Electronic Journal of Statistics
Vol. 9 (2015) 1335–1356
ISSN: 1935-7524
DOI: 10.1214/15-EJS1040

Normalized and standard Dantzig estimators: Two approaches

Jan Mielniczuk[†] **and Hubert Szymanowski**[*]

Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warsaw
e-mail: miel@ipipan.waw.pl; h.szymanowski@ipipan.waw.pl

Abstract: We reconsider the definition of the Dantzig estimator and show that, in contrast to the LASSO, standardization of an experimental matrix leads in general to a different estimator than in the case when it is based on the original data. The properties of the first method, resulting in what is called here the normalized Dantzig estimator are studied and the results on its estimation and prediction error are compared with similar results for the standard version. It is shown that in general the normalized version yields tighter estimation and prediction bounds than the other approach. In the correct specification case tighter bounds are obtained for the normalized Dantzig estimator than for the LASSO. Numerical examples indicate that in the case of imbalanced data the normalized estimator also performs better than the standard version.

MSC 2010 subject classifications: Primary 62J05, 62J07; secondary 90C25.
Keywords and phrases: Linear model, high dimensionality, Dantzig selector, LASSO, normalization, constrained optimization, Karush-Kuhn-Tucker conditions.

Received October 2013.

Contents

1 Introduction 1336
1.1 Regression model 1336
2 Preliminaries and auxiliary results 1341
3 Main results 1346
3.1 The normalized Dantzig estimator 1346
3.2 Case of correct model specification 1349
3.3 The standard Dantzig estimator 1351
3.4 The linear model with intercept 1352
3.5 Numerical examples 1353
Acknowledgment 1356
References 1356

[*]Supported by research fellowship within "Information technologies: research and their interdisciplinary applications" POKL.04.01.01-00-051/10-00.

[†]This author is also with Warsaw University of Technology, Koszykowa 75, Warsaw.

1. Introduction

1.1. Regression model

We consider a general regression model of real-valued responses having the following structure

$$y_i = \mu(x_{i.}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are iid $N(0, \sigma^2)$ and $x_{i.}$ are p-dimensional column vectors. In a vector form we have

$$y = \mu + \varepsilon, \tag{1}$$

where $\mu = (\mu(x_{1.}), \ldots, \mu(x_{n.}))^T, \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $y = (y_1, \ldots, y_n)^T$. Let $n \times p$ matrix $X = [x_{1.}, \ldots, x_{n.}]^T = [x_1, \ldots, x_p]$ be the matrix of experiment. Number of regressors $p = p_n$ may depend on n and may be larger than n. Throughout $||x||$ and $|x|$ will stand for ℓ^2 and ℓ^1 norm of a vector x. Define D as a diagonal matrix $\operatorname{diag}(||x_1||, \ldots, ||x_p||)$. In particular, when $\mu = X\beta$ for some β we obtain the linear model. However, we consider here much more general situation in which a general regression function μ is approximated by a linear model $X\beta$. Observe that the intercept is not singled out i.e. we treat $x_1, \ldots, x_p$ as genuine regressors. The case of the linear model including intercept is treated in Section 3.4.

We define standard Dantzig estimator as

$$\hat{\beta}_D = \arg\min_{\beta \in \mathbb{R}^p} \left\{ |\beta| : \left| D^{-1} X^T (Y - X\beta) \right|_\infty \le r \right\}. \tag{2}$$

Note that i^{th} coordinate of $D^{-1}X^T(Y - X\beta)$ is the Least Squares (LS) estimator of a slope when residuals $Y - X\beta$ are treated as a response and i^{th} column of XD^{-1} as a predictor. Thus we are looking for vector β having the minimal ℓ^1 norm for which the slopes of the residual regression, which should be negligible, do not exceed in absolute value a certain given threshold r. This, up to a constant term n^{-1} is exactly definition of the Dantzig estimator given in [1], compare also [4] where the estimator was introduced, in particular p. 2316 containing the remark on the case of general X. Theoretical properties of (2) are studied in [1].

However, it is not equality (2) which is usually used when the Dantzig estimator is considered and applied. Namely, in the papers discussing its properties it is common to assume from the beginning that columns of X are normalized to have norm 1 or $\sqrt{n}$ and properties of the Dantzig estimator are studied for such a case (see e.g. [7, 12]). Since such a condition is usually not met by the experimental matrix, the original matrix X has to be normalized i.e. is replaced by XD^{-1}. Consequently, in the linear case the Dantzig estimator defined in (2) for such a matrix is actually an estimator of $D\beta$ as $EY = X\beta = XD^{-1}D\beta$. Moreover, conditions on experimental design in such a case are stated for normalized matrix XD^{-1}. In order to obtain the estimator of underlying β when X is normalized a final rescaling is needed. This leads to the following quantity,

$$\hat{\beta}_N = D^{-1} \arg\min_{\beta \in V_N} |\beta|, \tag{3}$$

where

$$V_N = \{\beta : |D^{-1}X^T(Y - XD^{-1}\beta)|_\infty \le r\}, \tag{4}$$

which we call the normalized Dantzig estimator in contrast to the (standard) Dantzig estimator defined in (2). In order to appreciate the difference with regard to (2) it is worthwhile to view $\hat\beta_N$ as the vector such that

$$\hat\beta_N = \arg\min_{\beta\in\mathbb{R}^p} \big\{|D\beta| : \big|D^{-1}X^T(Y - X\beta)\big|_\infty \le r\big\},$$

i.e. $|D\beta|$ and not $|\beta|$ is minimized over $V_D = \{|D^{-1}X^T(Y - X\beta)\big|_\infty \le r\}$. Note also that in view of the definitions of the feasible sets V_D and V_N we have that $D\hat\beta_N, D\hat\beta_D \in V_N$. Moreover, $\hat\beta_N, \hat\beta_D \in V_D$. When design matrix X is orthonormal it follows from the definition of $\hat\beta_N$ and $\hat\beta_D$ that both estimators coincide with soft-thresholded values of components of X^TY, namely

$$\hat\beta_{N,i} = \hat\beta_{D,i} = \max(|(X^TY)_i| - r, 0)\operatorname{sgn}(X^TY)_i.$$

When X is orthogonal both estimators still coincide and their i^{th} components are equal to $D^{-2}(X^TY)_i$ soft-thresholded by $r/||x_i||$.

Note that the constraints for the feasible set V_D can be written as

$$\begin{aligned} g_i(\beta) &= I\{(XD^{-1})^T(Y - X\beta)_i \ge 0\}((XD^{-1})^T(Y - X\beta) - r)_i \\ &+ I\{(XD^{-1})^T(Y - X\beta)_i < 0\}(-(XD^{-1})^T(Y - X\beta) - r)_i) \le 0 \end{aligned}$$

for $i = 1, \dots, p$. Thus Karush-Kuhn-Tucker condition for $\hat\beta_D$ states that for some vector u with all nonnegative components $u_i \ge 0$

$$0 \in \frac{\partial}{\partial\beta}\{|\beta|\}_{|\beta=\hat\beta_D} - D^{-1}X^TXP_{\hat\beta_D}u, \tag{5}$$

where ∂/∂ denotes subderivative and P_β is diagonal matrix with i^{th} diagonal element equal to 1 or -1 depending on whether $(XD^{-1})^T(Y - X\beta)_i \ge 0$ or reversely. The corresponding condition for $\hat\beta_N$ is

$$0 \in \frac{\partial}{\partial\beta}\{|\beta|\}_{|\beta=\hat\beta_N} - D^{-1}X^TXD^{-1}P_{\hat\beta_N}u, \quad u \ge 0. \tag{6}$$

In the following example we point out that these two versions of the Dantzig estimator do *not* coincide and actually the difference between the Dantzig estimator $\hat\beta_D$ and the normalized Dantzig estimator $\hat\beta_N$ can be arbitrarily large.

Motivating example. Set $h \in (0, \sqrt{2}]$ and let

$$X = \frac{\sqrt{2}}{2h}\begin{bmatrix} h & 1 - h^2 - h\sqrt{2 - h^2} \\ h & 1 - h^2 + h\sqrt{2 - h^2} \end{bmatrix},$$

$$Y = \frac{\sqrt{2(2 - h^2)}}{2h^2}\begin{bmatrix} \sqrt{2 - h^2} - h \\ \sqrt{2 - h^2} + h \end{bmatrix}.$$

Note that the norms of columns of matrix X are equal respectively 1 and $1/h$. Therefore normalizing matrix $D = \mathrm{diag}(1, h^{-1})$. Moreover,

$$X^T X = \begin{bmatrix} 1 & \frac{1-h^2}{h} \\ \frac{1-h^2}{h} & \frac{1}{h^2} \end{bmatrix},$$

$$D^{-1}X^T X = \begin{bmatrix} 1 & \frac{1-h^2}{h} \\ 1-h^2 & \frac{1}{h} \end{bmatrix},$$

$$D^{-1}X^T X D^{-1} = \begin{bmatrix} 1 & 1-h^2 \\ 1-h^2 & 1 \end{bmatrix},$$

$$D^{-1}X^T Y = \begin{bmatrix} \frac{2-h^2}{h^2} \\ \frac{2-h^2}{h^2} \end{bmatrix}.$$

Note that every matrix X and every vector Y such that above equalities hold true yields the same form of $\hat{\beta}_D$ and $\hat{\beta}_N$.

