

On the Distribution, Model Selection Properties and Uniqueness of the Lasso Estimator in Low and High Dimensions

Ulrike Schneider and Karl Ewald

Vienna University of Technology

We derive expressions for the finite-sample distribution of the Lasso estimator in the context of a linear regression model with normally distributed errors 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. 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.

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 & 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 & 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 & Schneider (2015), 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 & Fu (2000). Moreover, Jagannath & Upadhye (2016) 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. We also exactly quantify the correspondence between the Lasso and least-squares (LS) estimator. 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 for some response vector. This structural set can be identified by how the row space of the regressor matrix intersects the cube 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. 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. All our results are based on properties of the optimization problem defining the estimator and do not hinge on the assumption of Gaussian errors.

The paper is organized as follows. We introduce setting and notation in Section 2. 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 and Notation

Consider the linear model

$$y = X\beta + \varepsilon, \quad (1)$$

where y is the observed $n \times 1$ data vector, X 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 with independent and identically distributed components that follow a $N(0, \sigma^2)$ -distribution. We consider the *weighted Lasso estimator* $\hat{\beta}_L$, defined as the unique solution to the minimization problem

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

where $\lambda_{n,j}$, are non-negative user-specified tuning parameters that will typically depend on n . To ease notation, however, 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 estimator is equal to the ordinary LS estimator $\hat{\beta}_{LS}$ and that $\lambda_1 = \dots = \lambda_p > 0$ 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 $\lambda = (\lambda_1, \dots, \lambda_p)'$ and define $\mathcal{M}_0 = \{j : \lambda_j = 0\}$, the index set of all unpenalized coefficients. If $\mathcal{M}_0 \neq \emptyset$, we speak of partial tuning. 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 e_j denote the j -th unit vector in \mathbb{R}^p and let $\phi_{(\mu, \Sigma)}$ denote the Lebesgue-density of a normally distributed random variable with mean μ and covariance matrix Σ and Φ be the cdf of a univariate standard normal distribution. For a vector $m \in \mathbb{R}^p$ and an index set $I \subseteq \{1, \dots, p\}$, the vector $m_I \in \mathbb{R}^{|I|}$ contains only the components of m corresponding to the elements of I and we write $|I|$ for the cardinality of I and I^c for $\{1, \dots, p\} \setminus I$, 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 $\text{sgn}(x) = \mathbf{1}_{\{x>0\}} - \mathbf{1}_{\{x<0\}}$ where $\mathbf{1}$ 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 column space and rank of a matrix C by $\text{col}(C)$ and $\text{rk}(C)$, 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 the indices in I only. We use $\mathbb{R}_{>0}$ for the positive and $\mathbb{R}_{\geq 0}$ for the non-negative real numbers.

Let $\{D_-, D_+, D_0\}$ 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 $d_j = \mathbf{1}_{\{j \in D_+\}} - \mathbf{1}_{\{j \in D_-\}}$. For such d , we denote by $\mathcal{O}^d = \{z \in \mathbb{R}^p : \text{sgn}(z_j) = d_j \text{ for } j = 1, \dots, p\} = \{z \in \mathbb{R}^p : z_j < 0 \text{ for } j \in D_-, z_j > 0 \text{ for } j \in D_+, z_j = 0 \text{ for } j \in D_0\}$. 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 $j \in D_+$ and $m_j = -\beta_j$ for $j \in D_0$. We write D_\pm^\dagger for $D_- \cup D_+$.

Finally, the directional derivative of a function $g : \mathbb{R}^p \rightarrow \mathbb{R}$ at m in direction $r \in \mathbb{R}^p$ with $\|r\| = 1$ is defined as

$$\frac{\partial g(m)}{\partial r} = \lim_{h \searrow 0} \frac{g(m + hr) - g(m)}{h}.$$

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 $p \leq n$. For our arguments, we use a reparametrized version of the objective function. Define

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

where $W = X'\varepsilon \sim N(0, \sigma^2 X'X)$ and note that V_β is minimized at $\hat{u} = \hat{\beta}_L - \beta$. We are interested in the distribution of the estimation error \hat{u} . To state the main theorem, we need the following lemma that characterizes a solution to the minimization problem.

Lemma 1. *Let $m \in \mathbb{R}^p$. The following two statements are equivalent.*

- (a) $\frac{\partial V_\beta(m)}{\partial r} \geq 0 \quad \forall r$ with $\|r\| = 1$
- (b) $\frac{\partial V_\beta(m)}{\partial e_j} \geq 0$ and $\frac{\partial V_\beta(m)}{\partial(-e_j)} \geq 0$ for $j = 1, \dots, p$.

Proof. Only (b) \Rightarrow (a) needs to be proved. Let $d \in \{-1, 0, 1\}^p$ such that $m + \beta \in \mathcal{O}^d$ and let $\{D_-, D_+, D_0\}$ be the corresponding partition of $\{1, \dots, p\}$. A straight-forward calculation shows that

$$\begin{aligned} \frac{\partial V_\beta(m)}{\partial r} &= 2r'X'Xm - 2r'W + 2 \sum_{j=1}^p \lambda_j (-\mathbf{1}_{\{j \in D_-\}} r_j + \mathbf{1}_{\{j \in D_+\}} r_j + \mathbf{1}_{\{j \in D_0\}} |r_j|) \\ &= \sum_{j=1}^p \mathbf{1}_{\{r_j \geq 0\}} [(2X'Xm - 2W)_j + 2\lambda_j (-\mathbf{1}_{\{j \in D_-\}} + \mathbf{1}_{\{j \in D_+ \cup D_0\}})] r_j \\ &\quad + \mathbf{1}_{\{r_j < 0\}} [-(2X'Xm - 2W)_j + 2\lambda_j (\mathbf{1}_{\{j \in D_- \cup D_0\}} - \mathbf{1}_{\{j \in D_+\}})] (-r_j) \\ &= \sum_{j=1}^p \mathbf{1}_{\{r_j \geq 0\}} \frac{\partial V_\beta(m)}{\partial e_j} r_j + \sum_{j=1}^p \mathbf{1}_{\{r_j < 0\}} \frac{\partial V_\beta(m)}{\partial(-e_j)} (-r_j) \geq 0. \end{aligned}$$

□

This complete characterization of minima of V_β now allows to state the following theorem on the distribution of the estimation error $\hat{u} = \hat{\beta}_L - \beta$.

