when the dimension q of significant predictors is of a rate of $o(n)$.

The rest of this paper is organized as follows. In Section 2, we provide Irrepresentable Conditions that are sufficient for a strong sign consistency and necessary for a weak sign consistency when p is fixed. When p grows at a polynomial rate in n, or even at an exponential rate, we prove that the Dantzig estimator is still strongly sign consistent under Irrepresentable Conditions. The conventional consistency of estimation is shown at the end of Section 2. Section 3 contains numerical studies that examine the theoretical results. Some concluding remarks are included in Section 4. The proofs of theorems are postponed to the Appendix.

2. Irrepresentable Conditions and Sign Consistencies

Consider the linear regression model

$$y = \mathbf{X}\beta + \varepsilon, \tag{2.1}$$

where $y = (Y_1, \ldots, Y_n)^\tau$ is a $n \times 1$ response, $\mathbf{X} = (X_1, \ldots, X_n)^\tau = (X^1, \ldots, X^p)$ is a $n \times p$ fixed design matrix with X_i as the ith row of $\mathbf{X}$, X^j as the jth column of $\mathbf{X}$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\tau$ an n-vector of i.i.d random errors with $E(\varepsilon_1) = 0$ and $E(\varepsilon_1^2) = \sigma^2$. Here τ stands for transposition of vector or matrix.

For simplicity, we assume the response $y = (Y_1, \ldots, Y_n)^\tau$ is centralized and the design matrix is standardized so that

$$\mathbf{1}_n \times y = 0,\ \mathbf{1}_n \times X^j = 0,\ \frac{(X^j)^\tau (X^j)}{n} = 1, j = 1, \ldots, p, \tag{2.2}$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of all components 1.

The Dantzig selector estimator $\hat{\beta}^D$ is defined as

$$\hat{\beta}^D = \text{argmin}\ _{\beta}\|\beta\|_1 \qquad \text{s.t. } \|\mathbf{X}^T(y - \mathbf{X}\beta)\|_\infty \leq \lambda. \tag{2.3}$$

Let $T^* = \{j : \beta_j \neq 0\}$ and $\hat{T} = \{j : \hat{\beta}_j \neq 0\}$. We refer to T^* as the true model. Let $|T^*| = q$ with $|T^*|$ the number of elements in the set T^*.

As Asif and Romberg (2010) commented, a dual problem to (2.3) is to have a maximizer $\hat{\mu}$ over all μ of the following function

$$-(\lambda\|\mu\|_1 + <\mu, \mathbf{X}^T y>) \qquad \text{s.t. } \|\mathbf{X}^T\mathbf{X}\mu\|_\infty \leq 1; \tag{2.4}$$

both $\hat{\beta}^D$ and $\hat{\mu}$ are for a given value of λ. Asif and Romberg (2010) showed that (2.5)-(2.8) are necessary and sufficient for $(\hat{\beta}^D, \hat{\mu})$ to be the unique primal-dual solution pair to (2.3) and (2.4):

$$\|\mathbf{X}^\tau_{\bar{E}}(y - \mathbf{X}\hat{\beta}^D)\|_\infty < \lambda, \tag{2.5}$$

$$\|\mathbf{X}^\tau_{\bar{T}}\mathbf{X}\hat{\mu}\|_\infty < 1, \tag{2.6}$$

$$\mathbf{X}^\tau_{T^*}\mathbf{X}\hat{\mu} = \text{sign}(\hat{\beta}^D_{T^*}), \tag{2.7}$$

$$\mathbf{X}^\tau_E(y - \mathbf{X}\hat{\beta}^D) = \lambda\,\text{sign}(\hat{\mu}_E), \tag{2.8}$$

In this section, we study the consistency of $\hat{\beta}^D$. Following the notations in Zhao and Yu (2006) and Wainwright (2006), $\hat{\beta}^D =_s \beta$ means that both $\hat{\beta}^D$ and β have the same sign element-wise.

Definition 1. The solution of (2.3) $\hat{\beta}^D$ is *strongly sign consistent* if there exists $\lambda = \lambda(n)$ such that

$$\lim_{n \to \infty} P(\hat{\beta}^D(\lambda) =_s \beta) = 1. \tag{2.9}$$

Definition 2. The solution of (2.3) $\hat{\beta}^D$ is *weakly sign consistent* if

$$\lim_{n \to \infty} P(\exists \lambda \geq 0, \hat{\beta}^D(\lambda) =_s \beta) = 1. \tag{2.10}$$

We need some notation. Let $C = (1/n)\mathbf{X}^\tau\mathbf{X}$. For any subset $T \subset \{1, \ldots, p\}$, $|T|$ denotes the number of elements in subset T, $\bar{T}$ is the complement of T in the set $\{1, \ldots, p\}$. Let $\beta_T = (\beta_j)_{j \in T}$ be the $|T| \times 1$ vector whose entries are those of β indexed by T. Similarly, $\mathbf{X}_T$ is defined as the $n \times |T|$ matrix whose columns are those of $\mathbf{X}$ indexed by T. Given a $p \times p$ matrix C and subsets $T_1,\ T_2 \subseteq \{1, \ldots, p\}$, let C_{T_1,T_2} be the $|T_1| \times |T_2|$ sub-matrix from C with rows corresponding to T_1 and columns corresponding to T_2. Let diag(β) be the diagonal matrix with diagonal β. For $\beta \in \mathbf{R}^p$, $\text{sign}(\beta) = (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p))^\tau$ is the signal function of β, where

$$\text{sign}(\beta_i) = \begin{cases} 1, & \beta_i > 0, \\ 0, & \beta_i = 0, \\ -1, & \beta_i < 0. \end{cases} \tag{2.11}$$

Assume $C_{T^*,E}$ is invertible for some $E \subset \{1,\ldots,p\}$ with $|E| = |T^*|$.

Irrepresentable Conditions. The inequality

$$|C_{\bar{T}^*,E} C_{T^*,E}^{-1} \mathrm{sign}(\beta_{T^*})| < \mathbf{1} \tag{2.12}$$

holds and there exists a positive constant vector η satisfying

$$\left| C_{\bar{E},T^*} C_{E,T^*}^{-1} \mathrm{sign}\Big(C_{T^*,E}^{-1} \mathrm{sign}(\beta_{T^*}) \Big) \right| \leq \mathbf{1} - \eta, \tag{2.13}$$

where $\mathbf{1}$ is a $(p-q) \times 1$ vector of all components 1 and $|\cdot|$ means the two inequalities hold element-wise in absolute value.

Weak Irrepresentable Conditions. The inequalities

$$|C_{\bar{T}^*,E} C_{T^*,E}^{-1} \mathrm{sign}(\beta_{T^*})| < \mathbf{1}, \tag{2.14}$$

$$\left| C_{\bar{E},T^*} C_{E,T^*}^{-1} \mathrm{sign}\Big(C_{T^*,E}^{-1} \mathrm{sign}(\beta_{T^*}) \Big) \right| < \mathbf{1} \tag{2.15}$$

hold, where $\mathbf{1}$ is the $(p-q) \times 1$ vector of all components 1 and $|\cdot|$ means the two inequalities hold element-wise in absolute value.

We note that the Irrepresentable Conditions for the Dantzig selector are different from those for the Lasso. The former are much more complex than the latter because of the differences in the Karush-Kuhn-Tucker (KKT) conditions for the two problems. The conditions we provide are different from those in Dicker and Lin (2009) for random design; their conditions cannot be extended to handle $p > n$ or the diverging dimension of the true model.