Assume that $r = 1$. Computing Dantzig estimator $\hat{\beta}_D$ is equivalent to finding the minimum of function $|\beta_1| + |\beta_2|$ with restrictions

$$\begin{cases} |\frac{2-h^2}{h^2} - \beta_1 - \frac{1-h^2}{h}\beta_2| & \leq & 1 \\ |\frac{2-h^2}{h^2} - (1-h^2)\beta_1 - \frac{1}{h}\beta_2| & \leq & 1. \end{cases}$$

We show that the Dantzig estimator equals

$$\hat{\beta}_D = \begin{cases} \left(0, \frac{2}{h}\right)^T & \text{for} \quad h \in (0, \frac{\sqrt{5}-1}{2}) \\ \left(\frac{2(h^2-1)}{h^2(h^2-2)}, \frac{2(h^2-1)}{h(h^2-2)}\right)^T & \text{for} \quad h \in (\frac{\sqrt{5}-1}{2}, 1) \\ (0,0)^T & \text{for} \quad h \in [1, \sqrt{2}]. \end{cases}$$

The case when $h \in [1, \sqrt{2}]$ is obvious, we give the detailed proof for the case $h \in (0, (\sqrt{5}-1)/2)$.

Set $h \in (0,1)$ and $\hat{\beta} = (0, 2h^{-1})^T$. Note that $D^{-1}X^T(Y - X\hat{\beta}) = (1, -1)^T$ and thus $g_i(\hat{\beta}) = 0$ for $i = 1, 2$. Karush-Kuhn-Tucker conditions (5) have the following form

$$\begin{cases} u_1 - (1-h^2)u_2 & \in & [-1, 1] \\ \frac{1-h^2}{h}u_1 - \frac{u_2}{h} & = & 1 \\ u_1 & \geq & 0 \\ u_2 & \geq & 0. \end{cases}$$

This yields that

$$\begin{cases} u_1 & \in & [\frac{h}{1-h^2}, \frac{1-(1-h^2)h}{1-(1-h^2)^2}] \\ u_2 & = & (1-h^2)u_1 - h. \end{cases}$$

Nonnegative u_1, u_2 satisfying the inequalities above exist if

$$\frac{h}{1-h^2} \leq \frac{1-(1-h^2)h}{1-(1-h^2)^2}$$

which is satisfied for $h \in (0, (\sqrt{5}-1)/2)$. Thus for such h we have $\hat{\beta}_D = \hat{\beta}$.

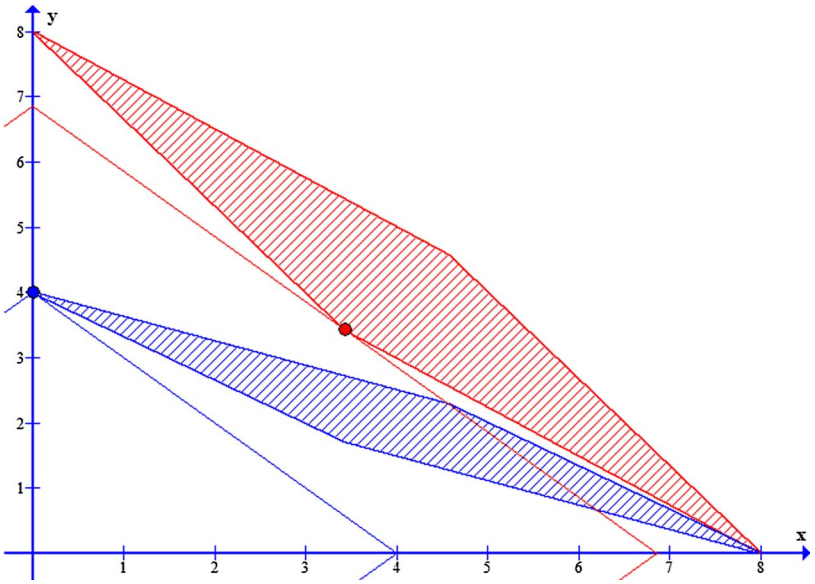

FIG 1. *Feasible regions V_D (blue) and V_N (red) satisfying Dantzig constraints for $h = 0.5$. Dantzig estimators lying on surfaces of ℓ^1 balls are represented by dots of corresponding color.*

Note that if $h = (\sqrt{5}-1)/2$ the Dantzig estimator is not uniquely defined.

Similarly, to find the normalized Dantzig estimator we need to minimize $|\beta_1| + |\beta_2|$ under restrictions

$$\begin{cases} |\frac{2-h^2}{h^2} - \beta_1 - (1-h^2)\beta_2| & \leq & 1 \\ |\frac{2-h^2}{h^2} - (1-h^2)\beta_1 - \beta_2| & \leq & 1. \end{cases}$$

Reasoning analogously as before we find that

$$D\hat{\beta}_N = \begin{cases} \left(\frac{2(h^2-1)}{h^2(h^2-2)}, \frac{2(h^2-1)}{h^2(h^2-2)}\right)^T & \text{for} \quad h \in (0,1) \\ (0,0)^T & \text{for} \quad h \in [1, \sqrt{2}]. \end{cases}$$

Thus the normalized Dantzig estimator equals

$$\hat{\beta}_N = \begin{cases} \left(\frac{2(h^2-1)}{h^2(h^2-2)}, \frac{2(h^2-1)}{h(h^2-2)}\right)^T & \text{for} \quad h \in (0,1) \\ (0,0)^T & \text{for} \quad h \in [1, \sqrt{2}] \end{cases}$$

and coincides with $\hat{\beta}_D$ for $h \in ((\sqrt{5}-1)/2, 1)$. For $h < (\sqrt{5}-1)/2$

$$|\hat{\beta}_D - \hat{\beta}_N|_1 = \left|\frac{2(h^2-1)}{h^2(h^2-2)}\right| + \left|\frac{2}{h} - \frac{2(h^2-1)}{h(h^2-2)}\right| = \frac{h^3-2h^2+2}{h^2(2-h^2)} \overset{h\to 0}{\longrightarrow} +\infty.$$

Whence ℓ^1 distance between estimators $\hat{\beta}_D$ and $\hat{\beta}_N$ can be arbitrarily large.

We note, however, that for $p = 2$ an analysis similar to that given above yields that if $\rho < \min(||x_1||/||x_2||, ||x_2||/||x_1||)$, where $\rho = x_1^T x_2/(||x_1||||x_2||)$ then $\hat{\beta}_N = \hat{\beta}_D$.

For the properties of Dantzig estimators in different settings we refer to [4, 3, 6, 5]. We also define the closely related LASSO estimator $\hat{\beta}_L = (\hat{\beta}_{L,1}, \ldots, \hat{\beta}_{L,p})$ (cf. [10])

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^p} \{||Y - X\beta||^2 + 2r \sum_{j=1}^{p} ||x_j|||\beta_j|\}, \tag{7}$$

where $r > 0$ is the tuning constant. For the recent book treatment of properties of the LASSO see [2]. We note that $\hat{\beta}_L \in V_D$ which follows from Karush-Kuhn-Tucker conditions (c.f. proof of Lemma 1) and consequently $D\hat{\beta}_L \in V_N$.

Note that in contrast to the Dantzig estimator, calculation of the LASSO for the normalized matrix X and then rescaling it yields exactly the same estimator as based on the original data.

Sufficient conditions under which the LASSO coincides with Dantzig selector for standardized X are given in [5] and [8], p. 2377. It is assumed in [8] that matrix X is normalized. Reconsidering the proof of Theorem 1 there without this assumption leads to following result. If matrix $M = D(X^T X)^{-1} D$ is diagonally dominant i.e. $M_{j,j} > \sum_{i \neq j} |M_{i,j}|$ for all $j = 1, \ldots, p$ then

$$D\hat{\beta}_N = \arg\min_{\beta \in \mathbb{R}^p} \{||Y - XD^{-1}\beta||^2 + 2r \sum_{j=1}^{p} |\beta_j|\}.$$

Since the LASSO estimator is invariant with respect to rescaling of columns of X, the RHS of the equality above equals $D\hat{\beta}_L$ and thus $\hat{\beta}_N = \hat{\beta}_L$. Analogous reasoning involving $D^{-1}X^T X$ instead of $X^T X$ yields a corresponding result for $\hat{\beta}_D$. Namely, under diagonal dominance of matrix $(X^T X)^{-1} D = D^{-1} M$ we have $\hat{\beta}_D = \hat{\beta}_L$. Whence estimators $\hat{\beta}_D$ and $\hat{\beta}_N$ coincide if both matrices M and $D^{-1}M$ are diagonally dominant.

Remark 1. In the case of equal pairwise correlations $x_i^T x_j/(||x_i||||x_j||) = \rho$ for $i \neq j$ we have $\hat{\beta}_D = \hat{\beta}_N$ provided $\rho > 1/(3 - 2p)$ and

$$x_{max} \sum_{i=1}^{p} ||x_i||^{-1} < \frac{(p-2)\rho + 1}{|\rho|} + 1,$$

where x_{max} is the maximal ℓ^2-norm of the columns of X. We verify condition of diagonal dominance of matrices M and $D^{-1}M$. In the equicorrelation case

$$M = \frac{1}{1-\rho}\big(I - \frac{\rho}{1 + (p-1)\rho} \mathbb{1}^T \mathbb{1}\big),$$

where I is the identity matrix and $\mathbb{1}$ the column of ones. It is easy to check that M is diagonally dominant for $\rho > (3 - 2p)^{-1}$. Now consider matrix $D^{-1}M$. We have

$$\sum_{i \neq j} |(D^{-1}M)_{i,j}| = \frac{|\rho|}{(1-\rho)((p-1)\rho + 1)} \sum_{i \neq j} ||x_i||^{-1}$$

and

$$(D^{-1}M)_{j,j} = \frac{(p-2)\rho + 1}{(1-\rho)((p-1)\rho + 1)||x_j||}.$$

Inequality $(D^{-1}M)_{j,j} > \sum_{i \neq j} |(D^{-1}M)_{i,j}|$ is satisfied if and only if

$$||x_j|| \sum_{i=1}^{p} \frac{1}{||x_i||} < \frac{(p-2)\rho + 1}{|\rho|} + 1.$$

Since this conditions has to be satisfied for each $j = 1, 2, \ldots, p$ it is enough to check it for j such that $||x_j|| = x_{max}$.