Theorem 2. *Let $z \in \mathbb{R}^p$. Let $d = \text{sgn}(z + \beta) \in \{-1, 0, 1\}^p$ and let $\{D_-, D_+, D_0\}$ be the corresponding partition of $\{1, \dots, p\}$. Then*

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

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

Proof. Since the function V_β is convex, $m \in \mathbb{R}^p$ is a minimizer of V_β if and only if $\frac{\partial V_\beta(m)}{\partial r} \geq 0$ for all $r \in \mathbb{R}$ with $\|r\| = 1$. We wish to find all minimizers m satisfying $m_j \leq z_j$ for $j \in D_-$, $m_j \geq z_j$ for $j \in D_+$ and $m_j = z_j$ for $j \in D_0$. Note that this implies that $m + \beta \in \mathcal{O}^d$ since $z + \beta \in \mathcal{O}$ by assumption. By Lemma 1 together with the fact that the condition $\frac{\partial V_\beta(m)}{\partial e_j} \geq 0$ and $\frac{\partial V_\beta(m)}{\partial(-e_j)} \geq 0$

reduces to $\frac{\partial V_\beta(m)}{\partial u_j} = 0$ if V_β is differentiable at m with respect to the j -th component, we get the following necessary and sufficient conditions for such m to be a minimizer of V_β .

$$\begin{cases} W_j = (X'Xm)_j - \lambda_j & \text{for } j \in D_- \\ W_j = (X'Xm)_j + \lambda_j & \text{for } j \in D_+ \\ (X'Xm)_j - \lambda_j \leq W_j \leq (X'Xm)_j + \lambda_j & \text{for } j \in D_0 \end{cases} \quad (4)$$

Therefore, m satisfying $m + \beta \in \mathcal{O}^d$ is a minimizer of V_β if and only if W lies in the set

$$\{s \in \mathbb{R}^p : s_j = (X'Xm)_j - \lambda_j \text{ for } j \in D_-, s_j = (X'Xm)_j + \lambda_j \text{ for } j \in D_+, \\ (X'Xm)_j - \lambda_j \leq s_j \leq (X'Xm)_j + \lambda_j \text{ for } j \in D_0\},$$

which equals

$$X'Xm + \{s_\lambda : (s_\lambda)_{D_-} = -\lambda_{D_-}, (s_\lambda)_{D_+} = \lambda_{D_+}, |s_{\lambda,j}| \leq \lambda_j \text{ for } j \in D_0\}.$$

Since we are interested in all minimizers m of V_β that satisfy $m_j \leq z_j$ for $j \in D_-$, $m_j \geq z_j$ for $j \in D_+$ and $m_j = z_j$ for $j \in D_0$ (that is, $m - z \in \mathcal{O}^d$), we let

$$A^d = \{X'Xm : m - z \in \mathcal{O}^d\} + \{s \in \mathbb{R}^p : s_{D_-} = -\lambda_{D_-}, s_{D_+} = \lambda_{D_+}, |s_j| \leq \lambda_j \text{ for } j \in D_0\}.$$

As W follows a $N(0, \sigma^2 X'X)$ -distribution, the sought-after probability is clearly given by

$$\int_{A^d} \phi_{(0, \sigma^2 X'X)}(u) du,$$

which is what was claimed. \square

Results on the distribution of $\hat{\beta}_L$ itself rather than the estimation error are of course a direct consequence of Theorem 2 and summarized in the following corollaries, the latter one giving the probability of the extreme event of the Lasso setting all components to zero.

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

$$\begin{aligned} & P(\hat{\beta}_{L,j} \leq z_j \text{ for } j \in D_-, \hat{\beta}_{L,j} \geq z_j \text{ for } j \in D_+, \hat{\beta}_{L,j} = 0 \text{ for } j \in D_0) \\ &= \int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \cdots \int_{\substack{m_j \geq z_j - \beta_j \\ j \in D_+}} \cdots \int_{\substack{m_j \leq z_j - \beta_j \\ j \in D_-}} \cdots \int \phi_{(0, \sigma^2 X'X)}(X'Xm_\beta + s_\lambda) dm_{D_-} dm_{D_+} ds_{D_0}, \end{aligned}$$

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

Corollary 4.

$$P(\hat{\beta}_L = 0) = \int_{-\lambda_p}^{\lambda_p} \cdots \int_{-\lambda_1}^{\lambda_1} \phi_{(X'X\beta, \sigma^2 X'X)}(s) ds$$

Remark 1. *To illustrate the structure behind the proof of Theorem 2, we rewrite Corollary 3 as*

$$P(\hat{\beta} \in B_z) = P(W \in A_\beta(B_z))$$

with $B_z = \{b \in \mathbb{R}^p : b_j \leq z_j \text{ for } j \in D_-, b_j \geq z_j \text{ for } j \in D_+, b_j = 0 \text{ for } j \in D_0\}$, $W \sim N(0, \sigma^2 X'X)$, and $A_\beta(B_z) = \bigcup_{b \in B_z} X'X(b - \beta) + \prod_{j=1}^p B_j(b_j)$ where

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

The events $\{\hat{\beta} \in B_z\}$ and $\{W \in A_\beta(B_z)\}$ are shown to be equivalent through Lemma 1. This equivalence holds due to the structure of the optimization problem defining $\hat{\beta}_L$ 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 $A_\beta(B_z)$ are illustrated in Figure 1. Note that A_β depends on λ whereas B_z does not.

We exploit the structure of the optimization problem by characterizing the minimum through directional derivatives rather than using the Kuhn-Karush Tucker conditions as in the high-level approach of Zhou (2014). This has the advantage that our distributional results give the joint distribution of the estimator only and are not “augmented” by the subdifferential vector.

Theorem 2 now puts us into a position to fully specify the distribution of the Lasso estimator. In case $\lambda_j > 0$ for all $j = 1, \dots, p$, 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, positive, and equal to zero – which we shall do in the sequel.

Proposition 5. *The distribution of $\hat{u} = \hat{\beta}_L - \beta$, conditional on the event $\{\hat{\beta}_L \in \mathcal{O}^d\}$, can be represented by a $\|d\|_1$ -dimensional Lebesgue-density given by*

$$f^d(z_{D^\pm}) = \frac{\mathbf{1}\{z_\beta + \beta \in \mathcal{O}^d\}}{P(\hat{\beta}_L \in \mathcal{O}^d)} \int \cdots \int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \phi_{(0, \sigma^2 X'X)}(X'X z_\beta + s_\lambda) ds_{D_0},$$

where z_β is defined by $(z_\beta)_{D^+} = z_{D^+}$ 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 $P(\hat{\beta}_L \in \mathcal{O}^d)$ can be calculated using Corollary 3.

Proof. Observe that

$$f^d(z_{D^\pm}) = \left(\frac{\partial}{\partial z_j} \right)_{j \in D^+} P\left(\hat{u}_j \leq z_j \text{ for } j \in D^+ | \hat{\beta}_L \in \mathcal{O}^d\right),$$

and note that by Theorem 2 for any $z \in \mathbb{R}^p$ we have with $z + \beta \in \mathcal{O}^d$ we have

