THREE SIMPLE HEURISTICS MATHEMATICAL PROOFS ON LASSO THEORY

Carlos José dos REIS[1]
Laerte Dias de CARVALHO[2]
Lucas Monteiro CHAVES[3]
Devanil Jaques de SOUZA[1]

■ ABSTRACT: Three relevant facts about the least absolute shrinkage and selection operator (Lasso) are studied: The estimatives follows piecewise linear curves in relation to tuning parameter, the number of nonzero selected covariates is an unbiased estimator of its degrees of freedom and when the number of covariates p is greater than the numbers of observations n at most n covariates are selected. These results are well known and described in the literature, but with no simple demonstrations. We present, based on a geometrical approach, simple and intuitive heuristics proofs for these results.

■ KEYWORDS: Covariates selection; degrees of freedom; piecewise linear; Shrinkage; Stein's lemma.

1 Introduction

Suppose the usual regression situation: data $\left(\boldsymbol{x}^i,y_i\right), i = 1,\ldots,n$, where $\boldsymbol{x}^i = (x_{i1},\ldots,x_{ip})'$ is a vector of predictors variables and y_i is the corresponding response. Consider as usual that the observations are independent and $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$. Tibshirani (1996) defines the Lasso estimative as the solution of the quadratic convex optimization problem:

[1]Universidade Federal de Lavras - UFLA, Departamento de Estatística, Caixa Postal: 3037, CEP: 37200-900, Lavras, MG, Brasil, E-mail: *carlosjreis17@yahoo.com.br; laerte@ufv.br; devaniljaques@ufla.br.*
[2]Universidade Federal de Viçosa - UFV, Departamento de Matemática, CEP: 36570-000 , Viçosa, MG, MG, Brasil, E-mail: *laerte@ufv.br.*
[3]Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, Caixa Postal: 3037, CEP: 37200-900, Lavras, MG, Brasil, E-mail: *lucas@ufla.br.*

$$\hat{\boldsymbol{\beta}}_{Lasso}(t) = \underset{\beta}{\arg\min} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 \quad \text{subject to } \left\|\boldsymbol{\beta}\right\|_1 = \sum_{i=1}^{p} \beta_i \leq t \quad . \tag{1}$$

The parameter restriction defines, for each t, a convex diamond shaped region K in $\mathbb{R}^p$. We can think the $n \times p$ design matrix $\boldsymbol{X} = (x_{ij})$ as a linear transformation from the Euclidean space $\mathbb{R}^p$ to $\mathbb{R}^n$. We will suppose, to avoid generalized inverses, that $\boldsymbol{X}$ is injective. In this case, we have a geometrical set-up described as: the image of $\mathbb{R}^p$, by the linear transform $\boldsymbol{X}$, is a p-dimensional subspace of $\mathbb{R}^n$, the image of the convex subset K is the convex subset $K_p = \boldsymbol{X}(K)$.

To obtain the Lasso estimative we have to find the point in K_p closest to the data vector $\boldsymbol{y}$. To do this we project $\boldsymbol{y}$ orthogonally into the subspace image of $\boldsymbol{X}$ ($\boldsymbol{y}_P^* = P_{\mathrm{Im}(\boldsymbol{X})} y$) and then find in K_p the point $\boldsymbol{y}_p$ closest to $\boldsymbol{y}_P^*$. As $\boldsymbol{X}$ is injective, the pre-image of this point defines the estimative $\hat{\boldsymbol{\beta}}_{Lasso}(t)$. In the parameter space, this is equivalent to find in K the point closest, in the Mahalanobis distance, $\langle \beta_1, \beta_2 \rangle_m = {\beta_1}' \boldsymbol{X}' \boldsymbol{X} \beta_2$, to the ordinary least squared estimative $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. This can be done by constructing several hyperboloids on $\hat{\boldsymbol{\beta}}_{\boldsymbol{ols}}$, until one of these reach a tangent point on K. (see Figure 1).