The aim of the paper is to study properties of $\hat{\beta}_N$ under assumptions imposed on X and compare them with properties of $\hat{\beta}_D$ and $\hat{\beta}_L$. We improve the results stated in [1] for $\hat{\beta}_D$ and we prove better bounds for $\hat{\beta}_N$ under more general conditions. The results specify upper bounds for prediction and estimation errors of $\hat{\beta}_N$ which are tighter than upper bounds for analogous quantities in the case of $\hat{\beta}_D$ (c.f. Theorems 5 and 8). We do not provide theoretical results which directly compare errors of both estimators although numerical experiments discussed in Section 3.5 indicate that in the case of imbalanced design plans when the norms of columns differ significantly $\hat{\beta}_N$ indeed performs better than $\hat{\beta}_D$.

Note that in the paper we deal with a particular normalization of columns, namely division by its norms so that the ℓ^2 norms of the transformed columns are equal to 1. In the case when distributions of columns exhibit e.g. pronounced skewness, different transformations such that as division by the columns' maximum absolute values might be preferable.

The paper is organized as follows. In Section 2 we state some auxiliary results on the considered estimators, in particular Lemma 3 yields a new bound on the size of the support of $\hat{\beta}_N$ and $\hat{\beta}_D$. Section 3 contains the main results, a brief discussion of the linear model with an intercept and numerical examples.

2. Preliminaries and auxiliary results

Let

$$\delta_{DL} = \hat{\beta}_D - \hat{\beta}_L \qquad \text{and} \qquad \tilde{\delta}_{DL} = D(\hat{\beta}_D - \hat{\beta}_L),$$
$$\delta_{NL} = \hat{\beta}_N - \hat{\beta}_L \qquad \text{and} \qquad \tilde{\delta}_{NL} = D(\hat{\beta}_N - \hat{\beta}_L),$$

with quantities δ_{DN} and $\tilde{\delta}_{DN}$ defined accordingly. We note that $\tilde{\delta}$ will always denote $D\delta$. Moreover, let $J_L = \{i : \hat{\beta}_{L,i} \neq 0\}$ and J_N and J_D defined analogously. $\bar{J}$ stands for $\{1, \ldots, p\} \setminus J$ and β_J for β restricted to J. Throughout the paper $X_0 = XD^{-1}$ will stand for the normalized matrix of experiment.

We start with a simple lemma, which shows the interplay between $\hat{\beta}_L, \hat{\beta}_D$ and $\hat{\beta}_N$.

Lemma 1. *We have (i)*

$$|\hat{\beta}_D| \leq |\hat{\beta}_L| \wedge |\hat{\beta}_N|, \quad |D\hat{\beta}_N| \leq |D\hat{\beta}_D| \wedge |D\hat{\beta}_L|$$

(ii)

$$\begin{aligned} |(\delta_{DL})_{\bar{J}_L}| &\leq |(\delta_{DL})_{J_L}|, \quad |(\delta_{DN})_{\bar{J}_N}| \leq |(\delta_{DN})_{J_N}| \\ |(\tilde{\delta}_{NL})_{\bar{J}_L}| &\leq |(\tilde{\delta}_{NL})_{J_L}|, \quad |(\tilde{\delta}_{DN})_{\bar{J}_D}| \leq |(\tilde{\delta}_{DN})_{J_D}| \end{aligned}$$

(iii)

$$|\hat{\beta}_N| \leq \frac{x_{max}}{x_{min}}|\hat{\beta}_D|, \quad |(\tilde{\delta}_{DL})_{\bar{J}_L}| \leq \frac{x_{max}}{x_{min}}|(\tilde{\delta}_{DL})_{J_L}|,$$

where x_{max} is the maximal and x_{min} minimal ℓ^2-norm of the columns of X.
(iv)

$$||Y - X\hat{\beta}_L|| \leq ||Y - X\hat{\beta}_N||.$$

Proof. The first inequality in (i) follows from the fact that $\hat{\beta}_L$ and $\hat{\beta}_N$ satisfy Dantzig constraint i.e. belong to V_D whereas second inequality follows from $D\hat{\beta}_D$, $D\hat{\beta}_L \in V_N$. The fact that $\hat{\beta}_L \in V_D$ follows from the fact that 0 has to belong to the subderivative of the right-hand side of (7) which implies that

$$|X^T(Y - X\hat{\beta}_L)|_j \leq r||x_j||$$

for $j = 1, \ldots, p$. This is equivalent to $\hat{\beta}_L \in V_D$. First inequality in (ii) follows from (i) and inequalities

$$|(\delta_{DL})_{\bar{J}_L}| = |(\hat{\beta}_D)_{\bar{J}_L}| \leq |(\hat{\beta}_L)_{J_L}| - |(\hat{\beta}_D)_{J_L}| \leq |(\hat{\beta}_L - \hat{\beta}_D)_{J_L}|.$$

The remaining ones are proved analogously.

Proof of the second inequality in (iii) follows from (ii) and the observation that

$$|(\tilde{\delta}_{DL})_{\bar{J}_L}| \leq x_{max}|(\hat{\beta}_L - \hat{\beta}_D)_{\bar{J}_L}| \leq x_{max}|(\hat{\beta}_L - \hat{\beta}_D)_{J_L}| \leq \frac{x_{max}}{x_{min}}|(\tilde{\delta}_{DL})_{J_L}|.$$

Inequality in (iv) follows from the definition of the LASSO and the second inequality in (i).

We define now the following restricted eigenvalue coefficient $\kappa(s, c)$

$$\kappa(s, c) = \min_{\substack{J \subset \{1,2,\ldots,p\} \\ |J| \leq s}} \min_{\substack{\delta \neq 0 \\ |\delta_{\bar{J}}| \leq c|\delta_J|}} \frac{||X_0\delta||}{||\delta_J||}. \tag{8}$$

This is modified version of $\tilde{\kappa}(s, c)$ introduced in [1] which differs from the original definition in that normalized matrix $X_0 = XD^{-1}$ is used instead of X and the constant $n^{-1/2}$ is omitted. We believe that the introduced modification is more convenient when dealing with the normalized Dantzig estimator, see e.g. proof of Theorem 2 below. Other measures used in sparse model selection are discussed e.g. in [11].

Note that obviously

$$\frac{||X_0\tilde{\delta}||}{||\tilde{\delta}_J||} \leq \frac{||X\delta||}{x_{min}||\delta_J||}$$

for $\delta = D^{-1}\tilde{\delta}$. However, as the condition $|\tilde{\delta}_{\bar{J}}| \leq c|\tilde{\delta}_J|$ is *not* equivalent to $|\delta_{\bar{J}}| \leq c(x_{max}/x_{min})|\delta_J|$, the inequality $\kappa(s, c) \leq \tilde{\kappa}(s, c(x_{max}/x_{min}))n^{1/2}x_{min}^{-1}$ does not necessarily hold in general as is seen from the following example.

Example 1. (i) Let $X^TX = diag(4,1)$ i.e. $x_{min} = 1$ and $x_{max} = 2$. Then it can be checked that

$$\begin{aligned}\kappa(1,1/8) &= \frac{57}{64} > n^{1/2}\tilde{\kappa}(1, \frac{x_{max}}{8x_{min}})/x_{min} = n^{1/2}\tilde{\kappa}(1,1/4) = \frac{3}{4}\\ \kappa(1,1) &= \frac{3}{4} = n^{1/2}\tilde{\kappa}(1, \frac{x_{max}}{x_{min}})/x_{min} = n^{1/2}\tilde{\kappa}(1,2)\\ \kappa(2,1) &= \frac{1}{2} < n^{1/2}\tilde{\kappa}(1, \frac{x_{max}}{x_{min}})/x_{min} = n^{1/2}\tilde{\kappa}(2,2) = (5-\sqrt{13})/2 \approx 0.7.\end{aligned}$$

(ii) We also note that the ratio $\kappa^2(s,3)/\kappa^2(s,1)$ can be arbitrarily small. This will be relevant in Remark 4 below. Let e.g. X^TX is 3×3 equicorrelated matrix with 1s on the diagonal and ρ otherwise, where $-1/2 < \rho < 0$. Then it can be checked that $\kappa^2(1,3) = 2\rho + 1$ whereas $\kappa^2(1,1) = 5\rho/3 + 1$ and thus $\lim_{\rho\to -1/2}\kappa^2(s,3)/\kappa^2(s,1) = 0$.

Observe also that

$$\kappa^2(s,c) \leq \min_{\substack{J\subset\{1,2,\ldots,p\}\\ |J|\leq s}} \min_{\substack{\delta\neq 0\\ \text{supp }\delta\subseteq J}} \frac{\delta^T X_0^T X_0 \delta}{\delta^T\delta} \leq 1,$$

since for any δ such that $\operatorname{supp}\delta \subseteq J$ we obviously have $|\delta_{\bar{J}}| \leq c|\delta_J|$ for any $c > 0$ and the minimal eigenvalue of of $X_{0J}^T X_{0J}$ does not exceed 1. Thus the last inequality follows from Rayleigh-Ritz theorem.

Positiveness of $\kappa(s,c)$, which due to restrictions on vectors δ over which minimization is performed can hold even for $p > n$, is a condition on a weak correlation of columns. We note however that it follows from analogous reasoning to that in [1], p. 1710 that $\kappa(s,c) > 0$ implies that any $2s$ columns of X are necessarily linearly independent. In the bounds we discuss in the following it is tacitly assumed that the value of κ appearing there is positive (i.e. restricted eigenvalue condition is satisfied), otherwise the bounds are trivially satisfied as κ appears in the denominators of the upper bounds.

