Electronic Journal of Statistics
Vol. 14 (2020) 944–969
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1687

On the distribution, model selection properties and uniqueness of the Lasso estimator in low and high dimensions

Karl Ewald and Ulrike Schneider

Department of Statistics and Mathematical Methods in Economics
Vienna University of Technology
Wiedner Hauptstrasse 8
A-1040 Vienna
e-mail: karl.ewald@tuwien.ac.at; ulrike.schneider@tuwien.ac.at

Abstract: We derive expressions for the finite-sample distribution of the Lasso estimator in the context of a linear regression model in low as well as in high dimensions by exploiting the structure of the optimization problem defining the estimator. In low dimensions, we assume full rank of the regressor matrix and present expressions for the cumulative distribution function as well as the densities of the absolutely continuous parts of the estimator. Our results are presented for the case of normally distributed errors, but do not hinge on this assumption and can easily be generalized. Additionally, we establish an explicit formula for the correspondence between the Lasso and the least-squares estimator. We derive analogous results for the distribution in less explicit form in high dimensions where we make no assumptions on the regressor matrix at all. In this setting, we also investigate the model selection properties of the Lasso and show that possibly only a subset of models might be selected by the estimator, completely independently of the observed response vector. Finally, we present a condition for uniqueness of the estimator that is necessary as well as sufficient.

MSC 2010 subject classifications: Primary 62E15; secondary 62J05, 62J07.
Keywords and phrases: Lasso, distribution, model selection, uniqueness.

Received April 2019.

1. Introduction

The distribution of the Lasso estimator (Tibshirani, 1996) has been an object of study in the statistics literature for a number of years. The often cited paper by Knight and Fu (2000) gives the asymptotic distribution of the Lasso in the framework of conservative model selection in a low-dimensional (fixed-p) framework by listing the limit of the corresponding stochastic optimization. Pötscher and Leeb (2009) derive explicit expressions of the distribution in finite samples as well as asymptotically for all large-sample regimes of the tuning parameter ("conservative" as well as "consistent model selection") in the framework of orthogonal regressors. More recently, Zhou (2014) gives high-level information on the finite-sample distribution for arbitrary designs in low and high dimensions, geared towards setting up a Monte-Carlo approach to infer about the distribution. In Ewald and Schneider (2018), the large-sample distribution of the Lasso is derived in a low-dimensional framework for the large-sample regime of the tuning parameter not considered in Knight and Fu (2000). Moreover, Jagannath and Upadhye (2018) consider the characteristic function of the Lasso to obtain approximate expressions for the marginal distribution of one-dimensional components of the Lasso when these components are "large", therefore not having to consider the atomic part of the estimator.

In this paper, we exactly and completely characterize the distribution of the Lasso estimator in finite samples in the context of a linear regression model with normal errors. In low dimensions, we give formulae for the cumulative distribution function (cdf), as well as the density functions conditional on which components of the estimator are non-zero. We do so assuming full column rank of the regressor matrix. Our results do not hinge on the normality assumption of the errors, but can easily be extended to more general error distributions. We also exactly quantify the correspondence between the Lasso and least-squares (LS) estimator, depending on the regressor matrix and tuning parameters only.

In a high-dimensional setting, we make absolutely no assumptions on the regressor matrix. We give formulae for the probability of the Lasso estimator falling into a given set and exactly quantify the relationship between the Lasso estimator and the data object $X'y$. Through this relationship, we also learn that the Lasso may never select certain models, this property depending only on the regressor matrix and the penalization weights and being independent of the observed response vector. In fact, we can characterize a so-called structural set that contains all covariates that are part of a Lasso model model for some response vector. This structural set can be identified by how the row space of the regressor matrix intersects a cube centered at the origin whose side lengths are determined by the penalization weights. The set may not contain all indices, in which case the Lasso estimator will rule out certain covariates for all possible observations of the dependent variable. This is related to the idea of so-called SAFE rules (Tibshirani et al., 2012) that can discard covariates for Lasso solutions for a fixed value of the dependent variable.

Finally, we present a condition for uniqueness of the Lasso estimator that is both necessary and sufficient, again related to how the row space of the regressor matrix intersects the above mentioned cube. Previously, only a sufficient condition for uniqueness has been known (see e.g. Tibshirani, 2013; Ali and Tibshirani, 2019). The results quantifying the relationship between the Lasso and the LS estimator or $X'y$, respectively, are in fact completely independent of the error distribution and merely utilize the given values of the dependent variable and the regressor matrix. The results on model selection properties and uniqueness use the regressor matrix and penalization weights only.

The paper is organized as follows. We introduce the setting and notation in Section 2 and state basic results used throughout the paper. The low-dimension case is treated in Section 3, whereas we consider the high-dimensional case in Section 4. We conclude in Section 5.

2. Setting, notation, and basic results

Consider the linear model

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

where y is the observed $n \times 1$ data vector, X is the $n \times p$ regressor matrix which is assumed to be non-stochastic, $\beta \in \mathbb{R}^p$ is the true parameter vector and ε the unobserved error term. We assume that $X'\varepsilon$ follows a $N(0, \sigma^2 X'X)$-distribution with $\sigma^2 > 0$, which, for instance, is the case when the components of ε are independent and identically distributed according to a $N(0, \sigma^2)$-distribution. Our results depend on the distribution of $X'\varepsilon$ only, and we chose the normal distribution for presentation purposes.[1]

We consider the *weighted Lasso estimator* $\hat{\beta}_{\mathrm{L}}$, defined as a solution to the minimization problem

$$\min_{\beta \in \mathbb{R}^p} L(\beta) = \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|^2 + 2\sum_{j=1}^{p} \lambda_{n,j} |\beta_j|, \tag{2}$$

where $\lambda_{n,j}$, are non-negative user-specified tuning parameters that will typically depend on n. To ease notation, we shall suppress this dependence for the most part and write $\lambda_{n,j} = \lambda_j$ for each j. Note that if $\lambda_j = 0$ for all j, the weighted Lasso is equal to LS estimation, and that $\lambda_1 = \cdots = \lambda_p > 0$ corresponds to the classical Lasso estimator as proposed by Tibshirani (1996), to which case we also refer to by uniform tuning. For later use, let $\lambda = (\lambda_1, \ldots, \lambda_p)'$ and define $\mathcal{M}_0 = \{j : \lambda_j = 0\}$, the index set of all unpenalized coefficients. If $\mathcal{M}_0 \neq \varnothing$, we speak of partial tuning. Note that $\mathcal{M}_0$ contains the indices of covariates that will be part of any model chosen by the Lasso. We stress dependence on the unknown parameter β when it occurs, but do not specify dependence on X, y or λ as these quantities are available to the user.

The following notation will be used throughout the paper. Let $\phi_{(\mu,\Sigma)}$ denote the Lebesgue-density of a normally distributed random variable with mean μ and covariance matrix Σ, and let Φ be the cdf of a univariate standard normal distribution. For a vector $m \in \mathbb{R}^p$ and an index set $I \subseteq \{1, \ldots, p\}$, the vector $m_I \in \mathbb{R}^{|I|}$ contains only the components of m corresponding to the elements of I. We write $|I|$ for the cardinality of I, and I^c for $\{1, \ldots, p\} \setminus I$, the complement of I. The 1-norm of m is denoted by $\|m\|_1$ whereas the 2-norm is simply denoted by $\|m\|$. For $x \in \mathbb{R}$, let $\mathrm{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$ where $\mathbb{1}$ is the indicator function. For a set $A \subseteq \mathbb{R}^p$, the set $m + A = A + m$ is defined as $\{m + z : z \in A\}$, with a analogous definitions for $A - m$ and $m - A$. We denote the Cartesian product by $\prod$ and the kernel, column space, and rank of a matrix C by $\ker(X)$, $\mathrm{col}(C)$, and $\mathrm{rk}(C)$, respectively. The columns of C are denoted by C_j whereas C_I, for some index set I, is the matrix containing the $|I|$ columns of C corresponding to indices in I only. For a quadratic matrix C, $|C|$ denotes the determinant of C. We use $\mathbb{R}_{>0}$ for the positive, and $\mathbb{R}_{\geq 0}$ for the non-negative real numbers.

[1] In fact, some of the following results are completely independent of the error distribution. These are Lemma 1 and Corollary 2, Theorems 8 and 12, Corollary 13 as well as Theorems 14 and 15.

Let $\{D_-, D_0, D_+\}$ be a partition of $\{1, \dots, p\}$ into three sets, some of which may be empty. It will be convenient to also describe this partition by a vector $d \in \{-1, 0, 1\}^p$ with $d_j = \mathbb{1}_{\{j \in D_+\}} - \mathbb{1}_{\{j \in D_-\}}$. For such d, we denote by $\mathcal{O}^d = \{z \in \mathbb{R}^p : \operatorname{sgn}(z_j) = d_j \text{ for } j = 1, \dots, p\} = \{z \in \mathbb{R}^p : z_j < 0 \text{ for } j \in D_-, z_j = 0 \text{ for } j \in D_0, z_j > 0 \text{ for } j \in D_+\}$. Note that $m + \beta \in \mathcal{O}^d$ is short-hand notation for $m_j < -\beta_j$ for $j \in D_-$, $m_j = -\beta_j$ for $j \in D_0$ and $m_j > -\beta_j$ for $j \in D_+$. We write $D_\pm$ for $D_- \cup D_+$.

Finally, we state the conditions that characterize solutions to (2), known as the Kuhn-Karush-Tucker (KKT) conditions for the Lasso (see e.g. Tibshirani, 2013, with the slight adaptation that we use componentwise tuning) which form the basis of most proofs and will be used throughout the article.

Lemma 1. *We have that*

$$b \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(\beta) \iff X'y \in X'Xb + \prod_{j=1}^{p} B_j(b_j),$$

where

$$B_j(b_j) = \begin{cases} \{\operatorname{sgn}(b_j)\lambda_j\} & b_j \neq 0 \\ [-\lambda_j, \lambda_j] & b_j = 0. \end{cases}$$

Plugging in $y = X\beta + u$, we rewrite this as

Corollary 2.

$$b \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(\beta) \iff X'\varepsilon \in A_\beta(b)$$

with $A_\beta(b) = X'X(b - \beta) + \prod_{j=1}^{p} B_j(b_j)$.

3. The low-dimensional case

Throughout this section, *we assume that X has full column rank p*, implying that we are considering the low-dimensional setting where $p \leq n$. This assumption is used in the following through the fact that $W = X'\varepsilon$ follows a non-degenerate normal distribution on $\mathbb{R}^p$ in the distributional statements on $\hat{\beta}_{\mathrm{L}}$ (Theorem 3, Corollary 5). It is additionally used through the fact that the true parameter is properly identified in the distributional statements concerning the estimation error $\hat{\beta}_{\mathrm{L}} - \beta$ (Corollary 4, Proposition 6, Theorem 7). We also rely on this assumption in Theorem 8 through the invertibility of $X'X$ and the existence of the LS estimator.