Proposition 1. *The Irrepresentable Condition yields*

$$P(\hat{\beta}^D =_s \beta) \geq P(A_n \cap B_n) \tag{2.16}$$

for

$$A_n = \{|Z_{\bar{E}} - C_{\bar{E},T^*} C_{E,T^*}^{-1} Z_E| \leq \frac{\lambda \eta}{\sqrt{n}}\},$$

$$B_n = \{|D C_{E,T^*}^{-1} Z_E| < \sqrt{n}\{|\beta_{T^*}| - \frac{\lambda}{n} |D C_{E,T^*}^{-1} \mathrm{sign}(\tilde{\mu}_E)|\}\},$$

where $Z_E = (\sqrt{n})^{-1} \mathbf{X}_E^{\tau} \varepsilon$, $Z_{\bar{E}} = (\sqrt{n})^{-1} \mathbf{X}_{\bar{E}}^{\tau} \varepsilon$, $D = \mathrm{diag}(\mathrm{sign}(\beta_{\mathrm{T}^*}))$, *and* $\tilde{\mu}_E = (\mathbf{X}_{T^*}^{\tau} \mathbf{X}_E)^{-1} \mathrm{sign}(\beta_{\mathrm{T}^*})$.

In the following, we show that the Irrepresentable Conditions are sufficient for strong sign consistency, and the Weak Irrepresentable Conditions are necessary for weak sign consistency.

2.1. Sign consistency with fixed p and q

For insight, we consider the case with fixed p and q first. In this setting, it is natural to assume the regularity conditions

(a) $C = C(n) \to C^*$as $n \to \infty$, where C^* is a positive definite matrix,

(b) $\dfrac{1}{n} \max\limits_{1 \le i \le n} X_i{}^\tau X_i \to 0,$ as $n \to \infty$.

When the X_i's are i.i.d. with a finite variance and p is fixed, then $C = C(n) \to C^*$ with $C^* = E X_1{}^\tau X_1$ and $\max_{1 \le i \le n} X_i{}^\tau X_i = o_p(n)$ (see Owen (2001)). Hence (a) and (b) hold trivially.

Theorem 1. *If p and q are fixed and* (a), (b), *and the Irrepresentable Conditions hold, for positive λ satisfying $\lambda/n \to 0$ and $\lambda/n^{(c+1)/2} \to \infty$ with $0 \le c < 1$,*

$$P(\hat{\beta}^D(\lambda) =_s \beta) = 1 - o(e^{-n^c}) \to 1 \textit{ as } n \to \infty.$$

Theorem 1 states that, for fixed p and q, under some mild conditions and the Irrepresentable Conditions, the probability that the Dantzig selector selects the true model approaches 1 at an exponential rate. Moreover, the Weak Irrepresentable Conditions are necessary for the weak sign consistency.

Theorem 2. *For fixed p and q, if* (a) *and* (b) *hold, and if the Dantzig selector based estimator is weakly sign consistent, then there exists an N such that Weak Irrepresentable Conditions hold for $n \ge N$.*

2.2. Sign consistency when both p and $q \to \infty$

Now we turn to the case where both p and q grow with n. Here we need new conditions. Let κ_{n1} be the largest eigenvalue of B, $B = C_{E,T^*}^{-1} C_{E,E} C_{T^*,E}^{-1}$, with E satisfying the Irrepresentable Conditions, and let τ_{n1} be the largest eigenvalue of the semi-positive definite matrix $(I - K)(I - K)^\tau$ with idempotent $K = \mathbf{X}_{T^*}(\mathbf{X}_E^\tau \mathbf{X}_{T^*})^{-1} \mathbf{X}_E^\tau$. Assume that there exist $0 \le d_1 < d_2 \le 1$ such that

(C1) $q = O(n^{d_1})$,

(C2) $n^{(1-d_2)/2} \min\limits_{i \in T^*} |\beta_i| \ge M_1 > 0$,

(C3) $\|C_{E,T^*}\alpha\|_2^2 \ge M_2 > 0$ for any unit vector α, with E satisfying Irrepresentable Conditions,

(C4) $0 < \kappa_{n1} \le \kappa_1 < \infty$,

(C5) $\tau_{n1} \le \tau_1 < \infty$.

Under these conditions and certain conditions on the moments of the error, the Dantzig selector of (2.3) can select the true model consistently provided that the Irrepresentable Conditions hold.

Theorem 3. *Assume the ε_i are i.i.d. random variables with $E(\varepsilon_i)^{2k} < \infty$ for some integer $k > 0$. If* (C1)−(C5) *hold, $p = o(n^{(d_2-d_1)k})$ for $d_2 > d_1$, and λ satisfies $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$, and $(\lambda/\sqrt{n})^{2k}/p \to \infty$, then the Irrepresentable Conditions imply that*

$$P(\hat{\beta}^D(\lambda) =_s \beta) \geq 1 - O\Big(\frac{pn^k}{\lambda^{2k}}\Big) \to \ 1 \text{ as n} \to \infty. \tag{2.17}$$

When E in the Irrepresentable Conditions satisfies $E = T^*$, (C4) is implied by (C3), and (C5) is satisfied automatically, since $(I-K)(I-K)^\tau$ is then a symmetric idempotent matrix whose eigenvalue is at most 1.

Corollary 1. *If the conditions of Theorem* 3 *hold with the exception of* (C4) *and* (C5)*, and if the Irrepresentable Conditions hold for $E = T^*$,*

$$P(\hat{\beta}^D(\lambda) =_s \beta) \geq 1 - O\Big(\frac{pn^k}{\lambda^{2k}}\Big) \to \ 1 \text{ as n} \to \infty. \tag{2.18}$$

2.3. Sign consistency for $p = \exp(n^c)$ and $q \to \infty$

From Theorem 3, we have that sign consistency for the Dantzig selector can be obtained under the assumption that the errors have finite $(2k)$-th order moment for an integer $k > 0$. However, even if all orders of moments of the error exist, p can only grow at a polynomial rate in n. In this section, we consider that p grows exponentially with n. As the cost for doing this, we assume a sub-Gaussian tail for the error.

Theorem 4. *Let* (C1)−(C5) *hold, and suppose the ε_i are i.i.d. random variables that, for some constants $1 \leq d \leq 2, C > 0$ and K, satisfy $P(|\varepsilon_i| > x) \leq K \exp(-Cx^d)$ for all $x \geq 0$ and $i = 1, 2, \ldots$. When $p = e^{n^{d_3}}$, $(1/\log n)^{I\{d=1\}}(\lambda/\sqrt{n})^d = O(n^{d_4})$, and $0 < d_3 < d_4 < d_2 d/2$, the Irrepresentable Conditions imply that*

$$P(\hat{\beta}^D(\lambda) =_s \beta) \geq 1 - O(e^{-n^\delta}) \to 1 \text{ as } n \to \infty, \tag{2.19}$$

where $\delta = \min\{d_4 - d_3, d_2 d/2\}$.

Corollary 2. *Assume the conditions of Theorem* 4 *except for* (C4) *and* (C5)*. Then the Irrepresentable Conditions with $E = T^*$ yield*

$$P(\hat{\beta}^D(\lambda) =_s \beta) = 1 - O(e^{-n^\delta}) \to 1 \text{ as } n \to \infty, \tag{2.20}$$

where $\delta = \min\{d_4 - d_3, d_2 d/2\}$.

Compared with Theorem 3, Theorem 4 and Corollary 2 tell us that p exponential in n is possible if the error terms have lighter tails, such as sub-Gaussian.

2.4. Conventional consistency of estimation after variable selection

In this section, we investigate the conventional consistency of the estimator after variable selection, with $q = o(n)$.

Write the $n \times |\hat{T}|$ design matrix as $\mathbf{X}_{\hat{T}} \triangleq (X_{1\hat{T}}, \ldots, X_{n\hat{T}})^{\tau}$. The post-selection least squares estimator $\hat{\beta}$ of β is

$$\hat{\beta}_{\hat{T}} = C^{-1}_{\hat{T},\hat{T}} \{ \frac{1}{n} X^{\tau}_{\hat{T}} Y \} = C^{-1}_{\hat{T},\hat{T}} \{ \frac{1}{n} \sum_{i=1}^{n} X_{i\hat{T}} Y_i \}, \ \hat{\beta}_{\bar{\hat{T}}} = 0. \tag{2.21}$$

Theorem 5. *Assume that the Dantzig selector estimator is strongly sign consistent, and the* $\max\limits_{1 \le i \le n, 1 \le j \le q} x^2_{iT^*,j} < \infty$ *holds, where* $X^{\tau}_{iT^*} = (x_{iT^*,1}, \ldots, x_{iT^*,q})$ *corresponding to the ith row of matrix* $\mathbf{X}_{T^*}$*. Then*

$$\|\hat{\beta} - \beta\|_2 = O_p\Big(\sqrt{\frac{q}{n}}\Big). \tag{2.22}$$

Consistency here is quite different from that of the classical least squares estimator, since the dimension of $X_{\hat{T}}$ is $|\hat{T}|$, also random.