We will rely on the following prediction error bound for the LASSO estimator proved in [9]. For any $\beta \in \mathbb{R}^p$ let $J(\beta) = \{i : \beta_i \neq 0\}$, $\mathcal{M}(\beta)$ its cardinality and $\mu_\beta = X\beta$. Consider the LASSO estimator $\hat{\mu}_L = X\hat{\beta}_L$ defined by (7) with $r = r_n = A\sigma\sqrt{\log p}$. Let $\mathcal{A} = \{|D^{-1}X^T\varepsilon|_\infty \leq \frac{r}{2}\}$.

Theorem 1. *Let ε_i be independent $\mathcal{N}(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \geq 1$, $p \geq 2$, $1 \leq s \leq p$ and $A > 2\sqrt{2}$. Then $P(\mathcal{A}) \geq 1 - p^{1-A^2/8}$ and we have on $\mathcal{A}$*

$$||\hat{\mu}_L - \mu|| \leq \inf_{\beta\in\mathbb{R}^p:\mathcal{M}(\beta)\leq s} \big(||\mu_\beta - \mu|| + C(\beta)\big),$$

with $C(\beta) = 3r\sqrt{\mathcal{M}(\beta)}/\kappa(\mathcal{M}(\beta),3)$.

Proof. Fix an arbitrary $\beta \in \mathbb{R}^p$ with $\mathcal{M}(\beta) \leq s$. Set $\tilde{\delta} = D(\hat{\beta}_L - \beta)$, $J_0 = J(\beta)$ and $\kappa = \kappa(\mathcal{M}(\beta),3)$. It is easily seen by invoking normality of errors (compare (B.4) in [1]) that $P(\mathcal{A}) \geq 1 - p^{1-A^2/8}$. Moreover, from Lemma B.1 there we

have on event $\mathcal{A}$

$$||\hat{\mu}_L - \mu||^2 \leq ||\mu_\beta - \mu||^2 + 3r\sqrt{\mathcal{M}(\beta)}||\tilde{\delta}_{J_0}||$$

which is equivalent to

$$\Big(||\hat{\mu}_L - \mu|| - ||\mu_\beta - \mu||\Big)\Big(||\hat{\mu}_L - \mu|| + ||\hat{\mu}_\beta - \mu||\Big) \leq 3r\sqrt{\mathcal{M}(\beta)}||\tilde{\delta}_{J_0}||.$$

When $3|\tilde{\delta}_{J_0}| \geq |\tilde{\delta}_{\bar{J}_0}|$ it follows from the definition of $\kappa = \kappa(\mathcal{M}(\beta), 3)$ that

$$||\tilde{\delta}_{J_0}|| \leq \frac{||XD^{-1}\tilde{\delta}||}{\kappa} = \frac{||\hat{\mu}_L - \mu_\beta||}{\kappa}.$$

Therefore, in such a case we have

$$\begin{aligned}\Big(\quad ||\hat{\mu}_L - \mu|| - ||\mu_\beta - \mu||\Big)\Big(||\hat{\mu}_L - \mu|| + ||\mu_\beta - \mu||\Big) &\leq \frac{3r\sqrt{\mathcal{M}(\beta)}}{\kappa}||\hat{\mu}_L - \mu_\beta|| \\ \leq \quad \frac{3r\sqrt{\mathcal{M}(\beta)}}{\kappa}\Big(||\hat{\mu}_L - \mu|| + ||\mu_\beta - \mu||\Big),&\end{aligned}$$

which yields the conclusion. When $3|\tilde{\delta}_{J_0}| \leq |\tilde{\delta}_{\bar{J}_0}|$, or equivalently, $4|\tilde{\delta}_{J_0}| \leq |\tilde{\delta}|$, the conclusion trivially follows from Lemma B.1 in [1] stating that

$$||\hat{\mu}_L - \mu||^2 + r|\tilde{\delta}| \leq ||\mu_\beta - \mu||^2 + 4r|\tilde{\delta}_{J_0}|. \tag{9}$$

We also state an analogue of Theorem 6.1 in [1], which will be used to prove our Theorem 5. Its proof relies on the crucial inequality (9), an analogue of which for the Dantzig estimator is not known. This is the reason why a similar result for the Dantzig estimator $\hat{\beta}_N$ is proved by comparing its prediction error to the prediction error of the Lasso estimator.

Theorem 2. *Under assumptions of Theorem 1 we have with probability at least* $1 - p^{1-A^2/8}$ *for any* $\eta > 0$ *that*

$$||\hat{\mu}_L - \mu||^2 \leq (1+\eta) \inf_{\beta \in \mathbb{R}^p : \mathcal{M}(\beta) \leq s} \Big(||\mu_\beta - \mu||^2 + \frac{C(\eta)\mathcal{M}(\beta)r^2}{\kappa^2(s, 3+4/\eta)} \Big), \tag{10}$$

where $C(\eta) = 2(2+\eta)/(1+\eta)$.

Proof follows the proof of Theorem 6.1 in [1] with one important change. Namely, the second display on p. 1728 there is now replaced by

$$\kappa^2||\tilde{\delta}_{J_0}||^2 \leq ||X_0\tilde{\delta}||^2 = ||\hat{\mu}_L - \mu||^2$$

with $\tilde{\delta} = D(\hat{\beta}_L - \beta)$ and $J_0 = J(\hat{\beta}_L)$, which follows from definition of κ. Recall that $\mathcal{A}$ is defined before Theorem 1.

Lemma 2. *(i) Let β be a vector satisfying Dantzig constraint V_D for $r = A\sigma\sqrt{\log p}$ with $A > 2\sqrt{2}$. Then on $\mathcal{A}$ we have*

$$|X_0^T(\mu - X\beta)|_\infty \leq \frac{3r}{2}. \tag{11}$$

(ii) If β satisfies normalized Dantzig constraint V_N then on $\mathcal{A}$ we have

$$|X_0^T(\mu - X_0\beta)|_\infty \leq \frac{3r}{2}. \tag{12}$$

Proof. On $\mathcal{A}$ we have

$$\begin{aligned} |X_0^T(\mu - X\beta)|_\infty &= |X_0^T(Y - \varepsilon - X\beta)|_\infty \\ &\leq |X_0^T(Y - X\beta)|_\infty + |X_0^T\varepsilon|_\infty \leq \frac{3r}{2}. \end{aligned}$$

The proof of part (ii) follows from (i) by noting that in this case $D^{-1}\beta \in V_D$. Observe that in particular it follows from part (ii) that if β is such that $\mu = X\beta$ then with $\tilde{\delta} = D(\hat{\beta}_N - \beta)$ on $\mathcal{A}$ it holds

$$|X_0^T X_0\tilde{\delta}|_\infty \leq \frac{3r}{2} \tag{13}$$

as $D\hat{\beta}_N \in V_N$.

Let λ_{max} be the maximal eigenvalue of $X_0^T X_0$. We show that the bound on size of $\{i : \hat{\beta}_{L,i} \neq 0\}$ established in [1] can be extended to both Dantzig estimators.

Lemma 3. *With probability at least $1 - p^{1-A^2/8}$ we have*

$$\mathcal{M}(\hat{\beta}) \leq \frac{4\lambda_{max}||\mu - X\hat{\beta}||^2}{r^2}, \tag{14}$$

where $\hat{\beta}$ is any of $\hat{\beta}_L$, $\hat{\beta}_D$ or $\hat{\beta}_N$.

Proof. In the case of $\hat{\beta}_L$ the proof follows the lines of the proof of similar result (B.3) in [1] but using the fact that $\hat{\beta}_{L,i} \neq 0$ implies $|X_0^T(Y - X\hat{\beta}_L)|_i = r$ and replacing X^TX by $X_0^TX_0$. For other estimators, it is sufficient to prove similar property, namely e.g. in case of $\hat{\beta}_D$ that

$$\mathcal{M}(\hat{\beta}_D) \leq |\{i : |X_0^T(Y - X\hat{\beta}_D)|_i = r\}|,$$

where $\mathcal{M}(\hat{\beta}_D)$ is cardinality of $J = J(\hat{\beta}_D) = \{i : \hat{\beta}_{D,i} \neq 0\}$. Let $\hat{\beta}_D^J \in \mathbb{R}^{\mathcal{M}(\hat{\beta}_D)}$ be $\hat{\beta}_D$ restricted to coordinates in J. Then

$$\hat{\beta}_D^J = \text{argmin}|\beta|, \quad \beta \in V_D^J, \tag{15}$$

where $V_D^J = \{\beta \in \mathbb{R}^{\mathcal{M}(\hat{\beta}_D)} : |D^{-1}X^T(Y - X_J\beta)|_\infty \leq r\}$. Note that this set is described by p conditions involving $\mathcal{M}(\hat{\beta}_D)$ variables. Equality (15) follows

from the observation that function $f(\beta) = |\beta|$ is minimized now over the set V_D^J which is embedded in V_D.

Consider the face of ℓ^1 ball in $\mathbb{R}^{\mathcal{M}(\hat{\beta}_D)}$ with radius $|\hat{\beta}_D^J|$ i.e. the set $\{\beta : |\beta| = |\hat{\beta}_D^J|\}$, which V_D^J touches. The face is determined by the equation $c^T\beta = |\hat{\beta}_D^J|$, where $c^T = (\mathrm{sgn}\hat{\beta}_{D,1}^J, \ldots, \mathrm{sgn}\hat{\beta}_{D,\mathcal{M}(\hat{\beta}_D)}^J)$. Thus $\hat{\beta}_D^J$ is a minimizer of $c^T\beta$ on V_D^J. Indeed, assume that there exists $\tilde{\beta} \in V_D^J$ such that $c^T\tilde{\beta} < c^T\hat{\beta}_D^J$. It follows easily that for any $\lambda \in [0,1]$ $\tilde{\beta}_\lambda = (1-\lambda)\hat{\beta}_D^J + \lambda\tilde{\beta} \in V_D^J$ and

$$c^T\tilde{\beta}_\lambda = (1-\lambda)c^T\hat{\beta}_D^J + \lambda c^T\tilde{\beta} < (1-\lambda)c^T\hat{\beta}_D^J + \lambda c^T\hat{\beta}_D^J = c^T\hat{\beta}_D^J.$$

However, for λ close to 0 we have $\mathrm{sgn}\tilde{\beta}_{\lambda,i} = \mathrm{sgn}\hat{\beta}_{D,i}^J$ for $i \in J$ what implies that $|\tilde{\beta}_\lambda| = c^T\tilde{\beta}_\lambda < c^T\hat{\beta}_D^J = |\hat{\beta}_D^J|$ contradicting (15). Thus $\hat{\beta}_D^J$ is an minimizer of $c^T\beta$ on convex polygon V_D^J and it follows that it is either vertex of V_D^J or a convex combination of some vertices at which minimal value of $c^T\beta$ is attained.