Theorem 3. *Let $z \in \mathbb{R}^p$ and let $d = \operatorname{sgn}(z)$ with $\{D_-, D_0, D_+\}$ being the corresponding partition of $\{1, \dots, p\}$.*

$$P(\hat{\beta}_{\mathrm{L},j} \leq z_j \textit{ for } j \in D_-,\ \hat{\beta}_{\mathrm{L},j} = 0 \textit{ for } j \in D_0,\ \hat{\beta}_{\mathrm{L},j} \geq z_j \textit{ for } j \in D_+)$$

$$= \underset{\substack{m_j \geq z_j - \beta_j \\ j \in D_+}}{\int \cdots \int} \ \underset{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}}{\int \cdots \int} \ \underset{\substack{m_j \leq z_j - \beta_j \\ j \in D_-}}{\int \cdots \int} \phi_{(0, \sigma^2 X'X)}(X'Xm_\beta + s_\lambda) |X'_{D_\pm} X_{D_\pm}| \, dm_{D_-} ds_{D_0} dm_{D_+},$$

where m_β and $s_\lambda \in \mathbb{R}^p$ are given by $(m_\beta)_{D_\pm} = m_{D_\pm}$, $(m_\beta)_{D_0} = -\beta_{D_0}$ and $(s_\lambda)_{D_-} = -\lambda_{D_-}, s_{D_+} = \lambda_{D_+}$, $(s_\lambda)_{D_0} = s_{D_0}$, respectively.

Proof. We need to compute the probability of the event $\{\hat{\beta}_{\mathrm{L}} \in B_z\}$ where $B_z = \{b \in \mathbb{R}^p : b_j \leq z_j \text{ for } j \in D_-,\, b_j = 0 \text{ for } j \in D_0,\, b_j \geq z_j \text{ for } j \in D_+\}$. By Corollary 2, this event is equivalent to the event $\{W \in A_\beta(B_z)\}$ with $A_\beta(B_z) = \cup_{b \in B_z} A_\beta(b)$ and $A_\beta(b)$ as defined in Corollary 2. As $W = X'\varepsilon \sim N(0, \sigma^2 X'X)$, the probability we are looking for is therefore given by

$$\int_{A_\beta(B_z)} \phi_{(0,\sigma^2 X'X)}(w) dw.$$

We now look at the structure of the set $A_\beta(B_z)$ more concretely. Since $\operatorname{sgn}(b) = \operatorname{sgn}(z)$ for all $b \in B_z$, we get

$$\begin{aligned} A_\beta(B_z) &= X'XB_z - X'X\beta + \prod_{j=1}^p B_j(z_j) \\ &= \{X'X(b-\beta) : b_j \leq z_j \text{ for } j \in D_-,\, b_j = 0 \text{ for } j \in D_0,\, b_j \geq z_j \text{ for } j \in D_+\} \\ &\quad + \{s : s_j = -\lambda_j \text{ for } j \in D_-,\, |s_j| \leq \lambda_j \text{ for } j \in D_0,\, s_j = \lambda_j \text{ for } j \in D_+\}, \end{aligned}$$

which, after applying the substitution $w = X'Xm_\beta + s_\lambda$, yields the claim. □

Theorem 3 gives the distribution of $\hat{\beta}_{\mathrm{L}}$. The dependence on the unknown parameter β arises in the shift $(m_\beta)_{D_0} = -\beta_{D_0}$ as well as in the limits for the variables of integration m_{D_+} and m_{D_-}. In case the regressors are orthogonal, more concretely, if $X'X = I_p$, the probability expression in Theorem 3 can be written as

$$\begin{aligned} &\prod_{j=1}^p \Bigg(\mathbb{1}_{\{j \in D_-\}} \int_{-\infty}^{z_j - \beta_j} \phi_{(0,\sigma^2)}(m - \lambda_j) dm + \mathbb{1}_{\{j \in D_0\}} \int_{-\lambda_j}^{\lambda_j} \phi_{(0,\sigma^2)}(s - \beta_j) ds \\ &\qquad + \mathbb{1}_{\{j \in D_+\}} \int_{z_j - \beta_j}^{\infty} \phi_{(0,\sigma^2)}(m + \lambda_j) dm \Bigg) \\ &= \prod_{j=1}^p \Bigg(\mathbb{1}_{\{j \in D_-\}} P(Z_j + \beta_j \leq z_j - \lambda_j) + \mathbb{1}_{\{j \in D_0\}} P(-\lambda_j \leq Z_j + \beta_j \leq \lambda_j) \\ &\qquad + \mathbb{1}_{\{j \in D_+\}} P(Z_j + \beta_j \geq z_j + \lambda_j) \Bigg), \end{aligned}$$

where $Z_j \overset{\text{iid}}{\sim} N(0, \sigma^2)$, which is consistent with the well-known fact that the Lasso is equivalent to componentwise soft-thresholding in this case, also treated in Pötscher and Schneider (2009).

The distribution of the estimation error $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$ can now be derived from Theorem 3 as a corollary.

Corollary 4. *Let $z \in \mathbb{R}^p$. Let $d = \operatorname{sgn}(z+\beta) \in \{-1, 0, 1\}^p$ and let $\{D_-, D_0, D_+\}$ be the corresponding partition of $\{1, \dots, p\}$. Then*

$$P(\hat{u}_j \leq z_j \text{ for } j \in D_-,\, \hat{u}_j = z_j \text{ for } j \in D_0,\, \hat{u}_j \geq z_j \text{ for } j \in D_+)$$

$$= \underset{\substack{m_j \geq z_j \\ j \in D_+}}{\int \cdots \int} \underset{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}}{\int \cdots \int} \underset{\substack{m_j \leq z_j \\ j \in D_-}}{\int \cdots \int} \phi_{(0,\sigma^2 X'X)}(X'Xm_\beta + s_\lambda)|X'_{D_\pm}X_{D_\pm}|\, dm_{D_-}\, ds_{D_0} dm_{D_+},$$

where m_β and $s_\lambda \in \mathbb{R}^p$ are given by $(m_\beta)_{D_\pm} = m_{D_\pm}$, $(m_\beta)_{D_0} = -\beta_{D_0}$ and $(s_\lambda)_{D_-} = -\lambda_{D_-}, s_{D_+} = \lambda_{D_+}$, $(s_\lambda)_{D_0} = s_{D_0}$, respectively.

Proof. Apply Theorem 3 using $z + \beta$ rather than z. □

Another direct consequence of Theorem 3 is a concrete formula for the probability of the extreme event of the Lasso setting all components equal to zero.

Corollary 5.

$$P(\hat{\beta}_{\mathrm{L}} = 0) = \int_{-\lambda_p}^{\lambda_p} \cdots \int_{-\lambda_1}^{\lambda_1} \phi_{(X'X\beta, \sigma^2 X'X)}(w)\, dw.$$

Proof. Apply Theorem 3 using $z = 0$. □

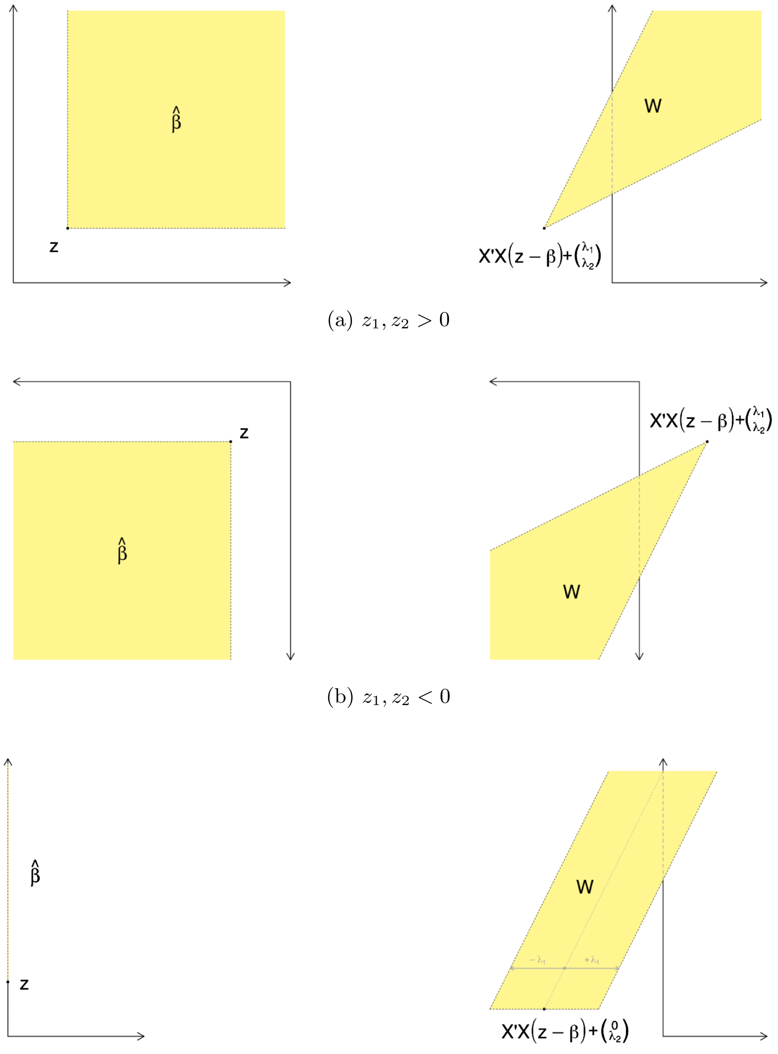
Remark 1. *To illustrate the structure behind the proof of Theorem 3, note that the equivalence of the events $\{\hat{\beta} \in B_z\}$ and $\{W \in A_\beta(B_z)\}$ is shown through through Corollary 2. The equivalence holds due to the structure of the optimization problem defining $\hat{\beta}_{\mathrm{L}}$ and does not depend on the distribution of $W = X'\varepsilon$. In this sense, the distributional results do not hinge on the normality assumption of the errors and can easily be generalized to other error distributions. The relationship and shape of the sets B_z and $A_\beta(B_z)$ is illustrated in Figure 1. Note that A_β depends on λ, whereas B_z does not.*

Remark 2. *Theorem 3, Corollary 4, Corollary 5, Proposition 6 and Theorem 7 do not rely on the normal distribution, as just mentioned. Indeed, the results equally hold for any other absolutely continuous distribution of $X'\varepsilon$ (with respect to Lebesgue measure), only the expression $\phi_{(0,\sigma^2 X'X)}$ would have to be replaced by the corresponding density function of $X'\varepsilon$. Moreover, the results also hold for discrete $X'\varepsilon$ in which case the integral would have to be replaced by a sum, and the density function by the corresponding probability mass function.*

Theorem 3 now puts us into a position to fully specify the distribution of the Lasso estimator. In case $\lambda_j > 0$ for all j, one easily sees from the preceding corollary that this distribution is not absolutely continuous with respect to the p-dimensional Lebesgue-measure, and thus no density exists. One can, however, represent the distribution through Lebesgue-densities after conditioning on which components of the estimator are negative, equal to zero, and positive, which we shall do in the sequel.