3. Numerical Studies

Simulation studies were conducted to examine the role of Irrepresentable Conditions in model selection consistency. For this purpose, we considered three examples. In the first example, the sample size was 1,000 and p and q were comparably small; the second featured a relatively large dimension p. In the third example, the dimension p was greater than the sample size, and we also considered a dataset analysed by Tibshirani (1996) using the Lasso. We use it to see, in practice, how we can check estimation consistency when the Dantzig selector is applied.

Consider $p = 6$ and $q = 5$ with the sample size 1,000. $x_1, x_2, x_3, x_4, x_5, e$, and ε were standard normal, then x_6 was generated as

$$x_6 = \frac{1}{8}x_1 + \frac{1}{4}x_2 + \frac{1}{2}x_3 + \frac{1}{2}x_4 + \frac{1}{2}x_5 + \frac{\sqrt{11}}{8}e. \tag{3.1}$$

The regression model was

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon. \tag{3.2}$$

And we took two cases: (a) $\beta = (4, 2, 0.5, -0.6, -0.7)^{\tau}$; (b) $\beta = (-4, -2, 0.5, 0.6, 0.7)^{\tau}$. Here $T^* = \{1, 2, 3, 4, 5\}$. It is easy to check that for (a), with $E = T^*$, $C_{\bar{T}^*,E} C^{-1}_{T^*,E} = C_{\bar{E},T^*} C^{-1}_{E,T^*} = (1/8, 1/4, 1/2, 1/2, 1/2)^{\tau}$, both Irrepresentable

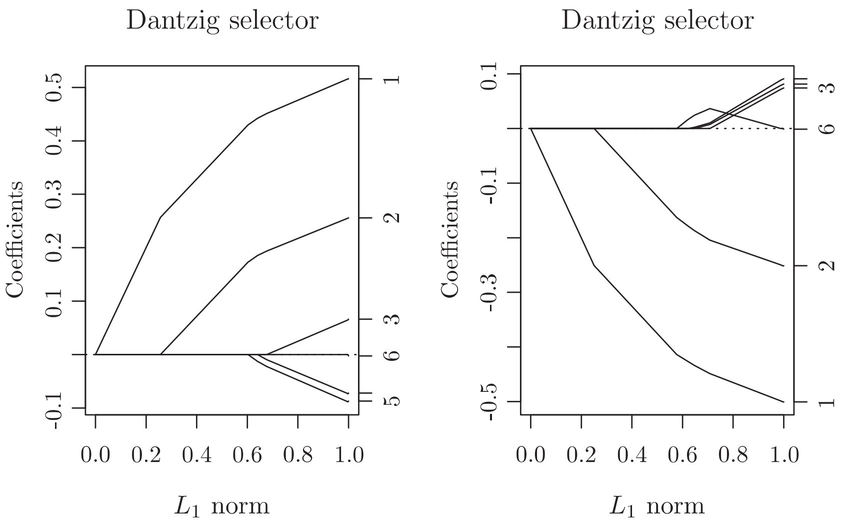

Figure 1. the Dantzig selector solution paths for settings (a) and (b). The left is for setting (a), and the right for setting (b).

Conditions and Weak Irrepresentable Conditions hold. For (b), the Irrepresentable Conditions do not hold, whatever be E. Figure 1 plots the estimated coefficients for various values of λ with the L_1 norm of the coefficient vector on the $X-$axis. The largest L_1 norm is attained when $\lambda = 0$ because this does no shrinkage of coefficients. For each plot, we here divided the original estimated coefficients by the corresponding largest L_1 norm among them. Figure 1 shows results that accord with our theoretical results. We conclude that in case (a), the Dantzig selector can select significant components successfully, while in case (b) it fails, as in a certain range of λ, X_6 is included in the working model.

Here $p = 50, 100$, and 200, with sample size 1,000. x_1, x_2, e, and ε were standard normal. x_3 was generated as

$$x_3 = \frac{2}{3}x_1 + \frac{2}{3}x_2 + \frac{1}{3}e,$$

with the true model assumed to be

$$Y = x_1\beta_1 + x_2\beta_2 + \varepsilon. \tag{3.3}$$

We generated i.i.d. $x_4, \ldots, x_p$ as standard normal and considered (a') $\beta = (-2, 3, 0, \ldots, 0)^\tau$; and (b') $\beta = (2, 3, 0, \ldots, 0)^\tau$.

Take $E = T^* = \{1, 2\}$. It is easy to obtain that $C_{T^*,E} = I_2$ and $C_{\bar{T}^*,E}C_{T^*,E}^{-1} = C_{\bar{E},T^*}C_{E,T^*}^{-1} = (2/3, 2/3)$. Then for case (a'), both the Irrepresentable and Weak Irrepresentable Conditions hold. For case (b'), whatever be E, at least one of

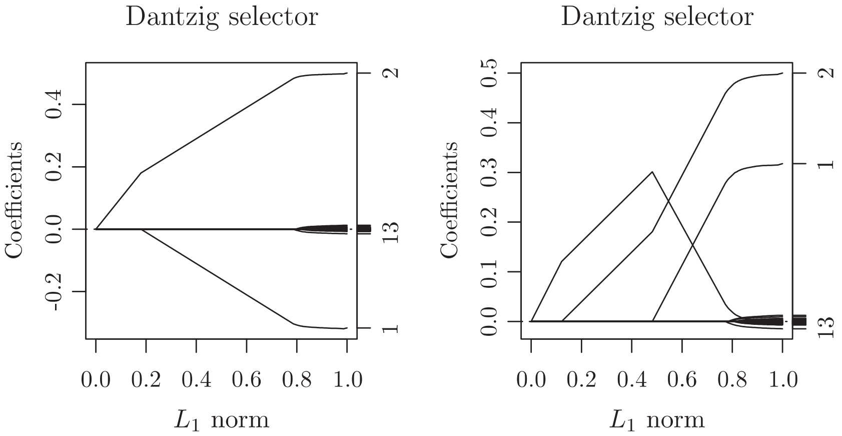

Figure 2. the Dantzig selector solution paths for settings (a') and (b') when $p = 50$. The left one is for setting (a'), and the right one for setting (b').

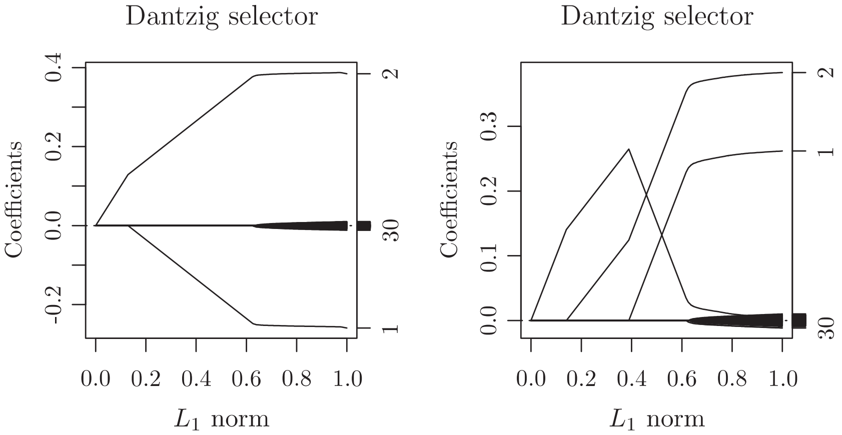

Figure 3. the Dantzig selector solution paths for settings (a') and (b') when $p = 100$. The left one is for setting (a'), and the right one for setting (b').

the equalities in the Weak Irrepresentable Conditions fails. Figures 2−4 show results that coincide with the theoretical analysis, all three values of p. Thus, in case (a'), the Dantzig selector selects the true model while for case (b') it does not, as X_3 is often included when λ is in a certain range.