Since the set V_D^J is $\mathcal{M}(\hat{\beta}_D)$-dimensional, each vertex is an intersection of at least $\mathcal{M}(\hat{\beta}_D)$ faces of V_D^J. Thus at least $\mathcal{M}(\hat{\beta}_D)$ coordinates of vector $D^{-1}X^T(Y - X\hat{\beta}_D)$ has absolute value equal to r.

3. Main results

3.1. The normalized Dantzig estimator

Consider the normalized Dantzig estimator $\hat{\beta}_N$ defined by (3) with $r = A\sigma\sqrt{\log p}$. The first result yields a bound for approximation error of $\hat{\beta}_N$. For any β let $J_0 = J(\beta) = \{i : \beta_i \neq 0\}$ and $\Lambda_N = \{\beta : |\tilde{\delta}_{\bar{J}}| \leq |\tilde{\delta}_J|\}$, where $\tilde{\delta} = D(\beta - \beta_N)$. Note that reasoning as in proof of Lemma 1 it is easily seen that if β is such that $|D\beta| \geq |D\hat{\beta}_N|$ then $\beta \in \Lambda_N$.

Theorem 3. *Let ε_i be independent $\mathcal{N}(0,\sigma^2)$ random variables with $\sigma^2 > 0$. Fix $n \geq 1$, $p \geq 2$ and $A > 2\sqrt{2}$. Then, with probability at least $1 - p^{1-A^2/8}$, we have*

$$||\hat{\mu}_N - \mu||^2 \leq \inf_{\beta\in\Lambda_N}\Big\{||\mu_\beta - \mu||^2 + \frac{9r^2\mathcal{M}(\beta)}{\kappa^2(\mathcal{M}(\beta),1)}\Big\}.$$

Proof. Observe that

$$||\mu - X\beta||^2 = ||\mu - X\hat{\beta}_N||^2 - 2\tilde{\delta}^T X_0^T(\mu - X\hat{\beta}_N) + ||X_0\tilde{\delta}||^2.$$

Applying Lemma 2 (i) we obtain as $\hat{\beta}_N \in V_D$

$$\begin{aligned} ||\mu - X\hat{\beta}_N||^2 &= ||\mu - X\beta||^2 + 2\tilde{\delta}^T X_0^T(\mu - X\hat{\beta}_N) - ||X_0\tilde{\delta}||^2 \\ &\leq ||\mu - X\beta||^2 + 3r|\tilde{\delta}| - ||X_0\tilde{\delta}||^2. \end{aligned}$$

Taking into account that $|\tilde{\delta}_{\bar{J}_0}| \leq |\tilde{\delta}_{J_0}|$ and the definition of $\kappa(s,1)$ we have in view of the Schwarz inequality

$$|\tilde{\delta}| \leq 2|\tilde{\delta}_{J_0}| \leq 2\sqrt{\mathcal{M}(\beta)}||\tilde{\delta}_{J_0}|| \leq 2\sqrt{\mathcal{M}(\beta)}\frac{||X_0\tilde{\delta}||}{\kappa(\mathcal{M}(\beta),1)}. \tag{16}$$

This yields

$$||\mu - X\hat{\beta}_N||^2 \le ||\mu - X\beta||^2 + 6r\sqrt{\mathcal{M}(\beta)}\frac{||X_0\tilde{\delta}||}{\kappa(\mathcal{M}(\beta),1)} - ||X_0\tilde{\delta}||^2$$
$$\le ||\mu - X\beta||^2 + \frac{9r^2\mathcal{M}(\beta)}{\kappa^2(\mathcal{M}(\beta),1)}, \tag{17}$$

where the last inequality is obtained by minimization w.r.t. $||X_0\tilde{\delta}||$. As β is an arbitrary element of Λ_N this yields the result.

We have the following corollary of the result which is an analogue of Theorem 5.1 in [1] for $\hat{\beta}_N$ with smaller constants. Specifically, the bound for an analogous expression with $\hat{\mu}_N$ replaced by $\hat{\mu}_D$ is $16r^2\mathcal{M}(\hat{\beta}_L)x_{max}/\kappa^2(\mathcal{M}(\hat{\beta}_L),1)$.

Corollary 1. *Provided that assumptions of Theorem 3 are satisfied and if* $\mathcal{M}(\hat{\beta}_L) \le s$ *then, with probability at least* $1 - p^{1-A^2/8}$*, we have*

$$\big|||\hat{\mu}_N - \mu||^2 - ||\hat{\mu}_L - \mu||^2\big| \le \frac{9r^2\mathcal{M}(\hat{\beta}_L)}{\kappa^2(\mathcal{M}(\hat{\beta}_L),1)} \le \frac{9r^2 s}{\kappa^2(s,1)}.$$

We observe that (17) is satisfied for $\beta = \beta_L$ as in view of Lemma 1 (ii) $\beta_L \in \Lambda_N$. Interchanging the roles of $\hat{\mu}_L$ and $\hat{\mu}_N$ in the proof above we obtain a symmetric result.

In particular it follows from Corollary 1 and Lemma 3 that on $\mathcal{A}$ we have

$$||\hat{\mu}_N - \mu||^2 \le ||\hat{\mu}_L - \mu||^2\Big(1 + \frac{36\lambda_{max}}{\kappa^2(\mathcal{M}(\hat{\beta}_L),1)}\Big). \tag{18}$$

In Theorem 4 below we provide opposite inequality. We note that λ_{max} in (18) can be quite large especially for large p as for random matrix X it behaves roughly as $\sqrt{p/n}$. In Theorem 6 below for the linear model we provide bounds for the prediction error which depend only on the size of the true model and κ.

Remark 2. By replacing $\hat{\beta}_L$ with $\hat{\beta}_D$ in the proof above and using $|(\tilde{\delta}_{DN})_{\bar{J}_D}| \le |(\tilde{\delta}_{DN})_{J_D}|$ from Lemma 1 (ii) we can easily obtain corresponding result for the pair $\hat{\beta}_D$ and $\hat{\beta}_N$. Namely, on $\mathcal{A}$ we have

$$\Big|||\mu - \hat{\mu}_N||^2 - ||\mu - \hat{\mu}_D||^2\Big| \le \frac{9r^2\mathcal{M}(\hat{\beta}_D)}{\kappa^2(\mathcal{M}(\hat{\beta}_D),1)}.$$

Remark 3. By triangle inequality we have

$$\Big|||\mu - \hat{\mu}_L||^2 - ||\mu - \hat{\mu}_D||^2\Big| \le 9r^2\Big(\frac{\mathcal{M}(\hat{\beta}_L)}{\kappa^2(\mathcal{M}(\hat{\beta}_L),1)} + \frac{\mathcal{M}(\hat{\beta}_D)}{\kappa^2(\mathcal{M}(\hat{\beta}_D),1)}\Big),$$

see also Corollary 2 below.

We give yet another bound of $||\mu - \hat{\mu}_L||^2$ in terms of $||\mu - \hat{\mu}_N||^2$. It is a generalization and improvement of Theorem 5.2 in [1] (modulo a slight difference

between κ and $n^{1/2}\tilde{\kappa}$) as X is not assumed to be normalized and we obtain tighter bound than in (5.2) there. For the sake of comparison we state that the bound in Theorem 5.2 derived for the case of not necessarily normalized matrix is

$$\left(1+\frac{9x_{max}}{x_{min}}\right)||\mu-\hat{\mu}_N||^2+\frac{81r^2\mathcal{M}(\hat{\beta}_D)x_{max}^2}{\kappa^2 x_{min}^2}.$$

Theorem 4. *Let $\kappa=\kappa(\mathcal{M}(\hat{\beta}_N),3)$. Then*

$$||\mu-\hat{\mu}_L||^2\le||\mu-\hat{\mu}_N||^2+\frac{36r^2\mathcal{M}(\hat{\beta}_N)}{\kappa^2}\le||\hat{\mu}_N-\mu||^2\Big(1+\frac{144\lambda_{max}}{\kappa^2}\Big). \quad (19)$$

Proof. We use (9) for $\beta=\hat{\beta}_N$, $\tilde{\delta}=D(\hat{\beta}_L-\hat{\beta}_N)$ and $J_0=J(\hat{\beta}_N)$. If $|\tilde{\delta}|\ge 4|\tilde{\delta}_{J_0}|$ then it implies that

$$||\mu-X\hat{\beta}_L||^2\le||\mu-X\hat{\beta}_N||^2.$$

If the opposite condition is satisfied then we have in view of the Schwarz inequality and the definition of κ that (compare (16))

$$|\tilde{\delta}|\le\frac{4(\mathcal{M}(\hat{\beta}_N))^{1/2}||X_0\tilde{\delta}||}{\kappa}.$$

Now we reason as in the previous proof using the inequality above to obtain

$$\begin{aligned}||\mu-X\hat{\beta}_L||^2 &\le ||\mu-X\hat{\beta}_N||^2+3r|\tilde{\delta}|-||X_0\tilde{\delta}||^2\\ &\le ||\mu-X\hat{\beta}_N||^2+\frac{12r(\mathcal{M}(\hat{\beta}_N))^{1/2}||X_0\tilde{\delta}||}{\kappa}-||X_0\tilde{\delta}||^2.\end{aligned}$$

Maximization of the RHS yields the first inequality in (19) and the second follows from Lemma 3.