Proposition 6. *The distribution of $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$, conditional on the event $\{\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d\}$, can be represented by a $\|d\|_1$-dimensional Lebesgue-density given by*

$$f^d(z_{D_\pm}) = \frac{\mathbb{1}\{z_\beta + \beta \in \mathcal{O}^d\}}{P(\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d)} \underset{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}}{\int \cdots \int} \phi_{(0,\sigma^2 X'X)}\left(X'Xz_\beta + s_\lambda\right)|X'_{D_\pm}X_{D_\pm}|\, ds_{D_0},$$


$\hat{\beta}$
z
W
$X'X(z-\beta)+\binom{\lambda_1}{\lambda_2}$
z
$X'X(z-\beta)+\binom{\lambda_1}{\lambda_2}$
$\hat{\beta}$
W
$\hat{\beta}$
z
W
$-\lambda_1$
$+\lambda_1$
$X'X(z-\beta)+\binom{0}{\lambda_2}$

(a) $z_1, z_2 > 0$

(b) $z_1, z_2 < 0$

(c) $z_1 = 0$ and $z_2 > 0$

FIG 1. *The sets B_z are displayed on the left-hand side, the corresponding sets $A_\beta(B_z)$ are displayed on the right-hand side. Illustration for $p = 2$ and various values of z, see Remark 1 for details.*

where z_β is defined by $(z_\beta)_{D_\pm} = z_{D_\pm}$, and $(z_\beta)_{D_0} = -\beta_{D_0}$ and s_λ is defined by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_+} = \lambda_{D_+}$, and $(s_\lambda)_{D_0} = s_{D_0}$. Note that the constants $P(\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d)$ can be calculated using Corollary 3.

Proof. Observe that

$$f^d(z_{D_\pm}) = \left(\frac{\partial}{\partial z_j}\right)_{j \in D_\pm} P\left(\hat{u}_j \le z_j \text{ for } j \in D_\pm | \hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d\right),$$

and note that by Corollary 4, for any $z \in \mathbb{R}^p$ with $z + \beta \in \mathcal{O}^d$ we have

$$\begin{aligned}
&P\left(\hat{u}_j \leq z_j \text{ for } j \in D_-, \hat{u}_j \geq z_j \text{ for } j \in D_+ | \hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d\right) P(\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d) \\
&\quad = P\left(\hat{u}_j \leq z_j \text{ for } j \in D_-,\ \hat{\beta}_{\mathrm{L},j} = 0 \text{ for } j \in D_0,\ \hat{u}_j \geq z_j \text{ for } j \in D_+\right) P(\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d) \\
&\quad = \underset{\substack{m_j \geq z_j \\ j \in D_+}}{\int \cdots \int} \underset{\substack{m_j \leq z_j \\ j \in D_-}}{\int \cdots \int} \underset{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}}{\int \cdots \int} \phi_{(0,\sigma^2 X'X)}(X'Xm_\beta + s_\lambda) |X'_{D_\pm} X_{D_\pm}| \, ds_{D_0} dm_{D_-} dm_{D_+},
\end{aligned}$$

where $m_\beta \in \mathbb{R}^p$ is defined by $(m_\beta)_{D_\pm} = m_{D_- \cup D_+}$, and $(m_\beta)_{D_0} = -\beta_{D_0}$, and $s_\lambda \in \mathbb{R}^p$ is defined by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_+} = \lambda_{D_+}$, and $(s_\lambda)_{D_0} = s_{D_0}$. Differentiating with respect to z_j with $j \in D_\pm$, and taking the absolute value gives the density, thus completing the proof. □

Besides the conditional densities, we can also specify the full cdf of $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$ which is done in the following theorem.

Theorem 7. *The cdf of $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$ is given by*

$$F(z) = P(\hat{u}_1 \leq z_1, \ldots, \hat{u}_p \leq z_p) = \sum_{d \in \{-1,0,1\}^p} \underset{\substack{m_j \leq z_j \\ j \in D_\pm}}{\int \cdots \int} h^d(m_{D_\pm}) \, d\nu_{\|d\|_1},$$

where ν_k denotes k-dimensional Lebesgue-measure, and where

$$h^d(m_{D_\pm}) = \mathbb{1}\{m_\beta + \beta \in \mathcal{O}^d\} \underset{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}}{\int \cdots \int} \phi_{(0,\sigma^2 X'X)}\left(X'Xm_\beta + s_\lambda\right) |X'_{D_\pm} X_{D_\pm}| ds_{D_0},$$

with $m_\beta \in \mathbb{R}^p$ given by $(m_\beta)_{D_\pm} = m_{D_\pm}$, $(m_\beta)_{D_0} = -\beta_{D_0}$, and $s_\lambda \in \mathbb{R}^p$ given by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_0} = s_{D_0}$, and $(s_\lambda)_{D_+} = \lambda_{D_+}$.

Proof. It is easily seen that

$$P\left(\hat{u}_1 \leq z_1, \ldots, \hat{u}_p \leq z_p\right) = \sum_{d \in \{-1,0,1\}^p} P(\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^d) \underset{\substack{m_j \leq z_j \\ j \in D_\pm}}{\int \cdots \int} f^d(m_{D_\pm}) \, d\nu_{\|d\|_1}.$$

Plugging in the formula for f^d completes the proof. □

For illustration of Proposition 6 and Theorem 7, consider Figures 2 and 3 which display an example of the distribution of $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$. One can see that the Lasso estimation error follows a shifted normal distribution, conditional on the event $\hat{u}_j \neq -\beta_j$ ($\hat{\beta}_{\mathrm{L},j} \neq 0$) for each j, with the shift depending on the signs of $\hat{\beta}_{\mathrm{L}}$, as can be seen in Figure 2. Figure 3 displays the mass which lies on the set $\{z \in \mathbb{R}^2 : z_1 = -\beta_1, z_2 \neq 0\}$, that is, the density functions $h^{(0,1)}$ and $h^{(0,-1)}$ on their corresponding domains. The mass on the set $\{z \in \mathbb{R}^2 : z_1 \neq 0, z_2 = -\beta_2\}$ looks qualitatively similar to Figure 3. Note that we also have point-mass at $-\beta$, as is pointed out by Corollary 5.

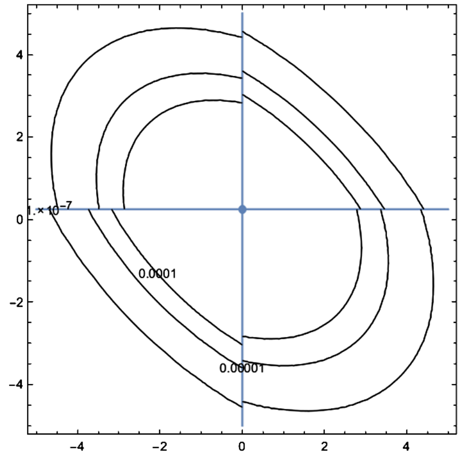


FIG 2. *The contour lines of the absolutely continuous part of the distribution of $\hat{\beta}_{\mathrm{L}} - \beta$ with respect to 2-dimensional Lebesgue-measure, for $X'X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\lambda = (0.75, 0.75)'$, and $\beta = (0, -0.25)'$. Note that the blue lines as well as the blue point also carry probability mass.*

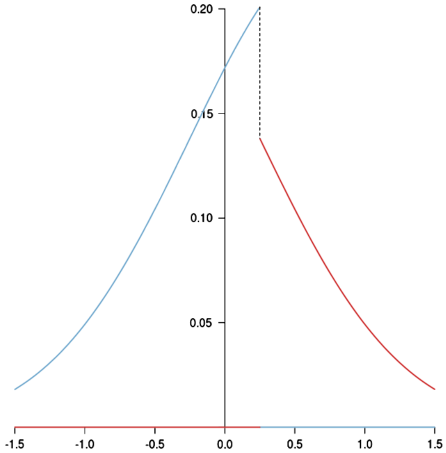


FIG 3. *The functions $h^{(0,-1)'}$ (in blue) and $h^{(0,1)'}$ (in red) for $X'X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\lambda = (0.75, 0.75)'$ and $\beta = (0, -0.25)'$, corresponding to the absolutely continuous part.*

3.1. Shrinkage areas

Using the conditions for minimality from Lemma 1, we can establish a direct relationship between the LS and the Lasso estimator in the following sense. For any $b \in \mathbb{R}^p$, there exists a set $S(b) \subseteq \mathbb{R}^p$, such that the Lasso estimator assumes the value b if and only if the LS estimator lies in $S(b)$. We refer to the set $S(b)$ as *shrinkage area* since the Lasso estimator can be viewed as a procedure that shrinks the LS estimates from the set $S(b)$ to the point b. Note that by shrinkage, we mean that $\|b\|_1 \leq \|z\|_1$ for each $z \in S(b)$, but $|b_j| > |z_j|$ could hold for certain components. The explicit form of $S(b)$ is formalized in the following theorem.

Theorem 8. *For each $b \in \mathbb{R}^p$ there exists a set $S(b) \subseteq \mathbb{R}^p$, such that*

$$\hat{\beta}_{\mathrm{L}} = b \Longleftrightarrow \hat{\beta}_{\mathrm{LS}} \in S(b).$$

Moreover, for $b \in \mathcal{O}^d$, the set $S(b)$ is given by

$$\begin{aligned} S(b) = \{z \in \mathbb{R}^p :& (X'Xz)_j = (X'Xb)_j + \operatorname{sgn}(b_j)\lambda_j \text{ for } j \in D_\pm, \\ & |(X'X(z-b))_j| \leq \lambda_j \text{ for } j \in D_0\}. \end{aligned}$$

Clearly, the sets $S(b)$ are disjoint for different b's.

Proof. Using Lemma 1, we find that $\hat{\beta}_{\mathrm{L}} = b$ holds if and only if $X'X\hat{\beta}_{\mathrm{LS}} \in X'Xb + \prod_{j=1}^p B_j(b_j)$, which, for $b \in \mathcal{O}^d$, holds if and only if

$$\begin{cases} (X'X\hat{\beta}_{\mathrm{LS}})_j = (X'X\hat{\beta}_{\mathrm{L}})_j - \lambda_j & \text{for } j \in D_- \\ |(X'X(\hat{\beta}_{\mathrm{LS}} - \hat{\beta}_{\mathrm{L}}))_j| \leq \lambda_j & \text{for } j \in D_0 \\ (X'X\hat{\beta}_{\mathrm{LS}})_j = (X'X\hat{\beta}_{\mathrm{L}})_j + \lambda_j & \text{for } j \in D_+, \end{cases}$$

or, $\hat{\beta}_{\mathrm{LS}} \in S(b)$, as required.

The sets are disjoint since, in case $\operatorname{rk}(X) = p$, all Lasso solutions are unique. If $S(b) \cap S(\tilde{b}) \neq \varnothing$ holds for some $b \neq \tilde{b}$, we can find $y \in \mathbb{R}^n$ such that $\hat{\beta}_{\mathrm{LS}} = (X'X)^{-1}X'y \in S(b) \cap S(\tilde{b})$ implying that both b and $\tilde{b}$ are Lasso solutions for the given y, yielding a contradiction. □

Remark 3. *Clearly, if $b \in \mathbb{R}^p$ satisfies $b_j \neq 0$ for all $j = 1, \ldots, p$, then $S(b)$ is the singleton*

$$S(b) = \{b + (X'X)^{-1}\tilde{\lambda}\},$$

where $\tilde{\lambda}_j = \operatorname{sgn}(b_j)\lambda_j$ for $j = 1, \ldots, p$. This implies that, in case $\hat{\beta}_{\mathrm{L},j} \neq 0$ for all j, the Lasso estimator is given by

$$\hat{\beta}_{\mathrm{L}} = \hat{\beta}_{\mathrm{LS}} + (X'X)^{-1}\tilde{\lambda}.$$

Note that aside from b, $S(b)$ depends on X and λ only.