Here we consider $p > n$ with $n = 30, p = 50$, and $q = 3$, and with $n = 50, p = 100$, and $q = 8$. The design matrix $\mathbf{X}$ was generated as multivariate standard normal, and $\beta = [\beta_{(1)}^{\tau}, \beta_{(2)}^{\tau}]^{\tau}$, where $\beta_{(1)}$ was a $q-$dimensional vector with all entries 1 and $\beta_{(2)}$ a $(p-q)$-dimensional zero vector. The true model was $Y = \mathbf{X}\beta + \varepsilon$, with ε being generated as the standard normal.

It is easy to verify that the Irrepresentable Conditions hold. Thus the Dantzig selector can select the significant predictors successfully. The selection

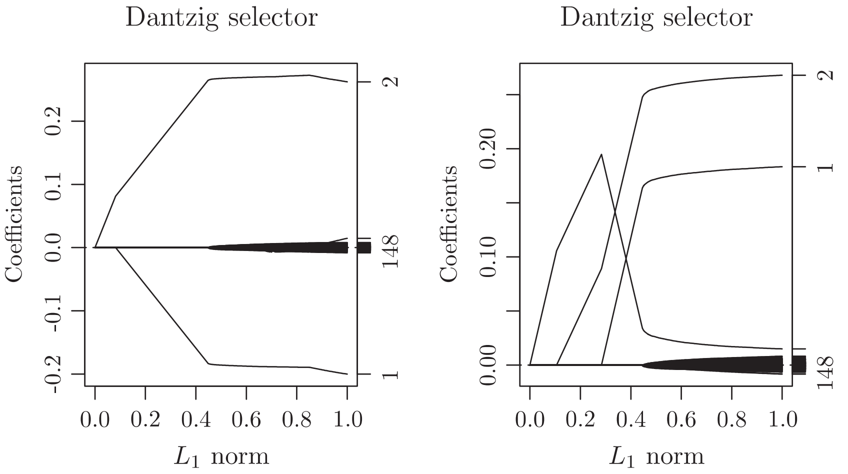

Figure 4. the Dantzig selector solution paths for settings (a') and (b') when $p = 200$. The left one is for setting (a'), and the right one for setting (b').

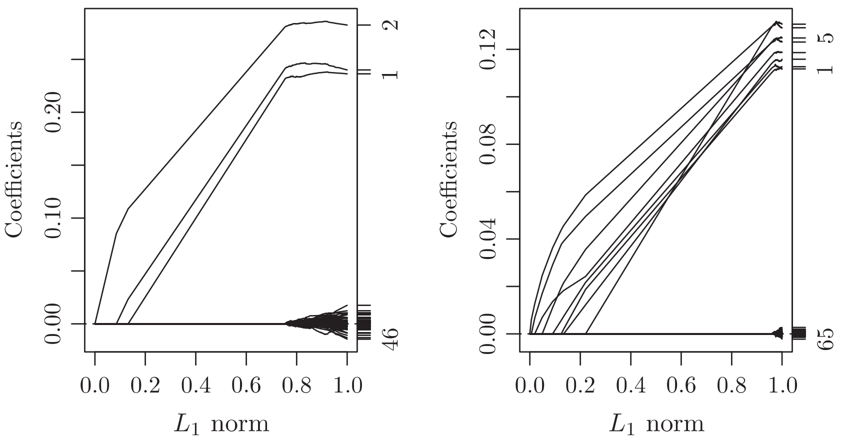

Figure 5. the Dantzig selector solution paths when $p > n$. The left hand side is with $n = 30, p = 50, q = 3$, and the right hand side is with $n = 50, p = 100, q = 8$.

paths are shown in Figure 5.

Data Example. We applied the Dantzig selector to the Prostate Cancer Data on which Tibshirani (1996) used the Lasso for variable selection. The data come from a study of the correlation between the level of prostate specific antigen and a number of clinical measures in men about to receive a radical prostatectomy. The data frame was 97 rows and 9 columns. As Tibshirani (1996) did, a linear model was fitted to log(prostate specific antigen) after standardizing the predictors.

To check both the Irrepresentable and Weak Irrepresentable Conditions, we

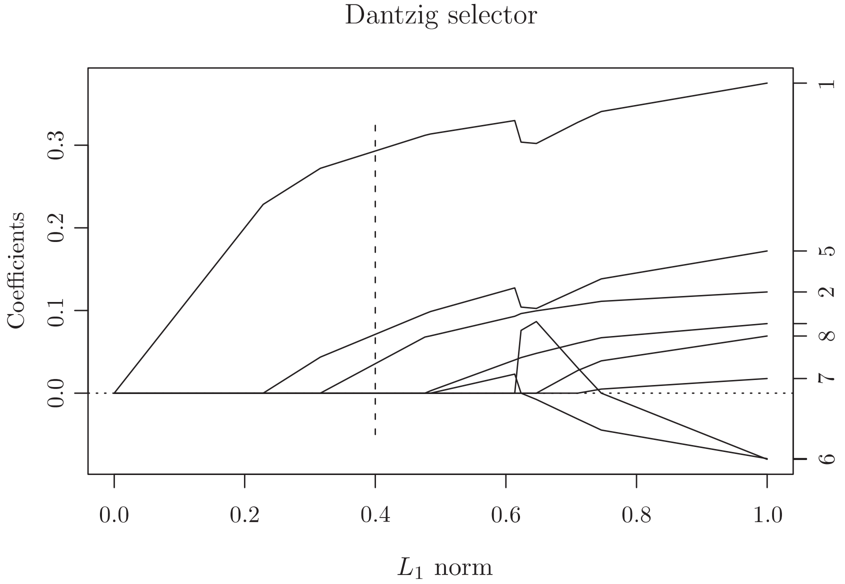

Figure 6. the Dantzig selector solution paths for Prostate Cancer Data.

find T^* and sign(β^*_{T}) first, then compute the estimators of the coefficients by fitting the linear model of Y against the predictors $x_1, \ldots, x_8$. Sorting the absolute values of the estimators in decreasing order, we have an order of $\{1, 5, 2, 4, 6, 3, 8, 7\}$. With $T^* = \{1, 5, 2\}$, the signs of β^*_T are taken as those of the estimators. Then it is easy to check that both the Irrepresentable and Weak Irrepresentable Conditions hold when $E = T^*$. Thus when the Dantzig selector is applied, theoretically, it is possible to correctly select the true model. From the selection path shown in Figure 6, we can see that the Dantzig selector does work. This selection is exactly as that of Tibshirani (1996). Based on this analysis, we can see that in practice, the Irrepresentable Conditions can be used as a tool to check whether the Dantzig selector works or not. As we have done , we can select the predictors first and assume them to be the significant ones to obtain the set T^*. Subsequently, we can check the Irrepresentable Conditions. Should the conditions not be satisfied, we may need a check for the rationality of the selection to see whether we need to try another variable selection method.

4. Concluding Remarks

We have investigated the model selection consistency of the Dantzig selector, and have found the dimension can be as much as exponential in sample size. Thus there are difference between the Lasso and the Dantzig selector as far as consistency is concerned. The Irrepresentable Conditions (ICs) for the Dantzig selector are more complex than those for the Lasso.

Whereas there are a couple of papers discussing the equivalence between the Lasso and the Dantzig selector in the literature, we suggest that such an equivalence only holds in some special scenarios. Asif and Romberg (2010) proved that the Lasso and the Dantzig Selector share the same homotopy path under some conditions, and the ICs for the dantzig selector are then identical to that for the Lasso. James, Radchenko, and Lv (2009) provided two types of design matrix $\mathbf{X}$ for which, under proper conditions, equivalence also holds. However, in general, the Danyzig selector is not equivalent to the Lasso.