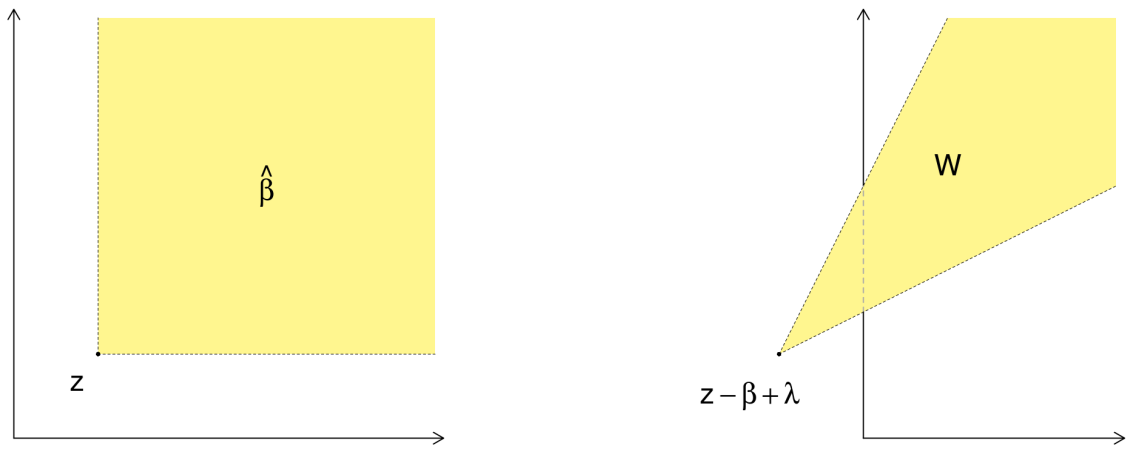
$$\begin{aligned} & P\left(\hat{u}_j \leq z_j \text{ for } j \in D_-, \hat{u}_j \geq z_j \text{ for } j \in D_+ | \hat{\beta}_L \in \mathcal{O}^d\right) \\ &= \frac{1}{P(\hat{\beta}_L \in \mathcal{O}^d)} P\left(\hat{u}_j \leq z_j \text{ for } j \in D_-, \hat{u}_j \geq z_j \text{ for } j \in D_+, \hat{\beta}_{L,j} = 0 \text{ for } j \in D_0\right) \\ &= \frac{1}{P(\hat{\beta}_L \in \mathcal{O}^d)} \int \cdots \int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \int \cdots \int_{\substack{m_j \leq z_j \\ j \in D^+}} \int \cdots \int_{\substack{m_j \geq z_j \\ j \in D^-}} \phi_{(0, \sigma^2 X'X)}(X'X m_\beta + s_\lambda) dm_{D_-} dm_{D^+} ds_{D_0}, \end{aligned}$$

where $m_\beta \in \mathbb{R}^p$ is defined by $(m_\beta)_{D^+} = m_{D_- \cup D^+}$, 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 $z_j : j \in D^+$ and taking the absolute value gives the density, thus completing the proof. \square

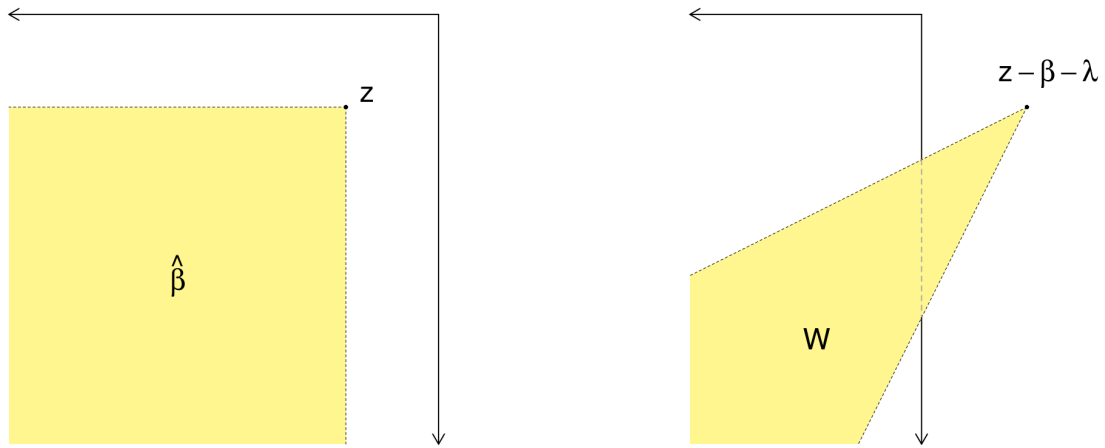
Besides the conditional densities, we can also specify the full cdf of $\hat{u} = \hat{\beta}_L - \beta$ which is done in the following theorem.

Theorem 6. *The cdf of $\hat{u} = \hat{\beta}_L - \beta$ is given by*

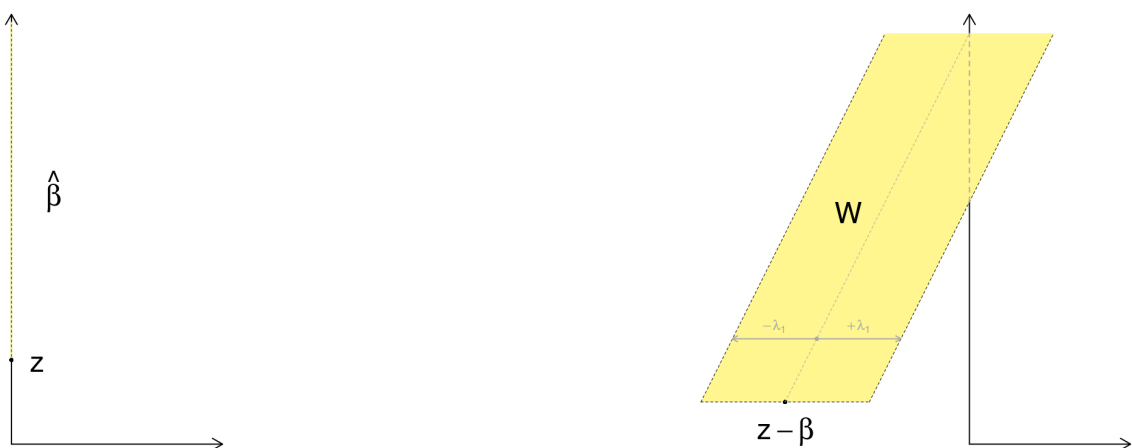
$$F(z) = P(\hat{u}_1 \leq z_1, \dots, \hat{u}_p \leq z_p) = \sum_{d \in \{-1, 0, 1\}^p} \int \cdots \int_{\substack{m_j \leq z_j \\ j \in D^+}} h^d(m_{D^+}) d\nu_{\|d\|_1},$$



(a) $z_1, z_2 > 0$



(b) $z_1, z_2 < 0$



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

Figure 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. Illustrated for $p = 2$ and various values of z , see Remark 1 for details.

where ν_k denotes k -dimensional Lebesgue-measure and where

$$h^d(m_{D_-^+}) = \mathbf{1}\{m_\beta + \beta \in \mathcal{O}^d\} \int \cdots \int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \phi_{(0, \sigma^2 X'X)}(X'X m_\beta + s_\lambda) ds_{D_0},$$

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

Proof. It is easily seen that

$$P(\hat{u}_1 \leq z_1, \dots, \hat{u}_p \leq z_p) = \sum_{d \in \{-1, 0, 1\}^p} P(\hat{\beta}_L \in \mathcal{O}^d) \int \cdots \int_{\substack{m_j \leq z_j \\ j \in D_-^+}} f^d(m_{D_-^+}) d\nu_{\|d\|_1}.$$

Plugging in the formula for f^d completes the proof. \square

For illustration of Proposition 5 and Theorem 6, consider Figures 2 and 3 which display an example of the distribution of $\hat{u} = \hat{\beta}_L - \beta$. One can see that the Lasso estimation error follows a shifted normal distribution conditional on the event $\hat{u}_j \neq -\beta_j$ ($\hat{\beta}_{L,j} \neq 0$) for each $j = 1, \dots, p$ with the shift depending on the signs of $\hat{\beta}_L$ as is to 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 4.