Given Theorem 8, we can identify areas in which components of the LS estimator are shrunk to zero by the Lasso. For $p = 2$, it leads to the image displayed in Figure 4. Clearly, the shrinkage areas are related to the polyhedral selection areas developed in Tibshirani and Taylor (2012) and employed for instance in Lee et al. (2016), but yield a different kind of information. Our results identify the regions of the LS estimator that lead to a particular value b of the Lasso estimator. The polyhedral regions in the above articles identify the regions of the dependent variable y that correspond to a particular Lasso model with specific signs of the active coefficients. Naturally, our regions are subsets of $\mathbb{R}^p$, while the polyhedral regions are subsets of $\mathbb{R}^n$ (the latter ones also allowing to interpret the Lasso fit as a projection, and not depending on full column rank).

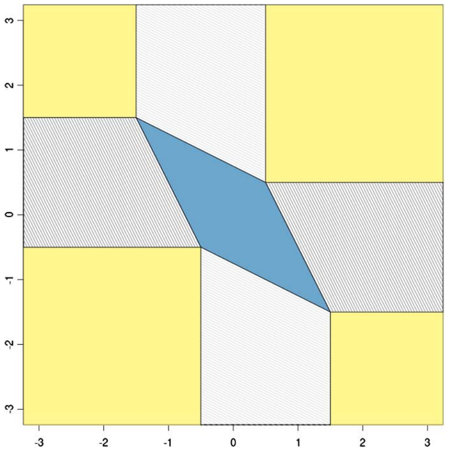


Fig 4. *The shrinkage areas from Theorem 8 for $p = 2$. The blue parallelogram equals the set $S(0)$. The dark gray area consists of lines parallel to the adjacent edge of the parallelogram, where each line equals a set $S\binom{0}{b_2}$ for $b_2 \neq 0$. Analogously, the light gray area consists of lines parallel to the adjacent edge of the parallelogram, and each of those lines equals a set $S\binom{b_1}{0}$ for $b_1 \neq 0$. The yellow areas contain all singletons $S\binom{b_1}{b_2}$, with $b_1, b_2 \neq 0$ as described in Remark 3. In this example, $X'X = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, and $\lambda = (0.75, 0.75)'$.*

4. High-dimensional case

We now turn to the main case of this this article, the high-dimensional setting where $p > n$. *We make no assumptions on the regressor matrix X in this section.* Using similar arguments as in the case $p \leq n$, we can again start by characterizing the distribution of the Lasso, albeit in a somewhat less explicit form. Note that we have $\text{rk}(X) < p$ and that the true parameter is not identified without further assumptions. We denote by $\mathcal{B}_0$ the set of all $\beta \in \mathbb{R}^p$ that yield the model given in (1), that is, $\mathcal{B}_0 = \{\beta \in \mathbb{R}^p : X\beta = \mathbb{E}(y) = \mu\}$. Furthermore, it is important to note that the Lasso solution need not be unique anymore. We give necessary and sufficient conditions for uniqueness later in Section 4.3.

All findings in this section also hold when $p \leq n$, but more explicit results for this case are found in Section 3. We start with a high-level result on the distribution of $\hat{\beta}_{\text{L}}$, which immediately follows from Corollary 2.

Theorem 9. *For any set $B \subseteq \mathbb{R}^p$ and any $\beta \in \mathcal{B}_0$, we have*

$$P(\underset{\beta \in \mathbb{R}^p}{\arg\min}\, L(\beta) \cap B \neq \varnothing) = P(W \in A_\beta(B)),$$

where $W \sim N(0, \sigma^2 X'X)$, and $A_\beta(B) = \bigcup_{b \in B} A_\beta(b)$, with $A_\beta(b) = X'X(b - \beta) + \prod_{j=1}^p B_j(b_j)$, and

$$B_j(b_j) = \begin{cases} \{\text{sgn}(b_j)\lambda_j\} & b_j \neq 0 \\ [-\lambda_j, \lambda_j] & b_j = 0. \end{cases}$$

In particular, the distribution of the estimator $\hat{\beta}_{\text{L}}$ does not depend on the choice of $\beta \in \mathcal{B}_0$.

To derive the analogue of the distribution of the estimation error in high dimensions for a fixed $\beta \in \mathcal{B}_0$, define the function $V_\beta(u) = L(u+\beta) - L(\beta)$ given by

$$V_\beta(u) = L(u+\beta) - L(\beta) = u'X'Xu - 2u'W + 2\sum_{j=1}^{p} \lambda_j \left[|u_j + \beta_j| - |\beta_j|\right], \quad (3)$$

which is minimized at $\hat{\beta}_{\text{L}} - \beta$, and where $\hat{\beta}_{\text{L}}$ may be any minimizer of $L(\boldsymbol{\beta})$. The high-dimensional version of Corollary 4 can now be formulated as

Theorem 10. *For any set $M \subseteq \mathbb{R}^p$ and any $\beta \in \mathcal{B}_0$, we have*

$$P(\operatorname*{arg\,min}_{u\in\mathbb{R}^p} V_\beta(u) \cap M \neq \varnothing) = P(W \in \bar{A}_\beta(M)),$$

where $W \sim N(0, \sigma^2 X'X)$ and $\bar{A}_\beta(M) = \bigcup_{m\in M} \bar{A}_\beta(m)$ with $\bar{A}_\beta(m) = X'Xm + \prod_{j=1}^{p} \bar{B}_{\beta,j}(m_j)$ and

$$\bar{B}_{\beta,j}(m_j) = \begin{cases} \{\operatorname{sgn}(m_j+\beta_j)\lambda_j\} & m_j + \beta_j \neq 0 \\ [-\lambda_j, \lambda_j] & m_j + \beta_j = 0. \end{cases}$$

Proof. As stated above, $m \in \mathbb{R}^p$ is a minimizer of V_β if and only if $m + \beta$ is a minimizer of $L(\boldsymbol{\beta})$. Corollary 2 then yields

$$m \in \operatorname*{arg\,min}_{u\in\mathbb{R}^p} V_\beta(u) \iff W \in A_\beta(m+\beta) \iff W \in \bar{A}_\beta(m). \qquad \square$$

Remark 4. *Note that, just as for the low-dimensional case discussed in Remark 2, the statements in Theorem 9 and Corollary 10 do not hinge on the normal distribution of $W = X'\varepsilon$. In fact, the both results equally hold for arbitrary distributions of W.*

While the distribution of $\hat{\beta}_{\text{L}} - \beta$ depends on the choice of $\beta \in \mathcal{B}_0$, the distribution of $\hat{\beta}_{\text{L}}$ does not, as it is determined by $y \sim N(\mu, \sigma^2 I_n)$. This is further formalized in the following corollary. As mentioned before, $\hat{\beta}_{\text{L}}$ need not be unique. Also remember that $\hat{\beta}_{\text{L}}$ itself minimizes the function $L(\boldsymbol{\beta})$ defined in (2).

As the random variable $W = X'\varepsilon$ has singular covariance matrix, some care needs to be taken when computing the probability from Corollary 10 through the appropriate integral of the corresponding density function.

Corollary 11. *Let the columns of U form a basis of $\operatorname{col}(X')$. The probability that a Lasso solution lies in the set $B \subseteq \mathbb{R}^p$ can be written as*

$$P(\operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta}) \cap B \neq \varnothing) = \mathbb{1}\{\operatorname{col}(X') \cap A_\beta(B) \neq \varnothing\} \int_{U'A_\beta(B)} \phi_{(0,\sigma^2 U'X'XU)}(w)dw.$$

Proof. Note that $U'W \sim N(0, \sigma^2 U'X'XU)$ and that $U'X'XU$ is invertible. Let N be a matrix whose columns form a basis of $\operatorname{col}(X')^{\perp}$, so that $N'W$ has covariance matrix $\sigma^2 N'X'XN = 0$, yielding $N'W = 0$ almost surely. We therefore have

$$\begin{aligned} W \in A_\beta(B) &\iff (U,N)'W \in (U,N)'A_\beta(B) \\ &\iff U'W \in U'A_\beta(B) \text{ and } 0 \in N'A_\beta(B) \\ &\iff U'W \in U'A_\beta(B) \text{ and } \operatorname{col}(X') \cap A_\beta(B) \neq \varnothing, \end{aligned}$$

which proves the claim. □

4.1. Selection regions and model selection properties

In the low-dimensional case, Theorem 8 gives what we call shrinkage areas of the Lasso with respect to the LS estimator. As the latter is never uniquely defined in the high-dimensional case, we instead look at the object $X'y$ and and consider so-called *selection regions* with respect to this quantity: for any $b \in \mathbb{R}^p$, we provide a set $T(b)$ such that a Lasso solution is equal to b if and only if $X'y$ lies in the set $T(b)$. The corresponding result turns out to be a restatement of Lemma 1, which we list again in the following for the sake of completeness.

Theorem 12. *For each $b \in \mathbb{R}^p$ there exists a set $T(b) \subseteq \mathbb{R}^p$ such that*

$$b \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(\beta) \iff X'y \in T(b).$$

Moreover, $T(b)$ is given by

$$T(b) = X'Xb + \prod_{j=1}^{p} B_j(b_j)$$

with

$$B_j(b_j) = \begin{cases} \{\operatorname{sgn}(b_j)\lambda_j\} & b_j \neq 0 \\ [-\lambda_j, \lambda_j] & b_j = 0. \end{cases}$$

Remark 5. *Analogously to the low-dimensional case, the sets $T(b)$ are singletons if $b \in \mathbb{R}^p$ satisfies $b_j \neq 0$ for all $j = 1, \ldots, p$:*

$$T(b) = \{X'Xb + \tilde{\lambda}\},$$

where $\tilde{\lambda}_j = \operatorname{sgn}(b_j)\lambda_j$ for $j = 1, \ldots, p$. Also, aside from b, the sets $T(b)$ depend on X and λ only.

Inspecting the sets $T(b)$ from Theorem 12 more closely, we see that they are, in general, not disjoint for different values of $b \in \mathbb{R}^p$. This illustrates the fact that, in contrast to the low-dimensional case, the Lasso solution need not be unique in high dimensions anymore. Indeed, we can have $T(b) \cap T(b') \neq \varnothing$, as long as $b - b' \in \ker(X)$ and $\{\operatorname{sgn}(b_j), \operatorname{sgn}(b'_j)\} \neq \{-1, 1\}$ for all j. This also makes apparent that b and b' may be Lasso solutions not corresponding to the same model, which has been noted by Tibshirani (2013) for the case of $\lambda_1 = \cdots = \lambda_p > 0$. We get deeper into the issue of (non-)uniqueness in Section 4.3.

Theorem 12 also sheds some light on which models $\mathcal{M} \subseteq \{1, \ldots, p\}$ may in fact be chosen by the Lasso estimator, where the Lasso model is given by $\{j : \hat{\beta}_{\mathrm{L},j} \neq 0\}$. We find that some models will, in fact, never be selected by the Lasso. This is illustrated in Figure 5 below, where the Lasso always sets the first component to zero, independently of y. This leads to the question on how to determine whether a particular model $\mathcal{M}$ may or may not be chosen.