When $p > n$ and the Irrepresentable Conditions are not satisfied, how to construct an adaptive Dantzig selector deserves further study.

Acknowledgements

The corresponding author: lzhu@hkbu.edu.hk. The research described here was supported by NNSF projects (11201499 and 11171188) of China, Mathematical Finance-Backward Stochastic Analysis and Computations in Financial Risk Control of China (10921101), NSF projects (ZR2010AZ001 and ZR2011AQ007) of Shandong Province of China and SRRF project (BS2011SF006) of Shandong Province of China and K C Wong-HKBU Fellowship Programme for Mainland China Scholars 2010-11, and a grant from the Research Grants Council of Hong Kong, and a grant from Hong Kong Baptist University, Hong Kong. The authors thank the Editor, an associate editor and two referees for their constructive comments and suggestions on the uniqueness of the Dantzig selector, and the kind help by the associate editor to provide us a relevant reference that led to a significant improvement of an early manuscript.

Appendix: Proofs

Proof of Proposition 1. For some subset $E \subset \{1, \ldots, p\}$ that satisfies the Irrepresentable Conditions, let

$$\tilde{\mu}_E = (\mathbf{X}^{\tau}_{T^*}\mathbf{X}_E)^{-1}\mathrm{sign}(\beta_{\mathrm{T}^*}), \quad \tilde{\mu}_{\bar{\mathrm{E}}} = 0, \tag{A.1}$$

$$\tilde{\beta}_{T^*} = (\mathbf{X}^{\tau}_{E}\mathbf{X}_{T^*})^{-1}\mathbf{X}^{\tau}_{E} y - \lambda(\mathbf{X}^{\tau}_{E}\mathbf{X}_{T^*})^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}}), \quad \tilde{\beta}_{\bar{\mathrm{T}}^*} = 0. \tag{A.2}$$

We have that if $(\tilde{\mu}, \tilde{\beta})$ satisfies (2.5)$-$(2.8), then $\tilde{\beta}$ is the unique solution to (2.3). Thus, we only need to prove that when both A_n and B_n happen, Proposition 2.1, $(\tilde{\mu}, \tilde{\beta})$ as (A.1) and (A.2) satisfy conditions (2.5) through (2.8). We check these conditions one by one. plugging (A.2) into (2.5), we have

$$\mathbf{X}^{\tau}_{E}(y - \mathbf{X}\tilde{\beta}) = \mathbf{X}^{\tau}_{E}(y - \mathbf{X}_{T^*}\tilde{\beta}_{T^*}) = \lambda \mathrm{sign}(\tilde{\mu}_{\mathrm{E}}). \tag{A.3}$$

Meanwhile, by (2.13), we have

$$\begin{aligned}
A_n &= \Big\{|Z_{\bar{E}} - C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E| \le \frac{\lambda\eta}{\sqrt{n}}\Big\} \\
&\subseteq \Big\{\|Z_{\bar{E}} - C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E\|_\infty < \frac{\lambda}{\sqrt{n}}(1 - \|C_{\bar{E},T^*}C_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})\|_\infty)\Big\} \\
&= \Big\{\|\mathbf{X}_{\bar{E}}^\tau(I - \mathbf{X}_{T^*}(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathbf{X}_E)\varepsilon\|_\infty \\
&\quad +\lambda\|\mathbf{X}_{\bar{E}}^\tau\mathbf{X}_{T^*}(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})\|_\infty < \lambda\Big\} \\
&\subseteq \Big\{\|\mathbf{X}_{\bar{E}}^\tau(I - \mathbf{X}_{T^*}(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathbf{X}_E^\tau)y + \lambda\mathbf{X}_{\bar{E}}^\tau\mathbf{X}_{T^*}(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})\|_\infty < \lambda\Big\} \\
&= \Big\{\|\mathbf{X}_{\bar{E}}^\tau(y - \mathbf{X}_{T^*}\tilde{\beta}_{T^*})\|_\infty < \lambda\Big\}. \qquad \text{(A.4)}
\end{aligned}$$

Thus A_n implies $\{\|\mathbf{X}_{\bar{E}}^\tau(y - \mathbf{X}_{T^*}\tilde{\beta}_T)\|_\infty < \lambda\}$.

Noting that $\mathbf{X}_E^\tau(y - \mathbf{X}\tilde{\beta}) = \mathbf{X}_E^\tau(y - \mathbf{X}_{T^*}\tilde{\beta}_{T^*}) = \lambda\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})$, (2.8) holds. Also, (2.12) implies $\|\mathbf{X}_{\bar{T}^*}^\tau\mathbf{X}\tilde{\mu}\|_\infty < 1$.

For (2.7), on the one hand,

$$\mathbf{X}_{T^*}^\tau\mathbf{X}\tilde{\mu} = \mathbf{X}_{T^*}^\tau\mathbf{X}_E\tilde{\mu}_E = \mathrm{sign}(\beta_{\mathrm{T}^*}), \qquad \text{(A.5)}$$

and on the other hand,

$$\begin{aligned}
B_n &= \Big\{|DC_{E,T^*}^{-1}Z_E| < \sqrt{n}\Big\{|\beta_{T^*}| - \frac{\lambda}{n}|DC_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})|\Big\}\Big\} \\
&\subseteq \Big\{ - DC_{E,T^*}^{-1}Z_E < \sqrt{n}\Big\{|\beta_{T^*}| - \frac{\lambda}{n}|DC_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})|\Big\}\Big\} \\
&\subseteq \Big\{ - DC_{E,T^*}^{-1}Z_E < \sqrt{n}\Big\{D\beta_{T^*} - \frac{\lambda}{n}DC_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}})\Big\}\Big\} \\
&= \Big\{D(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathbf{X}_E^\tau\varepsilon + D\beta_{T^*} > \lambda D(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathrm{sign}(\tilde{\mu}_E)\Big\} \\
&= \Big\{D(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathbf{X}_E^\tau y - \lambda D(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathrm{sign}(\tilde{\mu}_{\mathrm{E}}) > 0\Big\} \\
&\equiv \{D\tilde{\beta}_{T^*} > 0\}. \qquad \text{(A.6)}
\end{aligned}$$

Since $\{D\tilde{\beta}_{T^*} > 0\} = \{\mathrm{diag}(\mathrm{sign}(\beta_{\mathrm{T}^*}))\tilde{\beta}_{\mathrm{T}^*} > 0\} \subseteq \{\mathrm{sign}(\beta_{\mathrm{T}^*}) = \mathrm{sign}(\tilde{\beta}_{\mathrm{T}^*})\}$, together with (A.5) and (A.6), it is clear that B_n implies (2.7).

Thus, under the Irrepresentable Conditions, when A_n and B_n hold simultaneously, $(\tilde{\mu}, \tilde{\beta})$ as (A.1) and (A.2) satisfy (2.5)−(2.8).

This completes the proof.

Proof of Theorem 1. Let $\zeta = (\zeta_1, \ldots, \zeta_{p-q})^\tau = C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}}$, $\xi = (\xi_1, \ldots, \xi_q)^\tau = DC_{E,T^*}^{-1}Z_E$, $andh = (h_1, \ldots, h_n)^\tau = DC_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_E)$. By Proposition 2.1, $P(\hat{\beta}^D =_s \beta) \ge P(A_n \cap B_n)$. Thus we have

$$\begin{aligned}1 - P(\hat{\beta}^D =_s \beta) &\le 1 - P(A_n \cap B_n)\\ &= P(A_n^c \cup B_n^c)\\ &\le P(A_n^c) + P(B_n^c)\\ &\le \sum_{i=1}^{p-q} P\Big(|\zeta_i| \ge \frac{\lambda}{\sqrt{n}}\eta_i\Big) + \sum_{j=1}^{q} P\Big(|\xi_j| \ge \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j)\Big). \qquad \text{(A.7)}\end{aligned}$$

Since ε is an n-dimensional vector of i.i.d. random variables, under (a) and (b), we have

$$\begin{aligned}\zeta \ \to \ _dN\Big(0, \sigma^2\Big(&C^*_{\bar{E},\bar{T}^*}(C^*)^{-1}_{E,T^*}C^*_{E,E}(C^*)^{-1}_{T^*,E}C^*_{T^*,\bar{E}} - C^*_{\bar{E},\bar{T}^*}(C^*)^{-1}_{E,T^*}C^*_{E,\bar{E}} -\\ &C^*_{\bar{E},E}(C^*)^{-1}_{T^*,E}C^*_{T^*,\bar{E}} + C^*_{\bar{E},\bar{E}}\Big)\Big), \qquad \text{(A.8)}\\ \xi \to_d N\Big(0, \sigma^2\Big(&D(C^*)^{-1}_{E,T^*}C^*_{E,E}(C^*)^{-1}_{T^*,E}D\Big)\Big).\end{aligned}$$

It follows that all ζ_i's and ξ_j's converge in distribution to normal random variables with mean 0 and finite variances.