Thus we have from Corollary 1 and Theorem 4 that

$$||\mu-X\hat{\beta}_L||^2-L\le||\mu-X\hat{\beta}_N||^2\le||\mu-X\hat{\beta}_L||^2+U,$$

where $U=9r^2\mathcal{M}(\hat{\beta}_L)/\kappa^2(\mathcal{M}(\hat{\beta}_L),1)$ and $L=\min(U,36r^2\mathcal{M}(\hat{\beta}_N)/\kappa^2(\mathcal{M}(\hat{\beta}_N),3))$.

The following result is an analogue of oracle inequality for prediction loss of the Dantzig estimator in [1]; see Proposition 6.3 there. As it uses Corollary 1 instead of Theorem 5.1 in [1] the obtained bound is stricter.

Theorem 5. *Assume that for a certain $s\in\mathbb{N}$ and $\eta,B>0$ we have that the set*

$$\Lambda_{s,\eta,B}=\left\{\beta\in\mathbb{R}^p:\ \mathcal{M}(\beta)\le s,\ ||\mu_\beta-\mu||\le\frac{Br}{\kappa(s,3+4/\eta)}\sqrt{\mathcal{M}(\beta)}\right\}$$

is nonempty. Then with probability at least $1-p^{1-A^2/8}$ we have

$$\begin{aligned}&||\hat{\mu}_N-\mu||^2\\ &\le(1+\eta)\left(\inf_{\beta\in\mathbb{R}^p:\mathcal{M}(\beta)\le s}||\mu_\beta-\mu||^2+\frac{C(\eta)sr^2}{(1+\eta)\kappa^2(s,3+4/\eta)}+\frac{9r^2s_0}{(1+\eta)\kappa^2(s_0,1)}\right)\end{aligned} \quad (20)$$

where $s_0=4s\lambda_{max}[B+3]^2/\kappa^2(s,3+4/\eta)$ and $C(\eta)$ is defined in Theorem 2.

Proof. Note that in view of Theorem 1, Lemma 3 and assumptions we have that

$$\mathcal{M}(\hat{\beta}_L) \leq \frac{4\lambda_{max}}{r^2}(||\mu - X\bar{\beta}|| + C(\bar{\beta}))^2 \leq \frac{4\lambda_{max}\mathcal{M}(\bar{\beta})}{\kappa^2(s, 3+4/\eta)}\Big(B+3\Big)^2 \leq s_0$$

for some $\bar{\beta} \in \Lambda_{s,\eta,B}$. Thus using Corollary 1 and Theorem 2 we have

$$||\hat{\mu}_N - \mu||^2 \leq ||\hat{\mu}_L - \mu||^2 + \frac{9r^2\mathcal{M}(\hat{\beta}_L)}{\kappa^2(s_0, 1)}$$
$$\leq (1+\eta)\Bigg(\inf_{\beta\in\mathbb{R}^p:\mathcal{M}(\beta)\leq s} ||\mu_\beta - \mu||^2 + \frac{C(\eta)sr^2}{(1+\eta)\kappa^2(s, 3+4/\eta)} + \frac{9r^2 s_0}{(1+\eta)\kappa^2(s_0, 1)}\Bigg).$$

Remark 4. Observe that as Theorems 4–5 and Corollary 1 rely on relating prediction error of the normalized Dantzig estimator to that of LASSO, no clear picture emerges which one is preferable and whether it is advantageous to use $\hat{\beta}_N$ instead of $\hat{\beta}_L$ at all. In this context we refer the reader to the discussion following [4], where conflicting views are documented.

However, in the most important case of correct model specification discussed below obtaining direct bounds for estimation and prediction error of Dantzig estimator without relating them to those of LASSO is possible. Bounds stated in Theorem 6 are tighter than those for LASSO obtained in [1], Theorem 7.2, which are $16rt_n/\kappa^2(t_n, 3)$ for $|\hat{\beta}_L - \beta|_1$ and $16rt_n^2/\kappa^2(t_n, 3)$ for $||X\hat{\beta}_L - X\beta||^2$. The crucial difference between the bounds is appearance of the squared restricted eigenvalue coefficient $\kappa^2(t_n, 1)$ in the denominators of bounds for $\hat{\beta}_N$ in (21) instead of $\kappa^2(t_n, 3)$ in the case of LASSO. We have shown in Example 1 that the ratio $\kappa^2(t_n, 3)/\kappa^2(t_n, 1)$ can be arbitrarily small. The difference in bounds is related to the problem of determining, for both estimators, the best constant c for which $|(\hat{\beta} - \beta)_{\bar{T}}| \leq c|(\hat{\beta} - \beta)_T|$ holds with large probability. It is easily proved that $c \leq 1$ holds for the Dantizg estimator, however it seems that $c > 1$ is needed for LASSO (cf. e.g. discussion on p. 1364 in [11] and Lemma 11.2 there). We also note that a similar difference occurs when approximation error is considered in sup-norm, namely tighter bounds are obtained in the case of standardized X for Dantzig estimator than for LASSO (cf. Theorems 1 and 3 in [7]).

3.2. Case of correct model specification

Consider now the case when $\mu = X\beta$ and denote by T a minimal true model i.e. a model containing the minimal number of predictors such that $EY = X_T\beta$ for some β, where X_T denotes submatrix of X with columns restricted to subset T. Denote by t_n its cardinality. It is easy to see that if $\kappa(t_n, 1) > 0$ then the minimal true model is unique.

We first state the result on weighted ℓ^1 error of $\hat{\beta}_N$ and its squared ℓ^2 prediction error. The result is an improvement of Theorem 7.1 in [1] for $\hat{\beta}_N$ as constants 6 and 9 below replace 8 and 16 there, respectively. We do not assume that the columns of X are normalized.

Theorem 6. *We have with probability* $1-p^{1-A^2/8}$ *that*

$$\begin{aligned} |\tilde{\delta}| = |D(\hat{\beta}_N - \beta)| &\leq \frac{6rt_n}{\kappa^2(t_n,1)}, \\ ||X_0\tilde{\delta}||^2 = ||X(\hat{\beta}_N - \beta)||^2 &\leq \frac{9r^2t_n}{\kappa^2(t_n,1)}. \end{aligned} \tag{21}$$

Proof. The proof parallels that of Theorem 7.1 in [1]. Observe that as $D\beta \in V_N$ on $\mathcal{A}$ reasoning as in Lemma 1 we have $|\tilde{\delta}_{\bar{T}}| \leq |\tilde{\delta}_T|$. Thus in view of inequality (13) we have

$$\kappa^2(t_n,1)||\tilde{\delta}_T||^2 \leq ||X_0\tilde{\delta}||^2 \leq |X_0^TX_0\tilde{\delta}|_\infty|\tilde{\delta}| \leq \frac{3}{2}r(2|\tilde{\delta}_T|) \leq 3rt_n^{1/2}||\tilde{\delta}_T||, \tag{22}$$

from which it follows that

$$||\tilde{\delta}_T|| \leq \frac{3rt_n^{1/2}}{\kappa^2} \tag{23}$$

and thus

$$|\tilde{\delta}| \leq 2|\tilde{\delta}_T| \leq 2t_n^{1/2}||\tilde{\delta}_T|| \leq \frac{6rt_n}{\kappa^2(t_n,1)}.$$

Moreover, the bound on $||\tilde{\delta}_T||$ in (23) together with (22) yields the second required inequality.

Remark 5. Observe that for regular matrices we can expect $x_{min} \geq cn^{1/2}$ for some $c > 0$ and thus for such a case and constant p_n and t_n (21) implies that $|\hat{\beta}_N - \beta| = O(n^{-1/2})$.

Let $\hat{\theta}_N = D\hat{\beta}_N$, $\theta = D\beta$ and $\theta^*_{min} = \min_{i:\theta_i\neq 0}|\theta_i|$. The next result concerns truncated Dantzig estimator defined as follows. Let

$$S_0 = \{i : |\hat{\theta}_{i,N}| \geq a_0\}, \quad S_1 = \{i : |\hat{\theta}_{i,N}| \geq a_1\},$$

where $a_0 = 3r_n$ and $a_1 = 3r_n(|S_0| \vee 1)^{1/2}$. We call $\hat{\beta}^t_N = \hat{\beta}_N I_{S_1}$ the truncated normalized Dantzig estimator. The following result holds.

Theorem 7. *Assume that* $8r_nt_n^{1/2}\kappa^{-2}(t_n,1) \leq \theta^*_{min}$ *(θ^*_{min} condition). On $\mathcal{A}$ we then have (i)*

$$T \subseteq S_1 \qquad \text{and} \qquad |S_1| \leq t_n + \lfloor\sqrt{t_n}\kappa^{-2}\rfloor \tag{24}$$

(ii) Moreover,

$$|D(\hat{\beta}^t_N - \beta)| \leq |D(\hat{\beta}_N - \beta)|.$$

Proof. First inequality in (21) together with $2|\tilde{\delta}_{\bar{T}}| \leq |\tilde{\delta}|$ yields $|S_0 \setminus T| \leq |\tilde{\delta}_{\bar{T}}|/a_0 \leq t_n\kappa^{-2}$. Thus $|S_0| \leq t_n(1+\kappa^{-2})$ and $a_1 \leq 3r_n\sqrt{t_n(1+\kappa^{-2})}$. Using this and (23) we have $||\tilde{\delta}_T|| + a_1 \leq \theta^*_{min}$ or

$$||\tilde{\delta}_T||^2 \leq (\theta^*_{min} - a_1)^2. \tag{25}$$

Indeed, from θ^*_{min} condition, the fact that $\kappa \leq 1$ and (23) we have

$$\begin{aligned} ||\tilde{\delta}_T|| + a_1 &\leq 3r_n t_n^{1/2}\kappa^{-2} + 3r_n\sqrt{t_n(1+\kappa^{-2})} = 3r_n t_n^{1/2}\kappa^{-2}(1+\sqrt{\kappa^4+\kappa^2}) \\ &\leq 3(1+\sqrt{2})r_n t_n^{1/2}\kappa^{-2} \leq \theta^*_{min}. \end{aligned}$$

Evidently, $|T\setminus S_1|(\theta^*_{min}-a_1)^2 < ||\tilde{\delta}_T||^2$ and thus in view of (25) we have $T\subseteq S_1$ on $\mathcal{A}$. But $S_1\subseteq S_0$, thus $|S_0|\geq t_n$ and $a_1 \geq 3r_n t_n^{1/2}$. Thus using first inequality in (21) again, we have

$$|S_1\setminus T| \leq |\tilde{\delta}_{\bar{T}}|/a_1 \leq t_n^{1/2}\kappa^{-2}.$$

Whence on $\mathcal{A}$ we have $|S_1| \leq t_n + t_n^{1/2}\kappa^{-2}$.