Along these lines, define $\mathcal{B}_{\mathcal{M}} = \{b \in \mathbb{R}^p : b_j \neq 0 \text{ if and only if } j \in \mathcal{M}\}$. Then there exists a $y \in \mathbb{R}^n$ such that a corresponding corresponding Lasso solution chooses $\mathcal{M}$ if and only if there exists $y \in \mathbb{R}^n$ such that $X'y \in T(\mathcal{B}_{\mathcal{M}})$. In other words, this is the case if and only if $\operatorname{col}(X') \cap T(\mathcal{B}_{\mathcal{M}}) \neq \varnothing$ with $T(\mathcal{B}_{\mathcal{M}}) = \bigcup_{b \in \mathcal{B}_{\mathcal{M}}} T(b)$. Looking at the definition of $T(b)$ in Theorem 12, and noting that $X'Xb \in \operatorname{col}(X')$, we can deduce the following corollary.

Corollary 13. *Let $X \in \mathbb{R}^{n \times p}$ and $\lambda \in \mathbb{R}^p_{\geq 0}$ be given. There exist $y \in \mathbb{R}^n$ such that a corresponding Lasso solution selects model $\mathcal{M} \subseteq \{1, \ldots, p\}$ if and only if*

$$\operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}} \neq \varnothing,$$

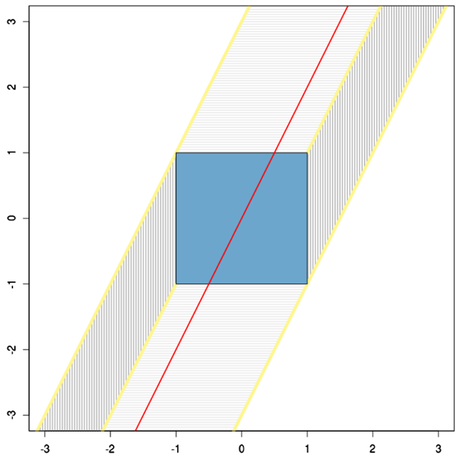


FIG 5. *The selection regions with respect to $X'y$ from Theorem 12, with $X = (1, 2)$ and $\lambda_1 = \lambda_2 = 1$ from Example 1. Displayed in red is $\operatorname{col}(X')$, the area on which the probability mass of $X'y$ is concentrated. The set $T\binom{0}{0}$ is displayed in blue, while the parallel light gray lines represent the sets $T\binom{0}{b_2}$ with $b_2 \neq 0$, and the parallel dark gray lines are the sets $T\binom{b_1}{0}$ with $b_1 \neq 0$. The yellow lines consist of the singletons $T\binom{b_1}{b_2}$ with $b_1, b_2 \neq 0$. Note that the red line does not intersect any of the sets $T\binom{b_1}{0}$ or $T\binom{b_1}{b_2}$ with $b_1, b_2 \neq 0$.*

where

$$\mathscr{B}_{\mathcal{M}} = \prod_{j=1}^{p} \begin{cases} \{-\lambda_j, \lambda_j\} & j \in \mathcal{M} \\ [-\lambda_j, \lambda_j] & j \notin \mathcal{M}, \end{cases}$$

which satisfies $\mathscr{B}_{\tilde{\mathcal{M}}} \subseteq \mathscr{B}_{\mathcal{M}}$ *for* $\mathcal{M} \subseteq \tilde{\mathcal{M}}$*.*

A model that may be chosen by the Lasso is called *accessible* in Sepehri and Harris (2017). (This reference who also provides a condition for when this is the case. The difference is that uses geometric considerations in $\mathbb{R}^n$ under a uniqueness assumption, whereas our approach operates in $\mathbb{R}^p$ with no assumptions on X.)

The sets $\mathscr{B}_{\mathcal{M}}$ are made up of the faces of the λ-cube. If $\mathcal{M}_0 = \varnothing$, $\mathscr{B}_{\varnothing}$ is the p-dimensional λ-box, $B_{\{j\}}$ is the union of two opposite facets of the λ-box, and for $1 < |\mathcal{M}| < p$, $\mathscr{B}_{\mathcal{M}}$ is a union of $(p - |\mathcal{M}|)$-dimensional faces of the λ-box. Finally, $\mathscr{B}_{\{1,\dots,p\}}$ simply contains the corners of the λ-box. These sets are illustrated in Figure 6 below.

For partial tuning with $\mathcal{M}_0 \neq \varnothing$, $\mathscr{B}_{\varnothing}$ is $(p - |\mathcal{M}_0|)$-dimensional and we have $\mathscr{B}_{\mathcal{M}} \subseteq \mathscr{B}_{\mathcal{M}_0}$ for all $\mathcal{M} \subseteq \{1, \dots, p\}$ as well as

$$\{0\} \subseteq \mathrm{col}(X') \cap \mathscr{B}_{\mathcal{M}_0} \neq \varnothing,$$

so that, not surprisingly, there always exist y such that the non-penalized components will be part of the model chosen by the Lasso solution.

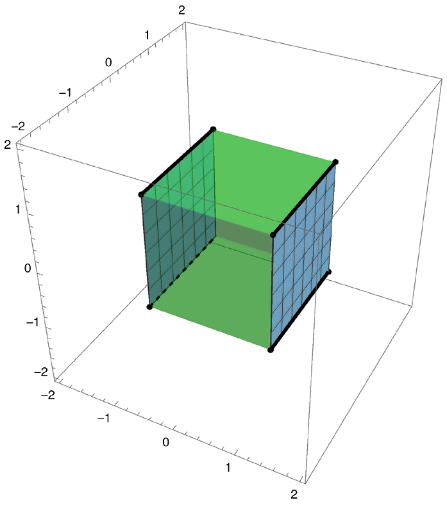


FIG 6. *Illustration of some of the sets* $\mathscr{B}_{\mathcal{M}}$ *for* $p = 3$ *and* $\lambda_1 = \lambda_2 = \lambda_3 = 1$*. The set* $\mathscr{B}_{\{1,2,3\}}$ *is given by the eight corners of the unit cube, whereas* $\mathscr{B}_{\{1,3\}}$ *contains the four parallel edges shown.* $\mathscr{B}_{\{1\}}$ *is the union of the two parallel 2-dimensional facets highlighted by a grid. The sets* $\mathscr{B}_{\{1,2\}}$ *and* $\mathscr{B}_{\{2,3\}}$ *are not depicted, but contain the four parallel vertical and horizontal edges of the cube, respectively. The sets* $\mathscr{B}_{\{2\}}$ *and* $\mathscr{B}_{\{3\}}$*, which are not shown either, each consist of the remaining two parallel vertical and horizontal facets.*

Example 1. *Suppose $X = (1, 2)$, so that $n = 1$, $p = 2$ and let $\lambda_1 = \lambda_2 = \lambda_3$ (uniform tuning). As can be learned from Figure 5,*

$$\operatorname{col}(X') \cap \mathscr{B}_{\{1\}} = \varnothing$$

for all $\bar{\lambda} > 0$, so that by Corollary 13, $\hat{\beta}_{\mathrm{L},1} = 0$ for any value of y, independent of $\mathcal{B}_0$ and σ^2. The distribution of the remaining component $\hat{\beta}_{\mathrm{L},2}$ is now given by a soft-thresholding rule (also see the paragraph following Theorem 3).

Example 2. *To look at a more complex example, suppose now that*

$$X = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix},$$

so that $n = 2$, $p = 3$. Let $\lambda_1 = \lambda_2 = \lambda_3 = \bar{\lambda}$ (uniform tuning). We have

$$\operatorname{col}(X') \cap \mathscr{B}_{\{3\}} = \varnothing$$

for all $\bar{\lambda} > 0$, so that by Corollary 13

$$P(\hat{\beta}_{\mathrm{L},3} = 0) = 1.$$

To say something about the distribution of the remaining components, note that the estimator is equivalent to the low-dimensional procedure using the matrix $\tilde{X} = X_{\{1,2\}}$, which contains the first and second regressor only. Let $\tilde{\beta} \in \mathbb{R}^2$ be such that $\tilde{X}\tilde{\beta} = X\beta$, where $X\beta = E(y)$, and let

$$V = (V_1, V_2)' \sim N(\tilde{\beta} - (\tilde{X}'\tilde{X})^{-1}\tilde{\lambda}, \sigma^2(\tilde{X}'\tilde{X})^{-1}),$$

where $\tilde{\lambda}$ will be specified below. We can now use Theorem 3 to find the following. The absolutely continuous parts of the distribution of $(\hat{\beta}_{\mathrm{L},1}, \hat{\beta}_{\mathrm{L},2})'$ can be determined by

$$\begin{aligned} P(\hat{\beta}_{\mathrm{L},1} \leq& z_1, \hat{\beta}_{\mathrm{L},2} \leq z_2, \hat{\beta}_{\mathrm{L},3} = 0) \\ &= \int_{-\infty}^{z_2 - \tilde{\beta}_2} \int_{-\infty}^{z_1 - \tilde{\beta}_1} \phi_{(0, \sigma^2 \tilde{X}'\tilde{X})}(\tilde{X}'\tilde{X}m + \tilde{\lambda})\, |\tilde{X}'\tilde{X}|\, dm_1 dm_2 \\ &= P(V_1 \leq z_1, V_2 \leq z_2) \end{aligned}$$

for $z_1, z_2 < 0$ and $\tilde{\lambda} = (-\bar{\lambda}, -\bar{\lambda})'$. Analogously, we get

$$P(\hat{\beta}_{\mathrm{L},1} \geq z_1, \hat{\beta}_{\mathrm{L},2} \geq z_2, \hat{\beta}_{\mathrm{L},3} = 0) = P(V_1 \geq z_1, V_2 \geq z_2)$$

for $z_1, z_2 > 0$ and $\tilde{\lambda} = (\bar{\lambda}, \bar{\lambda})'$. Moreover,

$$P(\hat{\beta}_{\mathrm{L},1} \leq z_1, \hat{\beta}_{\mathrm{L},2} \geq z_2, \hat{\beta}_{\mathrm{L},3} = 0) = P(V_1 \leq z_1, V_2 \geq z_2)$$

for $z_1 < 0, z_2 > 0$ and $\tilde{\lambda} = (-\bar{\lambda}, \bar{\lambda})'$, as well as

$$P(\hat{\beta}_{\mathrm{L},1} \geq z_1, \hat{\beta}_{\mathrm{L},2} \leq z_2, \hat{\beta}_{\mathrm{L},3} = 0) = P(V_1 \geq z_1, V_2 \leq z_2)$$

for $z_1 > 0, z_2 < 0$ and $\tilde{\lambda} = (\bar{\lambda}, -\bar{\lambda})'$.

This shows that the absolutely continuous parts of the estimator follow a normal distribution with the same covariance matrix as the LS estimator and a shift in expectation that depends on the regressor matrix as well as the tuning parameters. These findings are in line with Remark 3.