Assume that $\forall i, j$, and $E(\zeta_i)^2 \le t_0^2$, $E(\xi_i)^2 \le t_0^2$ for some constant $t_0 > 0$. Then for $x > 0$, it follows that the tail probability bound of the Gaussian distribution implies

$$P(\zeta_i > x) < x^{-1}e^{-x^2/2}, \quad P(\xi_j > x) < x^{-1}e^{-x^2/2} \qquad \text{(A.9)}$$

for $i = 1, \ldots, p-q, j = 1, \ldots, q$. Therefore, when $\lambda/n \to 0$, $\lambda/n^{(c+1)/2} \to \infty$ with $0 \le c < 1$, and p and q are fixed, we obtain that

$$\begin{aligned}\sum_{i=1}^{p-q} P\Big(|\zeta_i| \ge \frac{\lambda}{\sqrt{n}}\eta_i\Big) &< 2\sum_{i=1}^{p-q}\Big(\frac{\lambda}{\sqrt{n}t_0}\eta_i\Big)^{-1}\exp\Big(-\frac{1}{2}\frac{\lambda^2}{nt_0}\eta_i^2\Big)\\ &= o(e^{-n^c}), \qquad \text{(A.10)}\end{aligned}$$

$$\begin{aligned}&\sum_{j=1}^{q} P\Big(|\xi_j| \ge \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j)\Big)\\ &\quad= \sum_{j=1}^{q} P\Big(|\xi_j| \ge \sqrt{n}|\beta_j| + o(\sqrt{n}|\beta_j|)\Big)\\ &\quad\le 2\sum_{j=1}^{q}\Big(\sqrt{n}|\beta_j|(1+o(1))\Big)^{-1}\exp\Big(-\frac{1}{2}\Big(\sqrt{n}|\beta_j|(1+o(1))\Big)^2\Big)\\ &\quad= o(e^{-n^c}). \qquad \text{(A.11)}\end{aligned}$$

Combining (A.10) and (A.11) with (A.7), Theorem 1 follows immediately.

Proof of Theorem 2. Refer to T^* as the true model. Consider the event $\mathcal{C}_1 = \{\exists\ \lambda\ \ s.t.\ \hat{\beta}^D(\lambda) =_s \beta\}$. The duality of the Dantzig selector implies that there exists $\bar{\mu} \in \mathbb{R}^p$ with $\{j : \bar{\mu}_j \neq 0\} \triangleq E$ satisfied (2.5)$-$(2.8). Then on $\mathcal{C}_1$, we have

$$\mathbf{X}_{T^*}^{\tau}\mathbf{X}_E\bar{\mu}_E = \text{sign}(\hat{\beta}_{T^*}^D) = \text{sign}(\beta_{T^*}). \tag{A.12}$$

By (2.8), $\mathbf{X}_E^{\tau}\mathbf{X}_{T^*}\hat{\beta}_{T^*}^D = \mathbf{X}_E^{\tau}y - \lambda\,\text{sign}(\bar{\mu}_E)$. Since $\mathbf{X}_E'\mathbf{X}_T$ is invertible, plugging $\bar{\mu}_E$ and $\hat{\beta}_{T^*}^D$, solved by (A.12) and (2.7), into (2.5) and (2.6), respectively, and recalling that $C = (1/n)\mathbf{X}^{\tau}\mathbf{X}$, $Z_{T^*} = (1/\sqrt{n})\mathbf{X}_{T^*}^{\tau}\varepsilon$, $Z_{\bar{T}^*} = (1/\sqrt{n})\mathbf{X}_{\bar{T}^*}^{\tau}\varepsilon$, we get

$$\begin{aligned}\mathcal{C}_1 \subset \mathcal{C}_2 := \Big\{&|C_{\bar{T}^*,E}C_{T^*,E}^{-1}\text{sign}(\beta_{T^*})| < \mathbf{1},\\ &\Big|C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}} - \frac{\lambda}{\sqrt{n}}C_{\bar{E},T^*}C_{E,T^*}^{-1}\text{sign}\Big(C_{T^*,E}^{-1}\text{sign}(\beta_{T^*})\Big)\Big| < \frac{\lambda}{\sqrt{n}}\mathbf{1}\Big\}\\ &\triangleq \mathrm{H}_1 \cap \mathrm{H}_2,\end{aligned} \tag{A.13}$$

Rewrite $\mathrm{H}_2 = \{|C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}} - (\lambda/\sqrt{n})C_{\bar{E},T^*}C_{E,T^*}^{-1}\text{sign}(C_{T^*,E}^{-1}\text{sign}(\beta_{T^*}))| < (\lambda/\sqrt{n})\mathbf{1}\}$ as

$$\Big\{\frac{\lambda}{\sqrt{n}}L < C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}} < \frac{\lambda}{\sqrt{n}}R\Big\},$$

where

$$\begin{aligned}L &= C_{\bar{E},T^*}C_{E,T^*}^{-1}\text{sign}\Big(C_{T^*,E}^{-1}\text{sign}(\beta_{T^*})\Big) - \mathbf{1},\\ R &= C_{\bar{E},T^*}C_{E,T^*}^{-1}\text{sign}\Big(C_{T^*,E}^{-1}\text{sign}(\beta_{T^*})\Big) + \mathbf{1}.\end{aligned}$$

To prove the necessity of the Irrepresentable Conditions, we proceed by contradiction. If the Irrepresentable Conditions fail, then for any integer N there always exists n, $n > N$, such that at least one component of $|C_{\bar{E},T^*}C_{E,T^*}^{-1}\text{sign}(C_{T^*,E}^{-1}\text{sign}(\beta_{T^*}))|$ is no less than 1. Without loss of generality, assume it is the first component and vice versa for the less than -1 case. Then

$$(C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}})_1 \in \Big[\frac{\lambda}{\sqrt{n}}L_1, \frac{\lambda}{\sqrt{n}}R_1\Big] \subseteq [0,\infty). \tag{A.14}$$

On the other hand, by (A.8), as n increases, with a nonzero probability $(C_{\bar{E},T}C_{E,T}^{-1}Z_E - Z_{\bar{E}})_1$ is negative and the probability of C_2 does not tend to 1; hence we have

$$\liminf P(C_1) \leq \liminf P(C_2) < 1, \tag{A.15}$$

which conflicts with weak sign consistency. Therefore, the Weak Irrepresentable Conditions are necessary for weak sign consistency. This completes the proof.