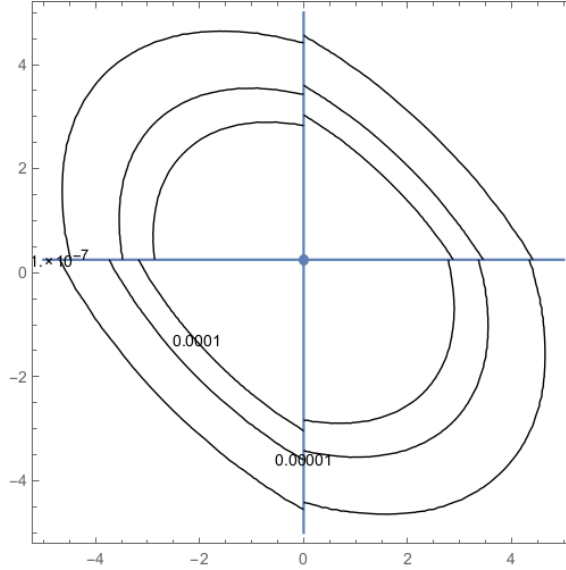


Figure 2: The contour lines of the absolutely continuous part with respect to 2-dimensional Lebesgue-measure of the distribution of $\hat{\beta}_L - \beta$ for $X'X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\lambda = (0.75, 0.75)'$ and $\beta = (0, -0.25)'$. Note that the blue lines as well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

3.1 Shrinkage Areas

Using the conditions for minimality from the proof of Theorem 2, 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.

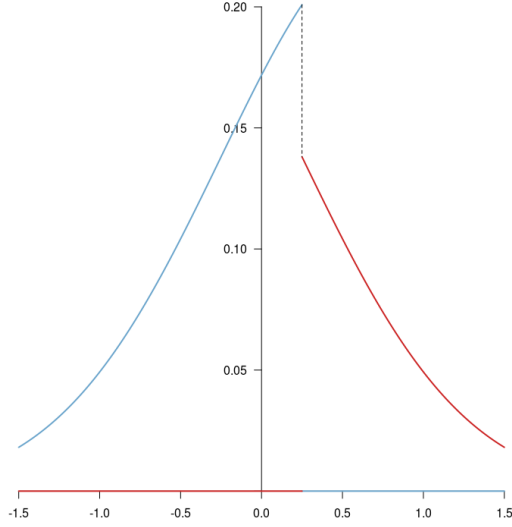


Figure 3: The functions $h^{(0,-1)'}$ (in blue) and $h^{(0,1)'}$ (in red) for $X'X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\lambda = (0.75, 0.75)'$ and $\beta = (0, -0.25)'$, corresponding to the absolutely continuous part.

Theorem 7. For each $b \in \mathbb{R}^p$ there exists a set $S(b) \subseteq \mathbb{R}^p$, such that

$$\hat{\beta}_L = b \iff \hat{\beta}_{LS} \in S(b).$$

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

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

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

Proof. Note that we have $\hat{\beta}_{LS} - \beta = (X'X)^{-1}X'\varepsilon = (X'X)^{-1}W$. With the minimality conditions in (4) from the proof of Theorem 2 together with the fact that $W = X'X(\hat{\beta}_{LS} - \beta)$ and some rearranging, we get that $m = \hat{\beta}_L - \beta$ minimizes V_β if and only if $\hat{\beta}_{LS}$ satisfies

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

or, $\hat{\beta}_{LS} \in S(b)$ for $\hat{\beta}_L = b$, as required. \square

Remark 2. Clearly, if $b \in \mathbb{R}^p$ satisfies $b_j \neq 0$ for all $j = 1, \dots, p$, then $S(b)$ is the singleton

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

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

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

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

Given Theorem 7, we can identify areas in which components of the LS estimator are shrunk to zero by the Lasso. For $p = 2$, this leads to the image displayed in Figure 4. Clearly, the shrinkage areas are related to the polyhedral selection areas of Lee et al. (2016) but yield a different level of information. Our results can identify the regions that lead to a given value b of $\hat{\beta}_L$ rather than to a general model (with given signs of the non-zero components of the estimator).

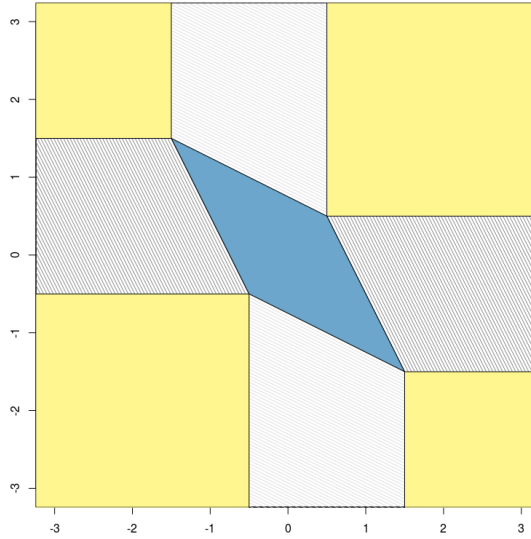


Figure 4: The shrinkage areas from Theorem 7 for $p = 2$. The blue parallelogram equals the set $S(0)$. The dark gray area should consists of lines parallel to the adjacent edge of the parallelogram where each line equals a set $S\left(\frac{0}{b_2}\right)$ for $b_2 \neq 0$. Analogously, the light gray area consists of lines parallel to the adjacent edge of the parallelogram and each of those lines equals a set $S\left(\frac{b_1}{0}\right)$ for $b_1 \neq 0$. The yellow areas contain all singletons $S\left(\frac{b_1}{b_2}\right)$ with $b_1, b_2 \neq 0$ as described in Remark 2. In this example, $X'X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ 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 on in Section 4.3.

Note that for any fixed $\beta \in \mathcal{B}_0$, the function V_β defined in (3) is minimized at $\hat{\beta}_L - \beta$, where $\hat{\beta}_L$ may be any solution of (2). 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.

Theorem 8. For any set $M \subseteq \mathbb{R}^p$ and any $\beta \in \mathcal{B}_0$, we have

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

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

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

Proof. Using the same necessary and sufficient conditions for $m \in \mathbb{R}^p$ to be a minimizer of V_β as in (4), we see that

$$m \in \arg \min_{u \in \mathbb{R}^p} V_\beta(u) \iff W \in A_\beta(m).$$

□

While the distribution of $\hat{\beta}_L - \beta$ depends on the choice of $\beta \in \mathcal{B}_0$, the distribution of $\hat{\beta}_L$ does not, as it is determined by $y \sim N(\mu, \sigma^2 I_n)$. This is formalized in the following corollary. As mentioned

before, $\hat{\beta}_L$ need not be unique. Also remember that $\hat{\beta}_L$ itself minimizes the function $L(\beta)$ defined in (2).