The pointmass part of $(\hat{\beta}_{\text{L},1}, \hat{\beta}_{\text{L},2})'$ at $(0,0)'$ has weight

$$P(\hat{\beta}_{\text{L},1} = \hat{\beta}_{\text{L},2} = \hat{\beta}_{\text{L},3} = 0) = \int_{-\bar{\lambda}}^{\bar{\lambda}} \int_{-\bar{\lambda}}^{\bar{\lambda}} \phi_{(\tilde{X}'\tilde{X}\tilde{\beta}, \sigma^2 \tilde{X}'\tilde{X})}(w) dw_1 dw_2.$$

For the remaining "mixed" terms, let

$$\tilde{X}'\tilde{X} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

and

$$Z = (Z_1, Z_2)' \sim N(\mu, \sigma^2 \Sigma) \text{ with } \Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

After some calculations, it can be shown that

$$\begin{aligned} &P(\hat{\beta}_{\text{L},1} = 0, \hat{\beta}_{\text{L},2} \leq z_2, \hat{\beta}_{\text{L},3} = 0) \\ &\quad = \int_{-\infty}^{z_2 - \tilde{\beta}_2} \int_{-\tilde{\lambda}}^{\tilde{\lambda}} \phi_{(0, \sigma^2 \tilde{X}'\tilde{X})}(\tilde{X}'\tilde{X}(-\tilde{\beta}_1, m_2)' + (s_1, -\bar{\lambda})')\, c\, ds_1 dm_2 \\ &\quad = P(-\bar{\lambda} \leq Z_1 \leq \bar{\lambda}, Z_2 \leq z_2) = P(-\bar{\lambda} \leq Z_1 \leq \bar{\lambda}) P(Z_2 \leq z_2) \end{aligned}$$

for $z_2 < 0$ and $\mu_1 = (|\tilde{X}'\tilde{X}|\tilde{\beta}_1 - b\bar{\lambda})/c$, $\sigma_1^2 = |\tilde{X}'\tilde{X}|/c$, $\mu_2 = (b\tilde{\beta}_1 + \bar{\lambda})/c + \beta_2$, and $\sigma_2^2 = 1/c$.

We get an analogous expression for $P(\hat{\beta}_{\text{L},1} = 0, \hat{\beta}_{\text{L},2} \geq z_2, \hat{\beta}_{\text{L},3} = 0)$ for $z_2 > 0$, but with $\bar{\lambda}$ replaced by $-\bar{\lambda}$ in μ_1 and μ_2. Moreover,

$$P(\hat{\beta}_{\text{L},1} \leq z_1, \hat{\beta}_{\text{L},2} = \hat{\beta}_{\text{L},3} = 0) = P(Z_1 \leq z_1) P(-\bar{\lambda} \leq Z_2 \leq \bar{\lambda})$$

for $z_1 < 0$ and $\mu_1 = \tilde{\beta}_1 + (b\tilde{\beta}_2 + \bar{\lambda})/a$, $\sigma_1^2 = 1/a$, $\mu_2 = (|\tilde{X}'\tilde{X}|\tilde{\beta}_2 - b\bar{\lambda})/a$, $\sigma_2^2 = |\tilde{X}'\tilde{X}|/a$.

Finally, we get an analogous expression for $P(\hat{\beta}_{\text{L},1} \geq z_1, \hat{\beta}_{\text{L},2} = \hat{\beta}_{\text{L},3} = 0)$ for $z_1 > 0$, but with $\bar{\lambda}$ replaced by $-\bar{\lambda}$ in μ_1 and μ_2. It might be interesting to note that in this example, the probabilities for the distribution that are made up of both a pointmass part and absolutely continuous parts can be represented by independent (normal) random variables.

In both examples above, the distribution of $\hat{\beta}_{\text{L}}$ is the same as the one of a Lasso estimator in a smaller model. This fact is, of course, only valid for the specific forms of X and λ considered here. The models considered by the Lasso do not depend on β and ε in the sense that certain values of X and λ may immediately rule out certain models, completely independently of y. (The choice between the accessible models does, of course, very much depend on β and ε.)

Sparked by Examples 1 and 2, this suggests that in the high-dimensional setting, model selection by the Lasso estimator may possibly not be a purely data-driven procedure insofar as there is a *structural model* or *structural set* $\mathscr{M} \subseteq \{1,\dots,p\}$, determined by X and λ only, that satisfies $\hat\beta_{\mathrm{L},j} = 0$ for any $j \notin \mathscr{M}$ *for all observations* $y \in \mathbb{R}^n$. In particular, the true parameter $\beta \in \mathcal{B}_0$, as well as the distribution of ε do not have any influence on this set. In other words, some models are never considered by the model selection procedure, completely independently of the data vector y. Put yet differently again, for a given regressor matrix X, one can restrict or choose this class of models by choice of λ.

Given all the considerations above, one might ask whether such a structural model $\mathscr{M}$ always satisfies $|\mathscr{M}| \leq n$ under certain conditions. Clearly, uniqueness would be a meaningful requirement in this context, as then all Lasso solutions will choose models of cardinality of at most n, as has been shown in Tibshirani (2013).[2] In that case, the Lasso estimator would be equivalent to a low-dimensional Lasso procedure, restricted to this structural model $\mathscr{M}$, and we could employ results from low-dimensional settings also for inference in high-dimensional models, such as Ewald and Schneider (2018) for constructing confidence regions.

In Examples 1 and 2, the Lasso solutions are always unique. It is not difficult, however, to construct an example where the solutions are not unique anymore.

Example 3. *Again, take the model from Example 1 with $X = (1,2)$. This time, choose $\lambda = (1,2)'$ (non-uniform tuning). It can easily be seen using Theorem 9 that for each $y < -1$,*

$$\hat\beta_{\mathrm{L}} = \begin{pmatrix} y+1 \\ 0 \end{pmatrix}, \ \hat\beta_{\mathrm{L}} = \begin{pmatrix} 0 \\ \frac{y+1}{2} \end{pmatrix}, \quad \textit{and} \ \ \hat\beta_{\mathrm{L}} = \begin{pmatrix} y+1-2c \\ c \end{pmatrix} \quad \textit{for} \ \ (y+1)/2 < c < 0$$

all are Lasso solutions for the same value of y. Similarly, for $y > 1$,

$$\hat\beta_{\mathrm{L}} = \begin{pmatrix} y-1 \\ 0 \end{pmatrix}, \ \hat\beta_{\mathrm{L}} = \begin{pmatrix} 0 \\ \frac{y-1}{2} \end{pmatrix}, \quad \textit{and} \ \ \hat\beta_{\mathrm{L}} = \begin{pmatrix} y+1-2c \\ c \end{pmatrix} \quad \textit{for} \ \ 0 < c < (y+1)/2$$

all are Lasso solutions for the same value of y. (Note that $\hat\beta_{\mathrm{L}} = 0$ for all y with $|y| \leq 1$.) The corresponding selection regions are illustrated in Figure 7.

Example 3 shows an already known property of the Lasso from another perspective: The solution to the Lasso problem is, in general, not unique. Moreover, if the solution is not unique, then, by convexity of the problem, there exists an uncountable set of solutions.[3] The example moreover shows that the set of y which yield non-unique Lasso solutions is not a null set. In fact, in this example, it occurs with probability $2\Phi(-1)$.

Of course, this problem could be overcome by slightly altering the choice of the tuning parameters, even though this would imply to make a choice of the class of models under consideration, as pointed out previously in this section.

[2] Note that this fact alone does not imply that the structural set has cardinality of at most n since the active sets may certainly vary over y.

[3] This fact has been pointed out by Tibshirani (2013) in Lemma 1 for the case of uniform tuning.

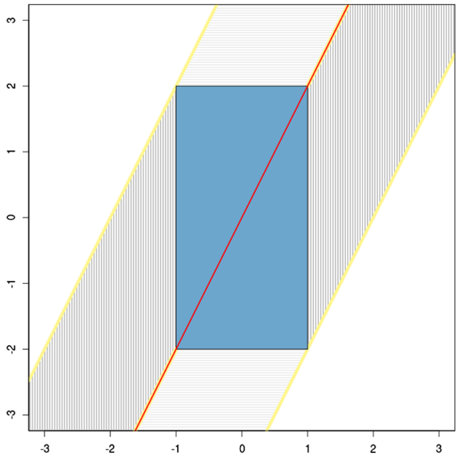


Fig 7. *The selection regions with respect to $X'y$ from Theorem 12, with $X = (1, 2)$ and $\lambda_1 = 1$, $\lambda_2 = 2$ from Example 3. Displayed in red is* $\mathrm{col}(X')$*, the area on which the probability mass of $X'y$ is concentrated. The set $T\binom{0}{0}$ is displayed in blue, while the parallel light gray lines represent $T\binom{0}{b_2}$ with $b_2 \neq 0$, and the parallel dark gray lines are $T\binom{b_1}{0}$ with $b_1 \neq 0$. The yellow lines consist of the singletons $T\binom{b_1}{b_2}$ with $b_1, b_2 \neq 0$. The red line passes through $T\binom{0}{0}$ where the solution is unique but also through the line where the light gray, the dark gray and the yellow areas intersect.*

4.2. Structural sets

Clearly, Example 3 shows that the structural set may be equal to the entire set of explanatory variables. It is easy to see that for $n = 1$ and $p = 2$, the Lasso estimator will always have a structural set with cardinality $n = 1$ whenever we have uniqueness. The question is, of course, whether the same can be said in more generality. Before answering this question, we show how the structural set can be determined given X and λ, by counting how many sets of parallel facets of the λ-box $\mathscr{B}_\varnothing$ are intersected by $\mathrm{col}(X')$.

Theorem 14. *Let $X \in \mathbb{R}^{n\times p}$ and $\lambda \in \mathbb{R}^p_{\geq 0}$ be given. Let $\mathscr{M}$ be the so-called structural set of X and λ that contains all $j \in \{1, \dots, p\}$, such that there exist $y \in \mathbb{R}^n$ so that a corresponding Lasso solution $\hat\beta_{\mathrm{L}}$ satisfies $\hat\beta_{\mathrm{L},j} \neq 0$, that is, $\mathscr{M}$ contains all regressors that are part of a Lasso solution for some observation y. This set is given by*

$$\mathscr{M} = \mathscr{M}(X,\lambda) = \big\{j \in \{1,\dots,p\} : \mathrm{col}(X') \cap \mathscr{B}_{\{j\}} \neq \varnothing\big\}.$$

Proof. By Corollary 13, there exist $y \in \mathbb{R}^n$ such that the corresponding Lasso solution chooses model $\mathcal{M}$ if and only if $\mathrm{col}(X') \cap B_{\mathcal{M}} \neq \varnothing$. For any $\{j\} \subseteq \mathcal{M}$, we have, by Corollary 13 also, $\mathscr{B}_{\mathcal{M}} \subseteq \mathscr{B}_{\{j\}}$, so that $\mathscr{B}_{\{j\}} \cap \mathrm{col}(X') \neq \varnothing$ also. □

Theorem 14 shows that in order to determine the structural set, only the intersection of $\mathrm{col}(X')$ with the $(p-1)$-dimensional faces, the so-called facets of the λ-cube, have to be considered. A strategy how to determine the structural set for a given X might be the following. Note that $\mathrm{col}(X') = \ker(X)^\perp$ and find vectors $V_1, \dots, V_k \in \mathbb{R}^p$ that span $\ker(X)$, where $k = p - \mathrm{rk}(X)$. Let $V = (V_1, \dots, V_k) \in \mathbb{R}^{p\times k}$ and check whether $V's = 0$ is solvable for $s \in \mathscr{B}_{\{j\}}$, where

$$\mathscr{B}_{\{j\}} = [-\lambda_1, \lambda_1] \times \cdots \times \{-\lambda_j, \lambda_j\} \times \cdots \times [-\lambda_p, \lambda_p].$$

If this is the case, then $j \in \mathscr{M}$, otherwise $j \notin \mathscr{M}$. So, determining the structural set amounts to identifying a basis of $\ker(X)$, and solving a linear system in $k = p - \mathrm{rk}(X)$ equations and p unknowns. After that, we have to check whether the resulting solution set contains any elements of $\mathscr{B}_{\{j\}}$ for $j = 1, \dots, p$. This approach is employed in Example 5 in the subsequent section.