Lemma A.1. *Let $\theta = (\theta_1, \dots, \theta_n)^\tau$ be a random vector with i.i.d. entries, and such that $E(\theta_1)^{2k} < \infty$ for some integer $k > 0$. Then, for constant vector α,*

$$E(\alpha^\tau \theta)^{2k} \le (2k-1)!!\|\alpha\|_2^2 E(\theta_1)^{2k}. \tag{A.16}$$

Proof of Theorem 3. By Proposition 1, $P(\hat{\beta}^D =_s \beta) \ge P(A_n \cap B_n)$. Thus we have

$$\begin{aligned} 1 - P(\hat{\beta}^D =_s \beta) &\le 1 - P(A_n \cap B_n) \\ &= P(A_n^c \cup B_n^c) \\ &\le P(A_n^c) + P(B_n^c) \\ &\le \sum_{i=1}^{p-q} P\Big(|\zeta_i^1| \ge \frac{\lambda}{\sqrt{n}}\eta\Big) + \sum_{j=1}^{q} P\Big(|\xi_j^1| \ge \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j)\Big), \end{aligned} \tag{A.17}$$

where $\zeta^1 = (\zeta_1^1, \zeta_2^1, \dots, \zeta_{p-q}^1)^\tau = C_{\bar{E},T^*}C_{E,T^*}^{-1}Z_E - Z_{\bar{E}}$, $\xi^1 = (\xi_1^1, \dots, \xi_q^1)^\tau = DC_{E,T^*}^{-1}Z_E$, and $h = (h_1, \dots, h_n)^\tau = DC_{E,T^*}^{-1}\text{sign}(\tilde{\mu}_E)$. Write $\zeta^1 = G^\tau \varepsilon$ with $G^\tau = (G_1, \dots, G_{p-q})^\tau = (1/\sqrt{n})(C_{\bar{E},T^*}C_{E,T^*}^{-1}\mathbf{X}_E^\tau - \mathbf{X}_{\bar{E}}^\tau)$. Then

$$\begin{aligned} G^\tau G &= \frac{1}{n}(C_{\bar{E},T^*}C_{E,T^*}^{-1}\mathbf{X}_E^\tau - \mathbf{X}_{\bar{E}}^\tau)(\mathbf{X}_E C_{T^*,E}^{-1}C_{T^*,\bar{E}} - \mathbf{X}_{\bar{E}}) \\ &= C_{\bar{E},T^*}C_{E,T^*}^{-1}C_{E,E}C_{T^*,E}^{-1}C_{T,\bar{E}} - C_{\bar{E},T^*}C_{E,T^*}^{-1}C_{E,\bar{E}} - C_{\bar{E},E}C_{T^*,E}^{-1}C_{T^*,\bar{E}} + C_{\bar{E},\bar{E}} \\ &= \frac{1}{n}\mathbf{X}_{\bar{E}}^\tau\Big(I - K - K^\tau + KK^\tau\Big)\mathbf{X}_{\bar{E}} \\ &= \frac{1}{n}\mathbf{X}_{\bar{E}}^\tau(I-K)(I-K)^\tau\mathbf{X}_{\bar{E}}, \end{aligned}$$

where $K = \mathbf{X}_{T^*}(\mathbf{X}_E^\tau\mathbf{X}_{T^*})^{-1}\mathbf{X}_E^\tau$.

Therefore, by (C5) and using $(X^j)^\tau(X^j)/n = 1$, we have

$$\|G_i\|_2^2 = G_i^\tau G_i = e_i^\tau G^\tau G e_i \le \tau_1 < +\infty, \tag{A.18}$$

for any $i = 1, \dots, p-q$.

Similarly, let $\xi^1 = H^\tau \varepsilon$ with $H^\tau = (H_1, H_2, \dots, H_q)^\tau = (1/\sqrt{n})DC_{E,T^*}^{-1}\mathbf{X}_E^\tau$. Then

$$H^\tau H = \frac{1}{n}DC_{E,T^*}^{-1}\mathbf{X}_E^\tau\mathbf{X}_E C_{T^*,E}^{-1}D = DC_{E,T^*}^{-1}C_{E,E}C_{T^*,E}^{-1}D.$$

By (C4) and the fact that $D^2 = I$, we have $\xi_j^1 = H_j^\tau \varepsilon$ with

$$\|H_j\|_2^2 \le \kappa_1 < +\infty \tag{A.19}$$

for any $j = 1, \dots, q$.

Given (A.18), (A.19), and $E(\varepsilon_1)^{2k} < \infty$, Lemma A.1 implies

$$E(\zeta_i^1)^{2k} < \infty,\ i = 1, \ldots, p-q,$$
$$E(\xi_j^1)^{2k} < \infty,\ j = 1, \ldots, q,$$

which gives

$$P(|\zeta_i^1| > t) = O(t^{-2k}), \quad P(|\xi_j^1| > t) = O(t^{-2k}), \tag{A.20}$$

for any $i = 1, \ldots, p-q,\ j = 1, \ldots, q$, by the Chebyshev inequality.

By making use of the first equation in (A.20) we have

$$\sum_{i=1}^{p-q} P\Big(|\zeta_i^1| \geq \frac{\lambda\eta}{\sqrt{n}}\Big) = (p-q)O\Big(\frac{n^k}{\lambda^{2k}}\Big) = O\Big(\frac{pn^k}{\lambda^{2k}}\Big). \tag{A.21}$$

Besides, (C3) has

$$\Big\|\frac{\lambda}{n}h\Big\|_\infty = \Big\|\frac{\lambda}{n}DC_{E,T^*}^{-1}\mathrm{sign}(\tilde{\mu}_\mathrm{E})\Big\|_\infty \leq \Big\|\frac{\lambda}{\mathrm{n}}\mathrm{DC}_{\mathrm{E,T^*}}^{-1}\mathrm{sign}(\tilde{\mu}_\mathrm{E})\Big\|_2 \leq \frac{\lambda\sqrt{\mathrm{q}}}{\mathrm{n}\sqrt{\mathrm{M}_2}}.$$

When $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$, (C1) yields

$$\|\frac{\lambda}{n}h\|_\infty = o(n^{(d_2-1)/2}). \tag{A.22}$$

Therefore, when $\lambda/\sqrt{n} = o(n^{(d_2-d_1)/2})$ and (C1) and (C2) both hold, using (A.22) and the second equation in (A.20), we have

$$\begin{aligned}\sum_{j=1}^{q} P\Big(|\xi_j^1| \geq \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j)\Big) &= \sum_{j=1}^{q} P\Big(|\xi_j^1| \geq \sqrt{n}|\beta_j| + o(\sqrt{n}|\beta_j|)\Big)\Big) \\ &= qO(n^{-kd_2}) = o(\frac{pn^k}{\lambda^{2k}}).\end{aligned} \tag{A.23}$$

For $p = o(n^{(d_2-d_1)k})$ and $(\lambda/\sqrt{n})^{2k}/p \to \infty$, combining (A.21) and (A.23) with (A.17), we have

$$P(\hat{\beta}^D =_s \beta) \geq 1 - O\Big(\frac{pn^k}{\lambda^{2k}}\Big). \tag{A.24}$$

Lemma A.2 (Huang, Ma, and Zhang (2008))**.** *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables with $E\varepsilon_i = 0$ and $Var(\varepsilon_i) = \sigma^2$. Further, suppose that $P(|\varepsilon_i| > x) \leq K\exp(-Cx^d)$, $i = 1, \ldots$ for positive constants C and K, and for $1 \leq d \leq 2$. Then, for all constants k_i satisfying $\sum_{i=1}^n k_i^2 = 1$, we have*

$$f_n(t) = \sup_{\sum_{i=1}^n k_i^2 = 1} P\Big\{|\sum_{i=1}^n k_i\varepsilon_i| > t\Big\} \leq \begin{cases} \exp\Big(-\frac{t^d}{M}\Big), & 1 < d \leq 2, \\ \exp\Big(-\frac{t^d}{\{M(1+\log n)\}}\Big), & d = 1, \end{cases} \tag{A.25}$$

for a positive constant M depending only on $\{d,K,C\}$.