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

$$P(\arg \min_{\beta \in \mathbb{R}^p} L(\beta) \cap B \neq \emptyset) = P(W \in A_0(B)),$$

where $W \sim N(0, \sigma^2 X'X)$ and $A_0(B) = \bigcup_{b \in B} A_0(b)$ with $A_0(b) = X'X(b - \beta) + \prod_{j=1}^p B_j(b_j)$ 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 $\hat{\beta}_L$ does not depend on the choice of $\beta \in \mathcal{B}_0$.

Proof. First note that $\arg \min_u V_\beta(u) = \arg \min_\beta L(\beta) - \beta$. We thus have

$$b \in \arg \min_{\beta} L(\beta) \iff b - \beta \in \arg \min_u V_\beta(u).$$

Using Theorem 8, we get for any $\beta \in \mathcal{B}_0$

$$b - \beta \in \arg \min_u V_\beta(u) \iff W \in X'X(b - \beta) + \prod_{j=1}^p B_j(b_j) = A_0(b). \quad (5)$$

Finally, note that A_0 depends on β only through $X'X\beta$ which assumes the same value for all $\beta \in \mathcal{B}_0$. \square

As the random variable $W = X'\varepsilon$ has singular covariance matrix, some care needs to be taken when computing the probability from Corollary 9 through the appropriate integral of the corresponding density function. This is specified in Corollary 10.

Corollary 10. *Let the columns of U form a basis of $\text{col}(X')$. The probability that a Lasso solution lies in the set $B \in \mathbb{R}^p$ can be written as*

$$P(\arg \min_{\beta \in \mathbb{R}^p} L(\beta) \cap B \neq \emptyset) = \mathbb{1}\{\text{col}(X') \cap A_0(B) \neq \emptyset\} \int_{U'A_0(B)} \phi_{(0, \sigma^2 U'X'XU)}(s) ds$$

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 $\text{col}(X')^\perp$, then $N'W$ has covariance matrix $\sigma^2 N'X'XN = 0$ and $N'W = 0$ almost surely. We therefore have

$$\begin{aligned} W \in A_0(B) &\iff (U, N)'W \in (U, N)'A_0(B) \iff U'W \in U'A_0(B) \text{ and } 0 \in N'A_0(B) \\ &\iff U'W \in U'A_0(B) \text{ and } \text{col}(X') \cap A_0(B) \neq \emptyset, \end{aligned}$$

which proves the claim. \square

4.1 Shrinkage Areas and Model Selection Properties

For the low-dimensional case, Theorem 7 gives shrinkage areas of the Lasso with respect to $\hat{\beta}_{LS}$. In the high-dimensional case, similar results are derived in Theorem 11 whose proof is based on Corollary 9. These shrinkage areas are now given with respect to $X'y$ as this quantity is always uniquely defined in the high-dimensional case.

Theorem 11. *For each $b \in \mathbb{R}^p$ there exists a set $\bar{S}(b) \subseteq \mathbb{R}^p$ such that*

$$b \in \arg \min_{\beta \in \mathbb{R}^p} L(\beta) \iff X'y \in \bar{S}(b),$$

Moreover, $\bar{S}(b)$ is given by

$$\bar{S}(b) = X'Xb + \prod_{j=1}^p B_j(b_j).$$

with

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

Proof. As $X'y = X'X\beta + X'\varepsilon = X'X\beta + W$, this follows immediately from (5). \square

Remark 3. *Inspecting the sets $\bar{S}(b)$ from Theorem 11 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 the Lasso solution need not be unique in high dimensions anymore. (This is of course in contrast to the low-dimensional case as can be seen in Theorem 7.) Indeed, we can have $\bar{S}(b) \cap \bar{S}(b') \neq \emptyset$ as long as $b - b' \in \ker(X)$ and $\{\text{sgn}(b_j), \text{sgn}(b'_j)\} \neq \{-1, 1\}$ for all $j = 1, \dots, p$. 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 $\lambda_1 = \dots = \lambda_p > 0$. We get deeper into the issue of (non-)uniqueness in Section 4.3.*

Moreover, just as for the low-dimensional case, note that aside from b , $\bar{S}(b)$ depends on X and λ only.

Theorem 11 also sheds some light on which models $\mathcal{M} \subseteq \{1, \dots, p\}$ may in fact be chosen by the Lasso estimator, where the model chosen by $\hat{\beta}_L$ is given by $\{j : \hat{\beta}_{L,j} \neq 0\}$. We find that some models will in fact never be selected by the Lasso.

Let $\mathcal{B}_{\mathcal{M}} = \{b \in \mathbb{R}^p : b_j \neq 0 \text{ if and only if } j \in \mathcal{M}\}$. Since $X'y$ lies in $\text{col}(X')$, any model $\mathcal{M} \subseteq \{1, \dots, p\}$ that satisfies that $\text{col}(X') \cap \bar{S}(\mathcal{B}_{\mathcal{M}}) = \emptyset$ with $\bar{S}(\mathcal{B}_{\mathcal{M}}) = \bigcup_{b \in \mathcal{B}_{\mathcal{M}}} \bar{S}(b)$ can be never be selected by the Lasso estimator. Looking at the definition of $\bar{S}(b)$ in Theorem 11 and noting that $X'Xb \in \text{col}(X')$, we get

$$\text{col}(X') \cap \bar{S}(\mathcal{B}_{\mathcal{M}}) = \emptyset \iff \text{col}(X') \cap \mathcal{B}_{\mathcal{M}} = \emptyset,$$

where

$$\mathcal{B}_{\mathcal{M}} = \bigcup_{b \in \mathcal{B}_{\mathcal{M}}} \prod_{j=1}^p \begin{cases} \text{sgn}(b_j)\lambda_j & j \in \mathcal{M} \\ [-\lambda_j, \lambda_j] & j \notin \mathcal{M} \end{cases} = \prod_{j=1}^p \begin{cases} \{-\lambda_j, \lambda_j\} & j \in \mathcal{M} \\ [-\lambda_j, \lambda_j] & j \notin \mathcal{M}. \end{cases}$$

We summarize this in the following corollary.

Corollary 12. *Let $X \in \mathbb{R}^{n \times p}$ and $\lambda \in \mathbb{R}_{\geq 0}^p$ be given. There exist $y \in \mathbb{R}^n$ such that the corresponding Lasso solution selects model $\mathcal{M} \subseteq \{1, \dots, p\}$ if and only if*

$$\text{col}(X') \cap \mathcal{B}_{\mathcal{M}} \neq \emptyset,$$

where

$$\mathcal{B}_{\mathcal{M}} = \begin{cases} \{-\lambda_j, \lambda_j\} & j \in \mathcal{M} \\ [-\lambda_j, \lambda_j] & j \notin \mathcal{M}. \end{cases}$$

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

What do the sets $\mathcal{B}_{\mathcal{M}}$ look like? If $\mathcal{M}_0 = \emptyset$, \mathcal{B}_{\emptyset} is the p -dimensional λ -box, $\mathcal{B}_{\{j\}}$ is the union of two opposite facets of the λ -box, for $1 < |\mathcal{M}| < p$, $\mathcal{B}_{\mathcal{M}}$ is the union of $p - |\mathcal{M}|$ dimensional faces of the λ -box, and $\mathcal{B}_{\{1, \dots, p\}}$ simply contains the corners of the λ -box.