We would like to point out the difference between the idea of a structural set and results concerning SAFE rules, such as Ghaoui, Viallon and Rabbani (2012), Tibshirani et al. (2012) and Ndiayee et al. (2017). Based on a SAFE rule, a regressor will be discarded by a Lasso solution for a given observation y. In contrast, if a covariate is not contained the the structural set, it will be excluded from the Lasso model for all observations y. This, on the one hand, implies that the result of Theorem 14 is much cruder than a safe rule, excluding (if any) less regressors. On the other hand, since the structural set is entirely independent of y, the corresponding Lasso problem can equivalently be viewed as a Lasso problem using covariates from $\mathscr{M}$ only, also regarding distributional results and inference. In particular, if $|\mathscr{M}| \leq n$, we can consider the low-dimensional Lasso problem using $X_{\mathscr{M}}$ as regressor matrix. If $X_{\mathscr{M}}$ has full rank, we can then use results from Ewald and Schneider (2018) to construct confidence regions. One has, of course, to be aware that inference is now on the parameter $\tilde{\beta}$ satisfying $X\beta = X_{\mathscr{M}}\tilde{\beta}$, as exemplified in Example 2.

Remark 6. *As indicated in Theorem 14 and as discussed above, the structural set $\mathscr{M}$ depends on X and λ only. Moreover, it can easily be seen that it depends on the tuning parameters λ only through the penalization weighting in the sense that whenever $\lambda = \bar{\lambda}\omega$ for some $\bar{\lambda} > 0$ and $\omega \in \mathbb{R}^p_{\geq 0}$, $\mathscr{M}(X, \lambda) = \mathscr{M}(X, \omega)$ follows. This implies that, in particular, in the common case of uniform tuning with $\bar{\lambda} = \lambda_1 = \cdots = \lambda_p$, the structural set* only depends on *X!*

Coming back to the conjecture whether the structural set always satisfies $|\mathscr{M}| \leq \min\{n, p\}$ in case the solutions are unique, using Theorem 14, we can list the following simple example with $n = 2$ and $p = 3$ to show that this cannot be the case in general. However, note that Theorem 14 allows to compute the structural set and that whenever $|\mathscr{M}| \leq n$, the resulting Lasso estimator is, in fact, just equivalent to a low-dimensional procedure.

Example 4. *Let*

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

and $\lambda = (1, 1, 1)'$. Then the structural set is clearly given by

$$\mathscr{M} = \{1, 2, 3\},$$

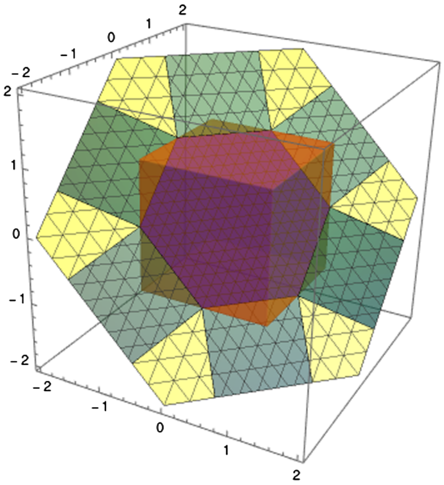


FIG 8. *The intersection of* $\mathrm{col}(X')$ *(in gray and yellow) with the unit cube (in orange) for* $\lambda_1 = \lambda_2 = \lambda_3 = 1$ *and* $X = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$*, see Example 4. The upper left edge is contained in* $\mathscr{B}_{\{1,2\}}$*, whereas the upper back edge is contained in* $\mathscr{B}_{\{2,3\}}$*. (Each* $\mathscr{B}_{\{i,j\}}$ *contains four parallel edges.) To view this figure in terms of selection regions, note that the areas corresponding to single-regressor models are displayed in gray, while the selection regions that correspond to two-regressor models are displayed in yellow. The intersection of the* λ*-cube with* $\mathrm{col}(X')$*, which corresponds to the zero estimator, is displayed in blue.*

as $(1,1,0)' \in \mathrm{col}(X') \cap \mathscr{B}_{\{1,2\}}$ *and* $(0,1,1)' \in \mathrm{col}(X') \cap \mathscr{B}_{\{2,3\}}$*, and* $\mathscr{B}_{\mathcal{M}} \subseteq \mathscr{B}_{\{j\}}$ *for any* $j \in \mathcal{M}$ *by Corollary 13, see Figure 8 for illustration. Yet the Lasso solutions for this* X *and* λ *are always unique which can be checked on the basis of Theorem 15 in the subsequent section.*

Finally, it is important to note that Theorem 14 also reveals that $\mathscr{M}$ contains all regressors if the columns of X are scaled to have unit length and the components are tuned uniformly: For $s = \bar{\lambda} X' X_j \in \mathrm{col}(X')$, by the Cauchy-Schwarz inequality, we have $s \in \mathscr{B}_{\{j\}}$ also, leading to $j \in \mathscr{M}$.

Example 2 (continued)**.** *If we rescale the columns of* X *from Example 2, we obtain*

$$\tilde{X} = \begin{pmatrix} \frac{2}{\sqrt{5}} & 0 & 1 \\ \frac{1}{\sqrt{5}} & 1 & 0 \end{pmatrix}.$$

Since the first row is an element of $\mathscr{B}_{\{3\}}$*, and the second row is an element of* $\mathscr{B}_{\{2\}}$*, clearly* $\{2,3\} \subseteq \mathscr{M}$*. To see that* $\{1\} \subseteq \mathscr{M}$ *also, note that the first row plus* $\sqrt{5} - 2$ *times the second row lies in* $\mathscr{B}_{\{1\}}$ *– yielding a full structural set.*

The above observation may be seen as an argument for rescaling the regressors before using the Lasso. It may, however, not always be desirable to so, such as in case the explanatory variables are observed in the same units, or in the presence of dummy variables. Also, rescaling the columns may result in changing whether or not the solutions are unique, an issue addressed in the following section.

4.3. A necessary and sufficient condition for uniqueness

We now turn to some results revolving around uniqueness of the Lasso estimator, which can be obtained with the same geometric approach, that is, studying the intersection of the λ-cube with $\operatorname{col}(X')$. Note that by uniqueness, we mean that for a given $X \in \mathbb{R}^{n \times p}$, and a given $\lambda \in \mathbb{R}^p$, the Lasso solutions are unique for all observations $y \in \mathbb{R}^n$.

Tibshirani (2013) showed that for a given regressor matrix X, Lasso solutions are unique in the above sense, if the columns of X are *in general position*,[4] which occurs when no k-dimensional affine[5] subspace for $k < \min(n,p)$ contains more than $k+1$ elements of the set $\{\pm X_1, \ldots, \pm X_p\}$, excluding antipodal pairs (see p. 1463 in Tibshirani, 2013). In fact, the solutions are then unique for all choices of the tuning parameter, provided that all components are tuned equally. As this condition is sufficient, one may ask whether it is also necessary. The answer to this question is, in fact, no, as can easily be seen from the example below.

When can non-unique solutions exist? For a given $X \in \mathbb{R}^{n \times p}$ and $\lambda \in \mathbb{R}^p$, this occurs if and only if there exist $b, \tilde{b} \in \mathbb{R}^p$ with $b \neq \tilde{b}$ and

$$\operatorname{col}(X') \cap T(b) \cap T(\tilde{b}) \neq \varnothing.$$

More concretely, by Theorem 12, and since the Lasso fit Xb is always unique,[6] this means that

$$X'Xb + v = X'X\tilde{b} + v,$$

where $v \in \operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}}$ for some $\mathcal{M} \subseteq \{1, \ldots, p\}$, and $\tilde{b}_{\mathcal{M}^c} = b_{\mathcal{M}^c} = 0$. Moreover, for $j \in \mathcal{M} \backslash \mathcal{M}_0$, we have $\operatorname{sgn}(b_j) = \operatorname{sgn}(v_j)$ whenever $b_j \neq 0$, as well as $\operatorname{sgn}(\tilde{b}_j) = \operatorname{sgn}(v_j)$ whenever $\tilde{b}_j \neq 0$. Note that we therefore have $Xb = X_{\mathcal{M}} b_{\mathcal{M}} = X_{\mathcal{M}} \tilde{b}_{\mathcal{M}} = X\tilde{b}$, implying that the columns of $X_{\mathcal{M}}$ must be linearly dependent. So non-uniqueness occurs only if $\operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}} \neq \varnothing$ for $\mathcal{M} \subseteq \{1, \ldots, p\}$ with linearly dependent columns in $X_{\mathcal{M}}$. The following example now immediately shows that the columns of X being in general position is not necessary for uniqueness.

Example 5. *Let*

$$X = \begin{pmatrix} 1 & 1 & 2 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix}.$$

Clearly, the columns are not general position, however, all Lasso solutions are unique for any choice of the tuning parameter, when the components are tuned uniformly: We have $\operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}} = \varnothing$ whenever $\{1,2\} \subseteq \mathcal{M}$ or $|\mathcal{M}| > 2$. This can easily be checked using the fact that $v \in \operatorname{col}(X')$ if and only if $v'w_1 =$

[4] Note that *general position* does not mean that any selection of n columns of X is linearly independent, as has sometimes been suggested in the literature, these two concepts are in fact unrelated.

[5] In Tibshirani (2013), the word "affine" is missing, which has caused some confusion in the literature.

[6] This has been shown in Lemma 1 in Tibshirani (2013) for uniform tuning and can easily be extended to non-uniform and partial tuning.

$v'w_2 = 0$ for $\ker(X) = \mathrm{col}\{w_1, w_2\}$. Therefore, the columns of $X_{\mathcal{M}}$ are linearly independent for any $\mathcal{M}$ that can be chosen by the Lasso, and all Lasso solutions must be unique.

The example above illustrates the commonly known fact that, if a Lasso solution is unique, it will contain at most n non-zero entries. We show that this fact can be sharpened to yield a necessary and sufficient condition for uniqueness of all Lasso solutions in the following way: first, we show that if the solution is unique, it in fact has at most $\mathrm{rk}(X) \leq n$ non-zero components. Second, we prove that this is not only a necessary, but also a sufficient criterion for uniqueness.