(ii) follows from the observation that if $|\hat{\theta}_{i,N}| \geq a_1$ then $\hat{\theta}^t_{i,N} = \hat{\theta}_{i,N}$ and if $|\hat{\theta}_{i,N}| < a_1$ and $\theta_i = 0$ then $\hat{\theta}^t_{i,N} = \theta_i = 0$. The event when $\theta_i \neq 0$ and $|\hat{\theta}_{i,N}| < a_1$ can occur on $\mathcal{A}^c$ only as $T\subseteq S_1$ on $\mathcal{A}$.

Observe that if $\kappa^{-2} < t_n^{1/2}$ then the size of the support of $\hat{\beta}^t_N$ does not exceed $2t_n$.

3.3. The standard Dantzig estimator

We now discuss results for the standard Dantzig estimator analogous to those stated above and show that in general weaker bounds are obtained for this case using the same methods as for the normalized Dantzig estimator. Still, results below improve bounds in the results of [1]. We note that the analogue of Theorem 3 holds with Λ_N replaced by $\Lambda_D = \{\beta : |\delta_{\bar{J}_0}| \leq \delta_{J_0}|\}$. We start with stating an analogue of Corollary 1.

Corollary 2. *Fix $n \geq 1$, $p \geq 2$. If $\mathcal{M}(\hat{\beta}_L) \leq s$ then, with probability at least $1 - p^{1-A^2/8}$, we have*

$$\begin{aligned} \left|||\hat{\mu}_D - \mu||^2 - ||\hat{\mu}_L - \mu||^2\right| &\leq \frac{9}{4}\left(\frac{x_{max}}{x_{min}}+1\right)^2 \frac{r^2\mathcal{M}(\hat{\beta}_L)}{\kappa^2(\mathcal{M}(\hat{\beta}_L), \frac{x_{max}}{x_{min}})} \\ &\leq \frac{9}{4}\left(\frac{x_{max}}{x_{min}}+1\right)^2 \frac{r^2 s}{\kappa^2(s, \frac{x_{max}}{x_{min}})}. \end{aligned}$$

Proof. We recall that $\tilde{\delta} = \tilde{\delta}_{DL} = D(\hat{\beta}_D - \hat{\beta}_L)$ and set $J_0 = J(\hat{\beta}_L)$. Reasoning as in the proof of Theorem 3 and using Lemma 2 (i) we obtain the following inequality holding with probability at least $1 - p^{1-A^2/8}$

$$||\mu - X\hat{\beta}_D||^2 \leq ||\mu - X\hat{\beta}_L||^2 + 3r|\tilde{\delta}| - ||X_0\tilde{\delta}||^2. \tag{26}$$

However, from Lemma 1 (iii) we have a weaker inequality now

$$\begin{aligned} |\tilde{\delta}| \leq \left(\frac{x_{max}}{x_{min}}+1\right)|\delta_J| &\leq \left(\frac{x_{max}}{x_{min}}+1\right)\sqrt{\mathcal{M}(\hat{\beta}_L)}||\tilde{\delta}_{J_0}|| \\ &\leq \left(\frac{x_{max}}{x_{min}}+1\right)\frac{\sqrt{\mathcal{M}(\hat{\beta}_L)}}{\kappa(\mathcal{M}(\hat{\beta}_L), \frac{x_{max}}{x_{min}})}||X_0\tilde{\delta}|| \end{aligned}$$

and this leads as in the proof of Corollary 1 to the conclusion. Note that the first inequality above can not be improved (set e.g. $p = 2$, $\delta = (1, 3)^T$ and $J = \{1\}$).

Observe that the bounds in Corollaries 1 and 2 coincide when the columns of X are normalized, however if $D \neq I$ then Corollary 2 yields weaker bounds than those in Corollary 1.

It is easy to check that Theorem 4 can be written in exactly the same form with $\hat{\beta}_N$ replaced by $\hat{\beta}_D$. However, analogue of Theorem 5 for $\hat{\beta}_D$ which relies on Corollary 2 is weaker again.

Theorem 8. *Assume that assumptions of Theorem 5 hold. Then with probability at least* $1 - p^{1-A^2/8}$ *we have*

$$||\hat{\mu}_D - \mu||^2$$
$$\leq (1+\eta)\left(\inf_{\beta \in \mathbb{R}^p : \mathcal{M}(\beta) \leq s} ||\mu_\beta - \mu||^2 + \frac{C(\eta) s r^2}{(1+\eta)\kappa^2(s, 3+4/\eta)} + \frac{9r^2(\frac{x_{max}}{x_{min}}+1)^2 s_0}{(1+\eta)\kappa^2(s_0, \frac{x_{max}}{x_{min}})} \right)$$

We omit an easy modification of the proof. Finally in the case of correct model specification we obtain the following analogue of Theorem 6 which uses $|D(\hat{\beta}_D - \beta)_{\bar{T}}| \leq (x_{max}/x_{min})|D(\hat{\beta}_D - \beta)_T|$.

Theorem 9. *We have with probability* $1 - p^{1-A^2/8}$ *that*

$$\begin{aligned} \tilde{\delta} = |D(\hat{\beta}_D - \beta)|_1 &\leq \frac{3}{2} \frac{(1 + \frac{x_{max}}{x_{min}})^2 r t_n}{\kappa^2(t_n, \frac{x_{max}}{x_{min}})} \\ ||X_0 \tilde{\delta}||^2 = ||X(\hat{\beta}_D - \beta)||^2 &\leq \frac{9}{4} \frac{(1 + \frac{x_{max}}{x_{min}})^2 r^2 t_n}{\kappa^2(t_n, \frac{x_{max}}{x_{min}})} \end{aligned} \tag{27}$$

An analogue of Theorem 7 with more complicated constants is omitted.

We proved upper bounds for estimation and prediction errors for both $\hat{\beta}_N$ and $\hat{\beta}_D$ which are sharper in the first case. As we do not provide lower bounds and we do not know that the upper bounds are tight we can not claim that performance of $\hat{\beta}_N$ is superior. However, numerical examples below suggest that in the cases of imbalanced experimental matrix X this is indeed the case.

3.4. The linear model with intercept

Consider now the case when $\mu = \alpha + x^T\beta$ i.e. the linear model with intercept. In this case when it is desirable that the estimator of the intercept is invariant with respect to shift of the data it is necessary to slightly modify the definitions of the Dantzig and the LASSO estimators. In the case of the LASSO this corresponds to the practical LASSO, when response and predictors are centered before calculation of (7). The modified definition of both Dantzig estimators is as follows.

Let $H_0 = I - \mathbb{1}\mathbb{1}^T/n$, where I is the identity matrix and $\mathbb{1}$ the column of ones, be the centering matrix, $D = \text{diag}(||H_0 x_j||)_{j=1}^p$ and $X_0 = H_0 X D^{-1}$. Thus X_0 is experimental matrix which columns are centered and standardized.

The standard Dantzig estimator is now defined as

$$\hat{\beta}_D = \arg\min_{\beta\in\mathbb{R}^p} \big\{|\beta| : \big|X_0^T(Y - H_0X\beta)\big|_\infty \le r\big\}, \tag{28}$$

whereas the normalized Dantzig estimator is defined as follows

$$\hat{\beta}_N = D^{-1}(\arg\min_{\beta\in\mathbb{R}^p} \big\{|\beta| : \big|X_0^T(Y - X_0\beta)\big|_\infty \le r\big\}). \tag{29}$$

Finally, the LASSO estimator is defined (cf [9])

$$\begin{aligned}\hat{\beta}_L &= \arg\min_{\beta\in\mathbb{R}^p} \big\{||H_0(Y - X\beta)||^2 + 2r\sum_{j=1}^p ||H_0x_j|| \cdot |\beta_j|\big\} \\ &= D^{-1}\arg\min_{\theta\in\mathbb{R}^p} \big\{||H_0Y - X_0\theta)||^2 + 2r\sum_{j=1}^p |\theta_j|\big\},\end{aligned}$$

where substitution $\theta = D\beta$ was used in the last line.

In all three cases α is estimated as $\bar{Y} - \bar{X}\hat{\beta}$, where $\bar{X} = (H_0x_1, \ldots, H_0x_p)$. Note that H_0 is a projection on a space orthogonal to $\mathbb{1}$ and when H_0 is replaced by identity in the above definitions we obtain original definitions of $\hat{\beta}_D$, $\hat{\beta}_N$ and $\hat{\beta}_L$. All the results proved above are also true for the case when modified definitions are considered when in the definition of κ matrix X_0 defined above is used. We omit easy details noting that this is due to the property that H_0 is idempotent.