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

$$\{0\} \subseteq \text{col}(X') \cap \mathcal{B}_{\mathcal{M}_0} \neq \emptyset,$$

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.

We illustrate the results above by the following simplistic yet instructive example.

Example 1. *Suppose $X = (1, 2)$, so that $n = 1$, $p = 2$ and let $\lambda_1 = \lambda_2 = \bar{\lambda}$ (uniform tuning). Note that*

$$\text{col}(X') \cap \mathcal{B}_{\{1\}} = \emptyset$$

for all $\bar{\lambda} > 0$, so that by Corollary 12, $\hat{\beta}_{L,1} = 0$ for any value of y , in fact independent of \mathcal{B}_0 and σ^2 ! This yields

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

To say something about the remaining distribution of $\hat{\beta}_L$ using Corollary 9, note that $W = X'\varepsilon = \binom{1}{2}\varepsilon$. We therefore get $P(\hat{\beta}_L = 0) = P(W \in A_0(0))$ with

$$A_0(0) = -X'X\beta + \begin{pmatrix} [-\bar{\lambda}, \bar{\lambda}] \\ [-\bar{\lambda}, \bar{\lambda}] \end{pmatrix} = -(\beta_1 + 2\beta_2) \binom{1}{2} + \begin{pmatrix} [-\bar{\lambda}, \bar{\lambda}] \\ [-\bar{\lambda}, \bar{\lambda}] \end{pmatrix},$$

so that the event $\{W \in A_0(0)\}$ is equivalent to the event that W lies in the set

$$\{(s - \mu) \binom{1}{2} : s \in [-\bar{\lambda}/2, \bar{\lambda}/2]\}$$

with $\mu = X\beta = \beta_1 + 2\beta_2$, whose probability is given by

$$P(\hat{\beta}_L = 0) = \Phi((\bar{\lambda}/2 - \mu)/\sigma) - \Phi((-\bar{\lambda}/2 - \mu)/\sigma).$$

Next, for $z < 0$, we have

$$\begin{aligned} P(\hat{\beta}_{L,1} = 0, \hat{\beta}_{L,2} \leq z) &= P(W \in \bigcup_{b_2 < z} A_0(\binom{0}{b_2})) = P(W \in \bigcup_{b_2 < z} b_2 \binom{2}{4} + \begin{pmatrix} [-\bar{\lambda}, \bar{\lambda}] \\ \{-\bar{\lambda}\} \end{pmatrix}) \\ &= P(W \in \bigcup_{b_2 < z} (2b_2 - \bar{\lambda}/2) \binom{1}{2}) = P(\varepsilon \leq 2z - \bar{\lambda}/2) = \Phi(2z - \bar{\lambda}/2). \end{aligned}$$

Similarly, we get that for $z > 0$

$$P(\hat{\beta}_{L,1} = 0, \hat{\beta}_{L,2} > z) = 1 - \Phi(2z + \lambda_1/2).$$

The distribution of $\hat{\beta}_L$ is thus given by

$$\hat{\beta}_{L,1} = 0$$

and $\hat{\beta}_{L,2}$ following the distribution given by

$$dF(z) = (\Phi(\lambda_1/2) - \Phi(-\lambda_1/2)) d\delta_0(z) + \mathbb{1}_{\{z < 0\}} 2\phi(2z - \lambda_1/2) dz + \mathbb{1}_{\{z > 0\}} 2\phi(2z + \lambda_1/2) dz.$$

It is interesting to note that the distribution of $\hat{\beta}_{L,2}$ is the same as the one of the Lasso estimator with the same penalization parameter in the smaller model $y_i = 2\beta_2 + \varepsilon$ where the first regressor is taken out. Indeed, using the Lasso in the smaller model is equivalent to using the Lasso in the larger model in our example, as the procedure only considers models that do not contain the first regressor. This fact is, of course, only valid for the specific form of X and λ . However, the models which are considered by the Lasso estimator in the first place 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 and is thus not to be considered a purely data-driven model selection procedure. The choice between the location model (when $\hat{\beta}_{L,1} = 0$ and $\hat{\beta}_{L,2} \neq 0$) and the pure noise model (when $\hat{\beta}_{L,1} = \hat{\beta}_{L,2} = 0$) in our example does of course very much depend on β and ε . These considerations are also illustrated in Figure 5 which depicts in which areas the quantity $X'y$ has to fall into in order for the Lasso estimator to choose a certain model. For clarification, note that the Lasso is always unique in this example.

The considerations sparked by Example 1 suggest that in the high-dimensional setting, model selection by the Lasso estimator may, indeed, not be a purely data-driven procedure in the sense that there is a *structural model* or *structural set* $\mathcal{M} \subseteq \{1, \dots, p\}$ determined by X and λ only, satisfying $\hat{\beta}_{L,j} = 0$ for any $j \notin \mathcal{M}$ 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 considered by the model selection procedure, completely independently of the data vector y . Put yet differently again, this means that for a given regressor matrix X , one can restrict or choose the class of models considered by choice of λ .

Given all the considerations above, one might ask whether such a structural model \mathcal{M} always satisfies $|\mathcal{M}| \leq n$ 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)¹. If this was indeed the case, the Lasso estimator would be equivalent to a low-dimensional Lasso procedure restricted to this structural model \mathcal{M} this could be used to employ results from low-dimensional models also for inference in high-dimensional settings.

¹Note 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 .

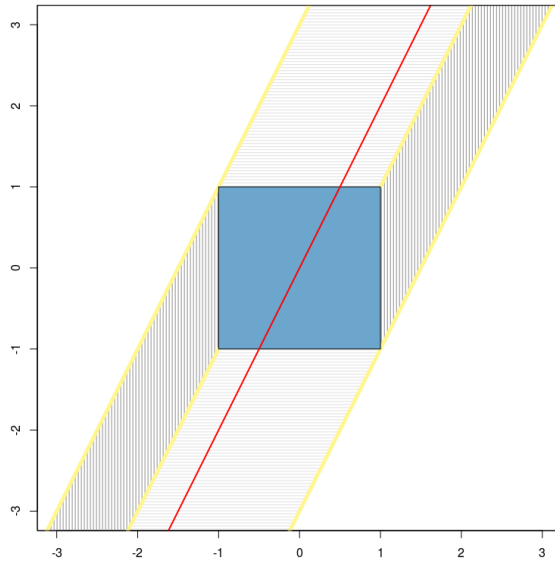


Figure 5: The shrinkage areas with respect to $X'y$ from Theorem 11 for Example 1. Displayed in red is $\text{span}(X')$, the area on which the probability mass of $X'y$ is concentrated. The set $\bar{S}\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ is displayed in blue, while the yellow lines correspond to $\bar{S}\left(\begin{smallmatrix} b_1 \\ b_2 \end{smallmatrix}\right)$ with $b_1, b_2 \neq 0$. The light gray area is $\bar{S}\left(\begin{smallmatrix} 0 \\ b_2 \end{smallmatrix}\right)$ with $b_2 \neq 0$, whereas the dark gray area equals $\bar{S}\left(\begin{smallmatrix} b_1 \\ 0 \end{smallmatrix}\right)$ with $b_1 \neq 0$.