Proof of Theorem 4. In the proof of Theorem 3, we have

$$\begin{aligned}1-P(\hat{\beta}^D =_s \beta) &\le 1-P(A_n \cap B_n)\\ &= P(A_n^c \cup B_n^c)\\ &\le P(A_n^c)+P(B_n^c). \qquad \text{(A.26)}\end{aligned}$$

By the proof of Theorem 3, $G^\tau G = (1/n)\mathbf{X}_{\bar{E}}^\tau (I-K)(I-K)^\tau \mathbf{X}_{\bar{E}}$, where $K = \mathbf{X}_{T^*}(\mathbf{X}_E^\tau \mathbf{X}_{T^*})^{-1}\mathbf{X}_E^\tau$. Then $\zeta_i^1 = G_i^\tau \varepsilon = e_i^\tau G^\tau \varepsilon$, where e_i is an unit vector with the ith component 1, others 0. By (C5), we have $\|G_i\|_2^2 = G_i^\tau G_i = e_i^\tau G^\tau G e_i \le \tau_{n1}$, $i=1,\ldots,p-q$. Hence under the conditions $p = e^{n^{d_3}}$, $(1/\log n)^{I\{d=1\}}(\lambda/\sqrt{n})^d = O(n^{d_4})$, and $0<d_3<d_4<d_2 d/2$, $1 \le d \le 2$, Lemma A.2 implies

$$\begin{aligned}P(A_n^c) &= \sum_{i=1}^{p-q} P\Big\{|\zeta_i^1| > \frac{\lambda}{\sqrt{n}}\eta\Big\}\\ &= \sum_{i=1}^{p-q} P\Big\{|\frac{1}{\sqrt{\tau_{n1}}}G_i^\tau \varepsilon| > \frac{\lambda}{\sqrt{n\tau_{n1}}}\eta\Big\}\\ &\le (p-q)P\Big\{|\frac{1}{\|G_1\|_2}G_1^\tau \varepsilon| > \frac{\lambda}{\sqrt{n\tau_1}}\eta\Big\}\\ &\le (p-q)f_n(\frac{\lambda}{\sqrt{n\tau_1}}\eta) = O(e^{n^{d_3-d_4}}). \qquad \text{(A.27)}\end{aligned}$$

Similarly, by (C4), we have $\|H_j\|_2^2 \le \kappa_{n1},\ j=1,\ldots,q$. Therefore, under (C1), (C2), (C4), and $0<d_3<\min\{d_4, d_2 d/2\}$, by Lemma A.2 again, we have

$$\begin{aligned}P(B_n^c) &= \sum_{j=1}^{q} P\Big(|\xi_j^1| \ge \sqrt{n}(|\beta_j| - \frac{\lambda}{n}h_j)\Big)\\ &= \sum_{j=1}^{q} P\Big(|H_j^\tau \varepsilon| \ge \sqrt{n}|\beta_j|(1+o(1))\Big)\\ &= \sum_{j=1}^{q} P\Big(|\frac{1}{\sqrt{\kappa_{n1}}}H_j^\tau \varepsilon| \ge \sqrt{\frac{n}{\kappa_{n1}}}|\beta_j|(1+o(1))\Big)\\ &\le qP\Big(|\frac{1}{\|H_1\|}H_1^\tau \varepsilon| \ge \sqrt{\frac{n}{\kappa_1}}|\beta_j|(1+o(1))\Big)\\ &\le qf_n\Big(\frac{M_1 n^{d_2/2}}{\sqrt{\kappa_1}}(1+o(1))\Big) = O(e^{-n^{d_2 d/2}}). \qquad \text{(A.28)}\end{aligned}$$

Combining (A.26) and (A.27) with (A.28), and invoking $\delta = \min\{d_4-d_3, d_2 d/2\}$, we obtain that

$$P(\hat{\beta}^D =_s \beta) \ge 1 - O(e^{-n^\delta}) \to 1, \text{as } n\to\infty. \qquad \text{(A.29)}$$

Proof of Theorem 5. Take $\Delta_n = \{\text{sign}(\hat{\beta}) = \text{sign}(\beta)\}$. On Δ_n, $\hat{T} = T^*$ is fixed. For any $\tau > 0$,

$$\begin{aligned} P(|\hat{\beta} - \beta| > \tau) &= P(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau) \\ &\leq P\Big(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau, \Delta_n\Big) + P(\Delta_n^c) \\ &= P\Big(|\hat{\beta}_{\hat{T}} - \beta_{T^*}| > \tau|\ \Delta_n\Big)P(\Delta_n) + P(\Delta_n^c). \end{aligned}$$

Under sign consistency, we know that $P(\Delta_n^c)$ goes to zero as n tends to infinity. Thus it is sufficient to prove $P\Big(|\hat{\beta}_{\hat{T}} - \beta_{\hat{T}}| > \tau|\Delta_n\Big)$ tends to zero. For this, it suffices to prove that, on Δ_n,

$$\Big\|C^{-1}_{T^*,T^*}\Big\{\frac{1}{n}\sum_{i=1}^{n} X_{iT^*}\varepsilon_i\Big\}\Big\|_2 = O_p\Big(\sqrt{\frac{q}{n}}\Big). \tag{A.30}$$

Since $\max\limits_{1\leq i\leq n, 1\leq j\leq q} x^2_{iT^*,j} < \infty$, we have

$$\begin{aligned} E\Big\|\Big\{\frac{1}{n}\sum_{i=1}^{n} X_{iT^*}\varepsilon_i\Big\}\Big\|_2^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\|X_{iT^*}\|_2^2 E\{\varepsilon_i^2\} \\ &= O\Big(\frac{q}{n}\Big). \end{aligned}$$

By the Markov inequality, we get that

$$\Big\|\Big(\frac{1}{n}\sum_{i=1}^{n} X_{iT^*}\varepsilon_i\Big)\Big\|_2^2 = O_p\Big(\frac{q}{n}\Big), \tag{A.31}$$

And it follows that

$$\begin{aligned} \Big\|C^{-1}_{T^*,T^*}\Big\{\frac{1}{n}\sum_{i=1}^{n} X_{iT^*}\varepsilon_i\Big\}\Big\|_2^2 &= \text{trace}\Big(\Big(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X}_{\mathrm{iT}^*}\varepsilon_{\mathrm{i}}\Big)^{\tau} \mathrm{C}^{-1}_{\mathrm{T}^*,\mathrm{T}^*}\mathrm{C}^{-1}_{\mathrm{T}^*,\mathrm{T}^*}\Big(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X}_{\mathrm{iT}^*}\varepsilon_{\mathrm{i}}\Big)\Big) \\ &= O\Big(\Big\|\Big(\frac{1}{n}\sum_{i=1}^{n} X_{iT^*}\varepsilon_i\Big)\Big\|_2^2\Big) = O_p\Big(\frac{q}{n}\Big). \end{aligned}$$

The second equality holds because of (C2). Hence, (A.30) holds.

References

Asif, M. S. and Romberg, J. (2010). On the lasso and dantzig selector equivalence. 44*th Annual Conference on Information Sciences and Systems.* 1-6.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2351.

Dicker, L. and Lin, X. H. (2009). A large sample analysis of the Dantzig selector and extensions. manuscript

Dicker, L. and Lin, X. H. (2011). Parallelism, uniqueness, and large-sample asymptotics for the Dantzig selector. manuscript.

Fan, J. (1997). Comments on 'Wavelets in Statistics: A review,' by A.Antoniadis. *J. Italian Statist. Soc.* **6**, 131-138.

Fan, J. and Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Huang, J. , Ma, S. and Zhang, C. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

James, G. M., Radchenko, P. and Lv, J. C. (2009). DASSO: connections between the Dantzig selector and lasso. *J. Roy. Statist. Soc. Ser. B* **71**, 127-142.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wainwright, M. (2006). Sharps thresholds for high-dimensional and noisy recovery of sparsity. Technical Report, Statistical Department, UC Berkley.

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7**, 2541-2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

School of Statistics, Central University of Finance and Economics, Beijing, China.

E-mail: gaiyujie83@gmail.com

Department of Mathematics, Hong Kong Baptist University, Hong Kong.

E-mail: lzhu@hkbu.edu.hk

School of Computer Science and Technology, Software Park Campus, Shandong University,
No.1500, Shunhua Road, Jinan, Shandong Province, 250101, P.R. China.

E-mail: llu@sdu.edu.cn

(Received February 2012; accepted May 2012)