3.5. Numerical examples

We present examples showing that discrepancy between the standard and the normalized Dantzig estimator is not merely theoretical curiosity. Let experimental matrix X be 72×256 matrix such that its rows are sampled from $N(0, \Sigma)$, where $\Sigma = \mathrm{diag}(1^\alpha, 2^\alpha, \ldots, 256^\alpha)$ with $\alpha \in [0, 1]$, errors are $N(0, 8)$-distributed. Thus the variances of attributes increase from 1^α to 256^α. We consider two vectors of coefficients $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0, 1, 1, 1, 1)^T$ and $\beta_2 = (1, 1, 1, 1, 1, 1, 1, 1, 0 \ldots, 0)^T \in \mathbb{R}^{256}$. Let PSR_D and FDR_D denote the positive selection rate and the false detection rate for $\hat{\beta}_D$, where positive selection means $\hat{\beta}_{D,i} \neq 0$ when $\beta_i \neq 0$ and false detection $\hat{\beta}_{D,i} \neq 0$ when $\beta_i = 0$. Thus

$$PSR_D = \frac{|T \cap J(\hat{\beta}_D)|}{|T|}$$

and

$$FDR_D = \frac{|J(\hat{\beta}_D) \setminus T|}{|J(\hat{\beta}_D)|},$$

where T denotes the set of indices of relevant variables. PSR_N and FDR_N are defined analogously. The upper-left panel in Figure 2 shows medians of $PSR_N - PSR_D$ (solid line) and $FDR_D - FDR_N$ (dashed line) as the function of α based on 200 repetitions computed in the case of β_1. The upper-right panel shows the respective means. It is seen that with increasing α performance of the normalized

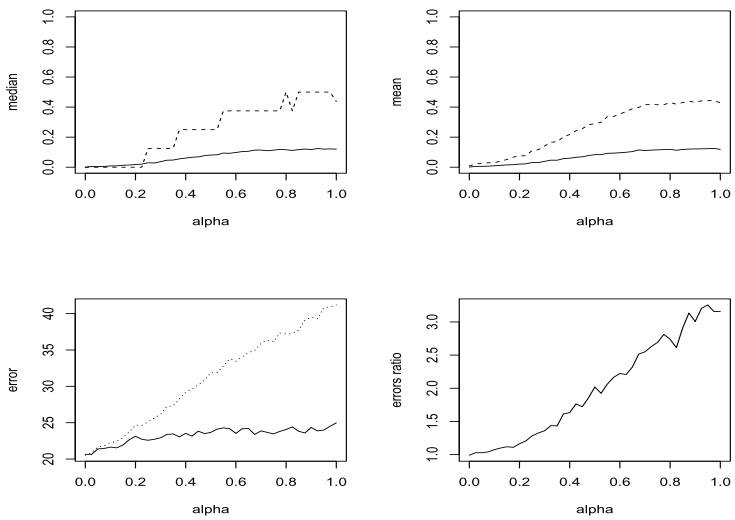

FIG 2. *Medians and means of differences between PSRs and FDRs for* $\hat{\beta}_N$ *and* $\hat{\beta}_D$ *(*β_1 *case).*

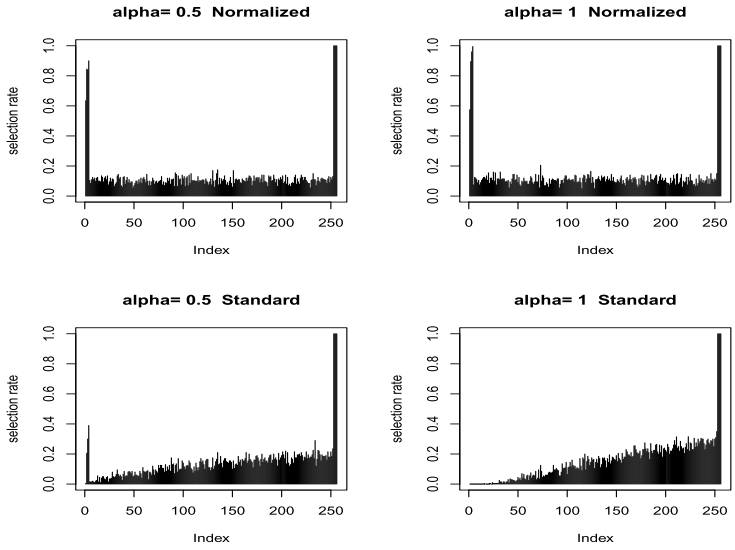

FIG 3. *Fraction of time when estimators are not equal 0 (*β_1 *case).*

Dantzig estimator becomes increasingly superior to performance of $\hat{\beta}_D$ with respect to both measures. The lower-left panel shows estimated value of the prediction error $||X_{test}(\beta - \hat{\beta})||$ for $\beta = \beta_1$, where X_{test} is an independent copy of X. Solid line corresponds to standard estimator and dotted line corresponds to the normalized estimator. The lower-right panel shows estimated value of ratio

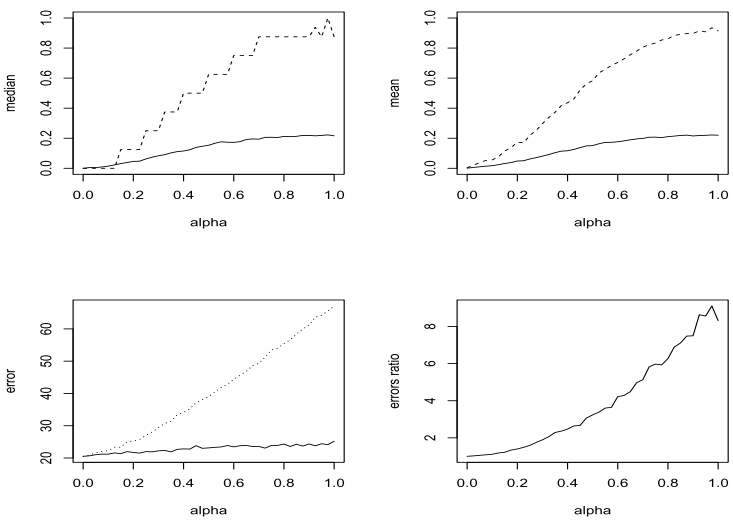

FIG 4. *Medians and means of differences between PSRs and FDRs for $\hat{\beta}_N$ and $\hat{\beta}_D$ (β_2 case).*

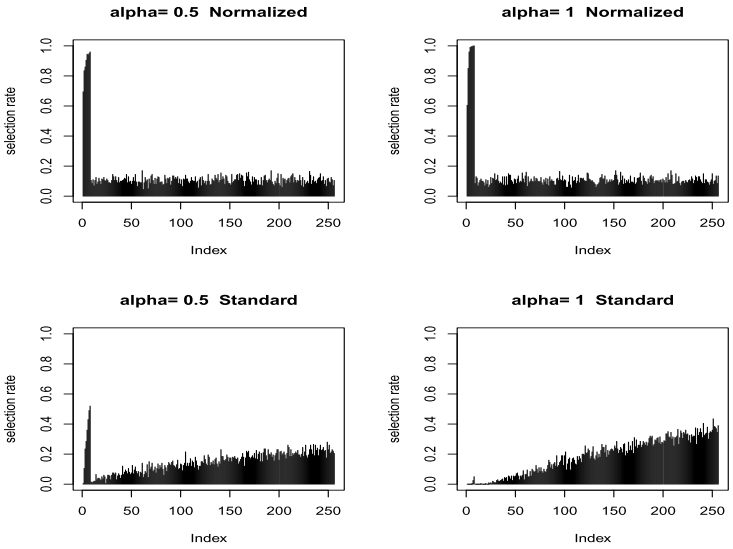

FIG 5. *Fraction of time when estimators are not equal 0 (β_2 case).*

of prediction errors $||X_{test}(\beta - \hat{\beta}_D)||/||X_{test}(\beta - \hat{\beta}_N)||$. It is seen that also in the terms of prediction error the standard estimator is inferior to the normalized one.

Figure 3 shows selection rate for each of the predictors for both estimators i.e. proportion of runs in which coefficient corresponding to a given variable is different from zero for α equal 0.5 and 1. It turns out that the standard estimator tends to select irrelevant variables with large variance and it ignores relevant

variables with low variance whereas normalized estimator does not have this drawback.

Figures 4 and 5 show corresponding results for vector of coefficients β_2. In this case the tendency is even stronger than for β_1. For large values of α the standard Dantzig selector chose irrelevant variables with large variance more frequently than relevant ones.

Acknowledgment

Comments of two referees and Associate Editor on the previous versions of the manuscript are appreciated.

References

[1] Bickel, P., Ritov, Y., and Tsybakov, A., Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009. MR2533469

[2] Bühlmann, P. and van de Geer, S., *Statistics for High-Dimensional Data*. Springer, New York, 2011. MR2807761

[3] Candès, E. and Plan, Y., Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37:2145–2177, 2009. MR2543688

[4] Candès, E. and Tao, T., The Dantzig Selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007. MR2382644

[5] James, G., Radchenko, P., and Lv, J., Dasso: Connections between the Dantzig selector and Lasso. *Journal of the Royal Statistical Society Series B*, 71:127–142, 2009. MR2655526

[6] Koltchinskii, V., The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009. MR2555200

[7] Lounici, K., Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008. MR2386087

[8] Meinshausen, N., Rocha, G., and Yu, B., A tale of three cousins: LASSO, L2boosting and Dantzig (discussion of Candès and Tao's Dantzig selector paper). *Annals of Statistics*, 35:2373–2384, 2007. MR2382649

[9] Pokarowski, P. and Mielniczuk, J., Combined ℓ_1 and ℓ_0 penalized least squares. *Journal of Machine Learning Research*, 16(May), 2015.

[10] Tibshirani, R., Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996. MR1379242

[11] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316

[12] Zhou, S., Thresholding procedures for high dimensional variable selection and statistical estimation. In *NIPS*, pages 2304–2312, 2009.