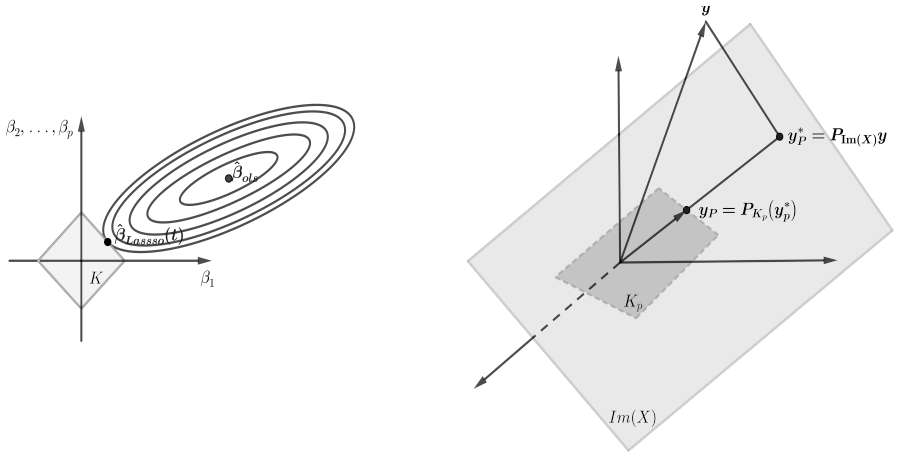

Figure 1 - Geometrical description of Lasso.

The Lasso estimative shrinks the coefficients towards zero as t goes to zero. Another main characteristic is that, with high probability, some coefficients are set exactly equal to zero since $\hat{\boldsymbol{\beta}}_{Lasso}(t)$ occurs in a singular face. Therefore the estimation process is also a model selection process.

LASSO regression theory has been successfully applied, for example, in healthcare (LEE et al., 2014; MUSORO et al., 2014), civil engineering (MANGALATHU et al., 2018) in new statistical methods (BELLONI et al., 2014;

LEE et al. 2016; CAVALCANTE et al., 2017).

The organization of this paper is as follows. In Section 2, we point out that, in relation to the tuning parameter t, the coordinates curves of $\hat{\boldsymbol{\beta}}_{Lasso}(t)$ are piecewise linear. We will present a simple but intuitive proof of this fact. In Section 3, we review the Stein's unbiased risk estimation and, with a very simple mathematical approach, we obtain the known unbiased estimator of the degrees of freedom for the Lasso. In Section 4 we prove that if the number n of observations are less than the number p of covariates then the Lasso selects at most n covariates.

2 Lasso trace curves are piecewise linear

Since the curves defined by Ridge estimator are named Ridge trace curves (HOERL and KENNARD, 1970) we will name $\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_i$, $i = 1, \ldots, p$ as Lasso trace curves. A typical picture of Lasso trace curves is given, for example, in Hastie and Tibshirani (2015) (Figure 2).

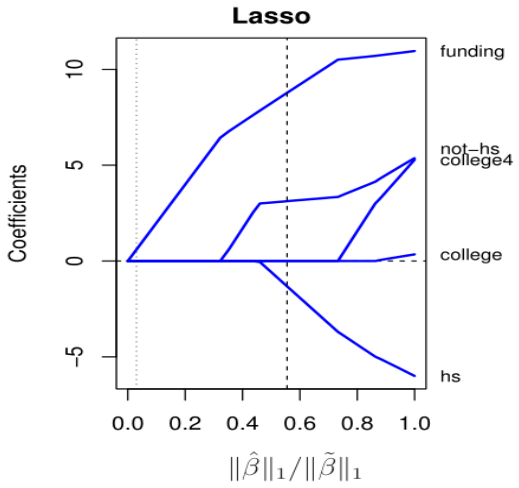

Figure 2 - Lasso trace curves.

Tibshirani (1996) showed that for the orthogonal case the Lasso trace curves are piecewise linear. Efron et al. (2004) presented a new model selection algorithm, named Least Angle Regression (LARS). This algorithm is piecewise linear by construction. The authors observed that the same geometry of the algorithm applies to the Lasso, despite of the fact that the two methods seems to be quite different.