Example 1 (continued). Here, $\mathcal{M} = \{2\}$ and the distribution of the Lasso estimator is equal to the distribution of the Lasso in the model

$$y = 2\beta_2 + \varepsilon.$$

Using this property, one can construct a confidence set for β_2 based on the Lasso using results from Pötscher & Schneider (2010).

In Example 1 the Lasso solution is always unique (for any value of y) since the column span of X' does not intersect overlapping shrinkage areas. It is not difficult, however, to construct an example where the Lasso solution is not unique anymore.

Example 2. Again, take the model from Example 1 with $X = (1, 2)$. This time, choose $\lambda = (1, 2)'$. It can easily be seen using Theorem 11 that for each $y < -1$,

$$\hat{\beta}_L = \begin{pmatrix} y+1 \\ 0 \end{pmatrix}, \hat{\beta}_L = \begin{pmatrix} 0 \\ \frac{y+1}{2} \end{pmatrix}, \text{ and } \hat{\beta}_L = (y+1-c^{-2c}) \text{ for } (y+1)/2 < c < 0.$$

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

$$\hat{\beta}_L = \begin{pmatrix} y-1 \\ 0 \end{pmatrix}, \hat{\beta}_L = \begin{pmatrix} 0 \\ \frac{y-1}{2} \end{pmatrix}, \text{ and } \hat{\beta}_L = (y+1-c^{-2c}) \text{ for } 0 < c < (y+1)/2.$$

all are Lasso solutions for the same value of y . (Note that $\hat{\beta}_L = 0$ for all y with $|y| \leq 1$.) The corresponding shrinkage areas are illustrated in Figure 6.

Example 2 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 solutions². This example shows, moreover, 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.

²This fact has been pointed out by Tibshirani (2013) in Lemma 1 for the case of uniform tuning.

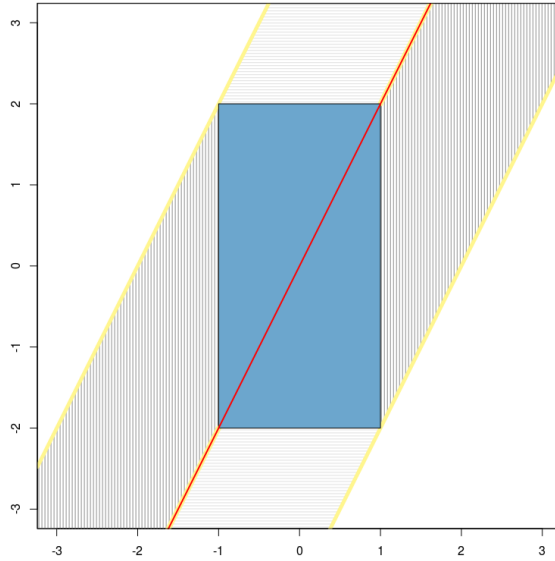


Figure 6: The shrinkage areas with respect to $X'y$ from Theorem 11 for Example 2. Displayed in red is $\text{span}(X')$, the area on which the probability mass of $X'y$ is concentrated. The set $S\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ is displayed in blue, while the yellow lines correspond to $S\left(\begin{smallmatrix} b_1 \\ b_2 \end{smallmatrix}\right)$ with $b_1, b_2 \neq 0$. The light gray area is $S\left(\begin{smallmatrix} 0 \\ b_2 \end{smallmatrix}\right)$ with $b_2 \neq 0$, whereas the dark gray area equals $S\left(\begin{smallmatrix} b_1 \\ 0 \end{smallmatrix}\right)$ with $b_1 \neq 0$. The red line passes through $S\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ 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 2 shows that the structural set may also be equal to the entire set of explanatory variables. In fact, 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 general. Before answering this question, we show how the structural set can fully be determined given X and λ by counting how many faces of the λ -box \mathcal{B}_\emptyset are intersected by $\text{col}(X')$.

Theorem 13. *Let $X \in \mathbb{R}^{n \times p}$ and $\lambda \in \mathbb{R}_{\geq 0}^p$ be given. Let \mathcal{M} be the 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 the corresponding Lasso solution $\hat{\beta}_L$ satisfies $\hat{\beta}_{L,j} \neq 0$, that is, the set of all regressors that are part of a Lasso solution for some observation y . This set is given by*

$$\mathcal{M} = \mathcal{M}(X, \lambda) = \{j \in \{1, \dots, p\} : \mathcal{B}_{\{j\}} \cap \text{col}(X') \neq \emptyset\}.$$

Proof. By Corollary 12, there exist $y \in \mathbb{R}^n$ such that the corresponding Lasso solution chooses model \mathcal{M} if and only if $\mathcal{B}_{\mathcal{M}} \cap \text{col}(X') \neq \emptyset$. For any $\{j\} \subseteq \mathcal{M}$, we have, by Corollary 12 also, $\mathcal{B}_{\mathcal{M}} \subseteq \mathcal{B}_{\{j\}}$, so that $\mathcal{B}_{\{j\}} \cap \text{col}(X') \neq \emptyset$ also. \square

Remark 4. *As indicated in Theorem 13 and as discussed above, the structural set \mathcal{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 $\omega \in \mathbb{R}_{\geq 0}^p$, $\mathcal{M}(X, \lambda) = \mathcal{M}(X, \omega)$ follows. This implies that, in particular, in the common case of uniform tuning with $\bar{\lambda} = \lambda_1 = \dots = \lambda_p$, the structural set only depends on X !*

Coming back to the conjecture whether the structural set always satisfies $|\mathcal{M}| \leq \min\{n, p\}$ in case the solutions are unique, using Theorem 13, 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 13 allows to compute the structural set and that whenever $|\mathcal{M}| \leq n$, the resulting Lasso estimator is, in fact, just equivalent to a low-dimensional procedure!