Theorem 15 (Uniqueness)**.** *Let $X \in \mathbb{R}^{n\times p}$ and $\lambda \in \mathbb{R}^p_{\geq 0}$. The Lasso solution is unique for all $y \in \mathbb{R}^n$ if and only if*

$$\mathrm{col}(X') \cap \mathscr{B}_{\mathcal{M}} = \varnothing \textit{ for all } \mathcal{M} \subseteq \{1, \dots, p\} \textit{ with } |\mathcal{M}| > \mathrm{rk}(X).$$

Proof. ($\Longrightarrow$) Assume the condition is not satisfied. Then there exists $v \in \mathscr{B}_{\mathcal{M}}$ with $|\mathcal{M}| > \mathrm{rk}(X)$ and $v = X'z$ for some $z \in \mathbb{R}^n$. We show that there is a $y \in \mathbb{R}^n$ such that the corresponding Lasso problem is not uniquely solvable.

If $X_j = 0$ for some $j \in \mathcal{M}_0$, we are done as the corresponding coefficient may be arbitrary. Note that $X_j = 0$ for $j \in \mathcal{M}\setminus\mathcal{M}_0$ is not possible: since $v \in \mathrm{col}(X')$, this would imply $v_j = 0$, but that contradicts $v \in \mathscr{B}_{\mathcal{M}}$. We therefore assume that $X_j \neq 0$ for all $j \in \mathcal{M}$.

Since $|\mathcal{M}| > \mathrm{rk}(X)$, there is a column of $X_{\mathcal{M}}$, say X_j ($X_j \neq 0$), that can be written as a linear combination of the other columns. In particular, we can write

$$dX_j = \sum_{l\in\mathcal{M}\setminus\{j\}} c_l X_l,$$

where $d = \mathrm{sgn}(v_j)$ if $\lambda_j \neq 0$ and $d = 1$ if $\lambda_j = 0$. Moreover, let $c = \max_{l\in\mathcal{M}\setminus\{j\}} |c_l| > 0$. Define $b \in \mathbb{R}^p$ by

$$b_l = \begin{cases} \frac{d}{2c} & l = j \\ \mathrm{sgn}(v_l) & l \in \mathcal{M} \setminus \{j\} \\ 0 & l \notin \mathcal{M}. \end{cases}$$

Then b is a Lasso solution for $y = z + Xb$ since

$$X'y = X'Xb + X'z = X'Xb + v \in T(b).$$

We now construct $\tilde{b} \in \mathbb{R}^p$, with $\tilde{b} \neq b$, that is also a Lasso solution for the same y by

$$\tilde{b}_l = \begin{cases} \mathrm{sgn}(v_l) + \frac{c_l}{2c} & l \in \mathcal{M} \setminus \{j\} \\ 0 & l = j \text{ or } l \notin \mathcal{M}. \end{cases}$$

Clearly, $b \neq \tilde{b}$, $\mathrm{sgn}(b_l) = \mathrm{sgn}(\tilde{b}_l) = \mathrm{sgn}(v_j)$ for $l \in \mathcal{M} \setminus \{j\}$ and

$$Xb = \sum_{l\in\mathcal{M}} b_l X_l = \sum_{l\in\mathcal{M}\setminus\{j\}} b_l X_l + \frac{d}{2c} X_j = \sum_{l\in\mathcal{M}\setminus\{j\}} (b_l + \frac{c_l}{2c}) X_l = X\tilde{b}.$$

We therefore get

$$X'y = X'Xb + v = X'X\tilde{b} + v \in S(\tilde{b})$$

also, implying that both b and $\tilde{b}$ are Lasso solutions for the given y.

($\Longleftarrow$) We now prove the other direction. Assume that there exists $y \in \mathbb{R}^n$ such that non-unique Lasso solutions $b \neq \tilde{b}$ exist. As discussed above, this implies the existence of $v \in \operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}}$ for some $\mathcal{M} \subseteq \{1, \dots, p\}$ with $X_{\mathcal{M}} b_{\mathcal{M}} = X_{\mathcal{M}} \tilde{b}_M$ and $b_{\mathcal{M}^c} = \tilde{b}_{\mathcal{M}^c} = 0$, entailing that the columns of $X_{\mathcal{M}}$ are linearly dependent.

If $|\mathcal{M}| > \operatorname{rk}(X)$, we are done. If $|\mathcal{M}| \leq \operatorname{rk}(X)$, we do the following. Since we have $\operatorname{rk}(X_{\mathcal{M}}) < |\mathcal{M}| \leq \operatorname{rk}(X)$, we can pick $z \in \mathbb{R}^n$ such that $z \in \operatorname{col}(X_{\mathcal{M}})^{\perp} \setminus \operatorname{col}(X_{\mathcal{M}^c})^{\perp}$. This is possible since

$$\begin{aligned} &\operatorname{col}(X_{\mathcal{M}})^{\perp} \setminus \operatorname{col}(X_{\mathcal{M}^c})^{\perp} = \varnothing \iff \operatorname{col}(X_{\mathcal{M}})^{\perp} \subseteq \operatorname{col}(X_{\mathcal{M}^c})^{\perp} \\ &\quad \iff \operatorname{col}(X_{\mathcal{M}^c}) \subseteq \operatorname{col}(X_{\mathcal{M}}) \iff \operatorname{col}(X_{\mathcal{M}}) = \operatorname{col}(X_{\mathcal{M}}, X_{\mathcal{M}^c}) = \operatorname{col}(X) \\ &\quad \iff \operatorname{rk}(X_{\mathcal{M}}) = \operatorname{rk}(X), \end{aligned}$$

which is not the case. This z satisfies $(X'z)_{\mathcal{M}} = (X_{\mathcal{M}})'z = 0$ and $(X'z)_{\mathcal{M}^c} = (X_{\mathcal{M}^c})'z \neq 0$, so that we can find $c \in \mathbb{R}$ such that

$$\tilde{v} = v + c\,X'z \in \mathscr{B}_{\tilde{\mathcal{M}}} \cap \operatorname{col}(X'),$$

with $\mathcal{M} \subseteq \tilde{\mathcal{M}}$ and $|\mathcal{M}| < |\tilde{\mathcal{M}}|$. As long as $|\tilde{\mathcal{M}}| \leq \operatorname{rk}(X)$, repeat the steps above with $v = \tilde{v}$ and $\mathcal{M} = \tilde{\mathcal{M}}$. □

Note that just as for Theorem 14, the result from the above theorem depends on λ only through the penalization weights, meaning that for any $\mathcal{M} \subseteq \{1, \dots, p\}$, whenever $\lambda = \bar{\lambda}\omega$ for some $\bar{\lambda} > 0$ and $\omega \in \mathbb{R}^p_{\geq 0}$, we have $\operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}}(\lambda) = \varnothing$ if and only if $\operatorname{col}(X') \cap \mathscr{B}_{\mathcal{M}}(\omega) \neq \varnothing$ (when indicating the dependence of $\mathscr{B}_{\mathcal{M}}$ on the tuning parameters).

As mentioned in the preamble of Section 4, Theorem 15 does not require $p > n$, so that it also covers the low-dimensional case. Clearly, the condition for uniqueness is trivially satisfied if $\operatorname{rk}(X) = p$.

5. Conclusion

We give explicit formulae regarding the distribution of the Lasso estimator in finite-samples, assuming a Gaussian distribution of $X'\varepsilon$. In the low-dimensional case, we consider the cdf as well as the density functions conditional on "active sets" of the estimator. Our results exploit the structure of the underlying optimization problem of the Lasso estimator and do not hinge on the normality assumption. We also explicitly characterize the correspondence between the Lasso and the LS estimator: It is shown that the Lasso estimator essentially creates shrinkage areas around the axes inside which the probability mass of the LS estimator is compressed into lower-dimensional densities that can be specified conditional on the active set of the estimator. As a result, the distribution looks like a pieced-together combination of Gaussian-like densities. Each active set has its own distributional piece with dimension depending on the number of nonzero components, resulting also in point mass at the origin and mass being distributed along the axes.

The form of the distribution is even more intricate in the high-dimensional case, in which the estimator may not be unique anymore. We quantify the relationship between a Lasso solution and the quantity $X'y$ (rather than the LS estimator as in the low-dimensional case). We gain valuable insights into the behavior of the estimator by illustrating that some models may never be selected by the estimator: The so-called structural set, that contains all covariates that are part of a Lasso solution for *some* response vector y, can be computed based on a geometric condition involving the regressor matrix and penalization weights only. In case this structural set has cardinality less than or equal to n, the Lasso is equivalent to a low-dimensional procedure and results from the $p \leq n$-framework can be used for inference. We also learn that in case of uniform tuning and the columns of X scaled to unit length, the structural set contains all covariates.

Finally, the previous insights allow us to close a gap in the literature by providing a condition for uniqueness of the Lasso estimator that is both necessary and sufficient.

Acknowledgements

The authors gratefully acknowledge support from the Deutsche Forschungsgemeinschaft (DFG) through grant FOR 916 and thank Thomas Hack and Nikos Dafnis for very insightful discussions on the geometric aspects of this paper. We also thank two anonymous referees for their comments.

References

Ali, A. and Tibshirani, R. J. (2019). The Generalized Lasso Problem and Uniqueness. *Electronic Journal of Statistics* **13** 2307–2347. MR3980959

Ewald, K. and Schneider, U. (2018). Uniformly Valid Confidence Sets Based on the Lasso. *Electronic Journal of Statistics* **12** 1358–1387. MR3802261

Ghaoui, L. E., Viallon, V. and Rabbani, T. (2012). Safe Feature Elimination in Sparse Supervised Learning. *Pacific Journal of Optimization* **8** 667–698. MR3026449

Jagannath, R. and Upadhye, N. S. (2018). The Lasso Estimator: Distributional Properties *Kybernetica (Prague)* **54** 778–797. MR3863256

Knight, K. and Fu, W. (2000). Asymptotics of Lasso-Type Estimators. *Annals of Statistics* **28** 1356–1378. MR1805787

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact Post-Selection Inference with an Application to the Lasso. *Annals of Statistics* **44** 907–927. MR3485948

NDIAYEE, E., FERCOQ, O., GRAMFORT, A. and SALMON, J. (2017). Gap Safe Screening Rules for Sparsity Enforcing Penalties. *Journal of Machine Learning Research* **18** 1–33. MR3763762

PÖTSCHER, B. M. and LEEB, H. (2009). On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding. *Journal of Multivariate Analysis* **100** 2065–2082. MR2543087

PÖTSCHER, B. M. and SCHNEIDER, U. (2009). On the Distribution of the Adaptive LASSO Estimator. *Journal of Statistical Planning and Inference* **139** 2775–2790. MR2523666

SEPEHRI, A. and HARRIS, N. (2017). The Accessible Lasso Models. *Statistics* **51** 711–721. MR3669285

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2013). The Lasso Problem and Uniqueness. *Electronic Journal of Statistics* **7** 1456–1490. MR3066375

TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics* **40** 1198–1232. MR2985948

TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2012). Strong Rules for Discarding Predictors in Lasso-Type Problems. *Journal of the Royal Statistical Society Series B* **74** 245–266. MR2899862

ZHOU, Q. (2014). Monte Carlo Simulation for Lasso-Type Problems by Estimator Augmentation. *Journal of the American Statistical Association* **109** 1495–1516. MR3293606