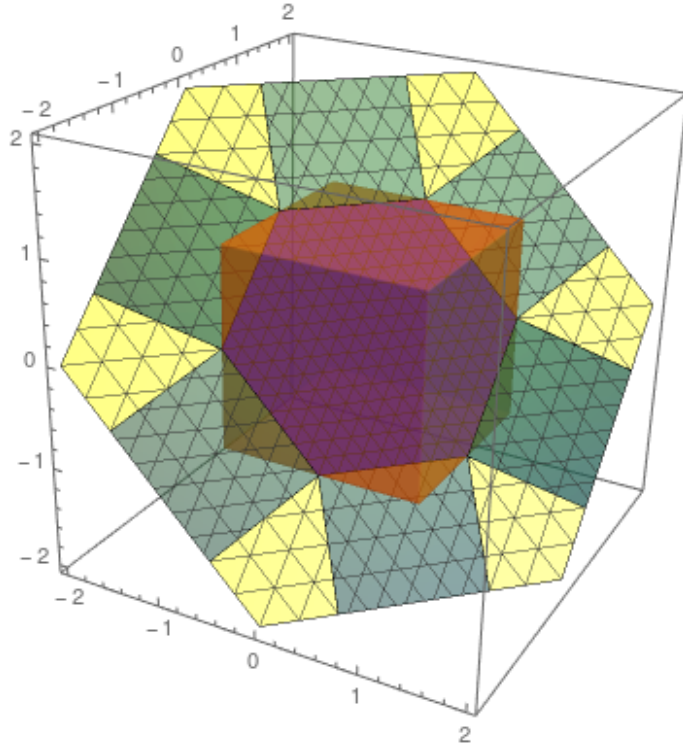


Figure 7: The intersection of $\text{col}(X')$ (in gray and yellow) with the λ -cube (in orange). The upper left edge is contained in $\mathcal{B}_{\{1,2\}}$ whereas the upper back edge is contained in $\mathcal{B}_{\{2,3\}}$. (Each $\mathcal{B}_{\{i,j\}}$ contains four parallel edges.) To view this figure in terms of shrinkage areas, note that the areas corresponding to single-regressor models are displayed in gray while the shrinkage areas that correspond to two-regressor models are displayed in yellow. The intersection of the λ -cube with $\text{col}(X')$ which corresponds to the zero estimator is displayed in blue.

Example 3. *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

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

as $(1, 1, 0)' \in \text{col}(X') \cap \mathcal{B}_{\{1,2\}}$ and $(0, 1, 1)' \in \text{col}(X') \cap \mathcal{B}_{\{2,3\}}$ and $\mathcal{B}_{\mathcal{M}} \subseteq \mathcal{B}_{\{j\}}$ for any $j \in \mathcal{M}$ by Corollary 12, see Figure 7 for illustration. Yet the Lasso solutions for this X and λ are always unique which can be checked on the basis of Theorem 14 in the subsequent 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 a similar geometric approach, that is, studying the intersection of the λ -cube with $\text{col}(X')$. 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 for all observations y if the columns of X are *in general position*³ which occurs when no k -dimensional affine⁴ subspace for any $k < \min(n, p)$ contains more than $k + 1$ elements of the

³Note 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.

⁴In Tibshirani (2013), the word “affine” is missing which has caused some confusion in the literature.

the set $\{\pm X_1, \dots, \pm X_p\}$, 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 uniformly. As the 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 exist non-unique solutions? 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

$$S(b) \cap S(\tilde{b}) \cap \text{col}(X') \neq \emptyset.$$

More concretely, by Theorem 11, and since the Lasso fit Xb is always unique⁵, this means that

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

where $v \in \text{col}(X') \cap \mathcal{B}_{\mathcal{M}}$ for some $\mathcal{M} \subseteq \{1, \dots, p\}$ and $\tilde{b}_{\mathcal{M}^c} = b_{\mathcal{M}^c} = 0$. Moreover, for $j \in \mathcal{M} \setminus \mathcal{M}_0$, we have $\text{sgn}(b_j) = \text{sgn}(v_j)$ whenever $b_j \neq 0$ as well as $\text{sgn}(\tilde{b}_j) = \text{sgn}(v_j)$ 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 if and only if $\text{col}(X') \cap \mathcal{B}_{\mathcal{M}} \neq \emptyset$ for $\mathcal{M} \subseteq \{1, \dots, p\}$ 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. 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. This is the case since $\text{col}(X') \cap \mathcal{B}_{\mathcal{M}} = \emptyset$ whenever $\{1, 2\} \subseteq \mathcal{M}$ or $|\mathcal{M}| > 2$ which can easily be checked using the fact that $v \in \text{col}(X')$ if and only if $v'w_1 = v'w_2 = 0$ for $\ker(X) = \text{span}\{w_1, w_2\}$.

We find that the following criterion is in fact sufficient as well as necessary for uniqueness of all Lasso solutions for a given X and λ .

Theorem 14 (Uniqueness). *Let $X \in \mathbb{R}^{n \times p}$ and $\lambda \in \mathbb{R}_{\geq 0}^p$. The Lasso solution is unique for all $y \in \mathbb{R}^n$ if and only if*

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

Proof. (\implies) Assume the condition is not satisfied. Then there exists $v \in \mathcal{B}_{\mathcal{M}}$ with $|\mathcal{M}| > \text{rk}(X)$ 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 that would imply $v_j = 0$ as $v \in \text{col}(X')$, but that contradicts $v \in \mathcal{B}_{\mathcal{M}}$. We therefore assume that $X_j \neq 0$ for all $j \in \mathcal{M}$.

Since $|\mathcal{M}| > \text{rk}(X)$, 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,

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

where $d = \text{sgn}(v_j)$ 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

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

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

$$X'y = X'Xb + X'z = X'Xb + v \in S(b).$$

⁵This has been shown in Lemma 1 in Tibshirani (2013) for uniform tuning and can easily be extended to non-uniform and partial tuning.

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

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

Clearly, $b \neq \tilde{b}$, $\operatorname{sgn}(b_l) = \operatorname{sgn}(\tilde{b}_l) = \operatorname{sgn}(v_j)$ 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 the given y .

(\Leftarrow) 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 $v \in \mathcal{B}_{\mathcal{M}} \cap \operatorname{col}(X')$ for some $\mathcal{M} \subseteq \{1, \dots, p\}$ with $X_{\mathcal{M}}b_{\mathcal{M}} = X_{\mathcal{M}}\tilde{b}_{\mathcal{M}}$ 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

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

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

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

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 $\mathcal{M} = \tilde{\mathcal{M}}$. \square

Remark. Note that just as for Theorem 13, 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 $\omega \in \mathbb{R}_{\geq 0}^p$, we have $\operatorname{col}(X') \cap \mathcal{B}_{\mathcal{M}}(\lambda) = \emptyset$ if and only if $\operatorname{col}(X') \cap \mathcal{B}_{\mathcal{M}}(\omega)$ (when indicating the dependence of $\mathcal{B}_{\mathcal{M}}$ on the tuning parameters).

As mentioned in the preamble of Section 4, Theorem 14 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 Gaussian errors. 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 of the error term. 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 with each active set having its own distributional piece whose dimension depends on the number of nonzero components in that it is given by 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. In fact, a structural set based on the regressor matrix and penalization weights only can be computed that contains all covariates that are part of a Lasso solution for some response vector y . 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. It remains to be seen whether one can quantify that this is in some sense a generic case.

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.

References

- EWALD, K. & SCHNEIDER, U. (2015). Confidence sets based on the Lasso. Tech. Rep. 1507.05315, arXiv.
- JAGANNATH, R. & UPADHYE, N. S. (2016). The Lasso estimator: Distributional properties. Tech. Rep. 1605:03280, arXiv.
- KNIGHT, K. & FU, W. (2000). Asymptotics of Lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.
- LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference with an application to the Lasso. *Annals of Statistics* **44**, 907–927.
- PÖTSCHER, B. M. & LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* **100**, 2065–2082.
- PÖTSCHER, B. M. & SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electronic Journal of Statistics* **4**, 334–360.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.
- TIBSHIRANI, R. J. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics* **7**, 1935–1490.
- ZHOU, Q. (2014). Monte carlo simulation for lasso-type problems by estimator augmentation. *Journal of the American Statistical Association* **109**, 1495–1516.