Since authors couldn't find a simple approach for this fundamental property, a very elementary proof, although incomplete, using only undergraduate analytical geometry, will be present.

There are two families of hipersurfaces in $\mathbb{R}^p$: the family of parallel ellipsoids centered in $(a_1, \ldots, a_p)$ and the family of diamond shaped convex sets (or simplest

diamond set) of the form $\sum_{i=1}^{p} |\beta_i| = t$, where t is a parameter. It is clear that each ellipsoid has only two tangent points with planes that are faces in the family of diamond sets, both are border of convex subsets $\mathbb{R}^p$. It is intuitive that by the convexity of these two subsets. That they are tangent in two points. Of course, we are considering only regular points, that is, points in hyperfaces of the diamond set to avoid measure theory that is necessary if we consider singular faces. We will be concerned only with the tangent point closest to the origin. It is a typical problem in Mathematical analysis to show that tangents points between these two families defines a smooth curve, which is called by definition $\hat{\beta}_{(lasso)}(t)$. We have to show that this curve is a straight line. First we will suppose that the family of ellipsoids have principal axes parallel to the coordinates axes.

Therefore the family of ellipsoids is of the form:

$$\frac{(\beta_1 - a_1)^2}{r_1^2} + \ldots + \frac{(\beta_n - a_p)^2}{r_p^2} = r,$$

where $r_1, \ldots, r_p$ are fixed number and $r > 0$ is a family parameter.

Clearly for each r there is only two values t such that the ellipsoid and the hyperplane are tangents. We will consider only the tangent point closest to the origin and will suppose also that this tangent point is in a $(p-1)$-dimensional face of the hypercube. In tangency point the hyperplane and the ellipsoid have a common normal vector. If this tangent point has positive coordinates a normal vector of the hyperplane is the vector $(1, \ldots, 1)$, and this vector is also normal to the ellipsoid. Let's $(\beta_1(s), \ldots, \beta_p(s))$ be a curve in the ellipsoid, such that $(\beta_1(0), \ldots, \beta_p(0)) = \hat{\boldsymbol{\beta}}_{(Lasso)}(t))$ is the tangent point. By implicit differentiation of

$$\frac{(\beta_1(s) - a_1)^2}{r_1^2} + \ldots + \frac{(\beta_p(s) - a_p)^2}{r_p^2} = r,$$

follows that:

$$\frac{2(\beta_1(0) - a_1)\beta'_1(0)}{r_1^2} + \ldots + \frac{2(\beta_p(0) - a_p)\beta'_p(0)}{r_p^2} = 0.$$

Hence, the vector $\left(\frac{(\beta_1(0)-a_1)}{r_1^2}, \ldots, \frac{(\beta_p(0)-a_p)}{r_p^2}\right)$ is perpendicular to the tangent vector $\left(\beta'_1(0), \ldots, \beta'_p(0)\right)$. As this vector is a generic vector on the tangent space of the ellipsoid, it is necessarily parallel to the vector $(1, \ldots, 1)$. That is,

$$\left(\frac{(\beta_1(0) - a_1)}{r_1^2}, \ldots, \frac{(\beta_p(0) - a_p)}{r_p^2}\right) = \alpha(1, \ldots, 1).$$

As α is dependent to the tangent point $\hat{\boldsymbol{\beta}}_{(Lasso)}(t)$ it is also a function of t and then

$$\left(\frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_1 - a_1\right)}{r_1^2}, \ldots, \frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_p - a_p\right)}{r_p^2} \right) = \alpha(t)\,(1,\ldots,1)\,.$$

Thus, the tangent point satisfies the equation

$$\left(\frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_1 - a_1\right)}{r_1^2}, \ldots, \frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_p - a_p\right)}{r_p^2} \right) = \alpha(t)\,(1,\ldots,1)\,.$$

This shows that the tangent points $\hat{\boldsymbol{\beta}}_{Lasso}(t)$ lies in a straight line. If the ellipsoids don't have its principal axes parallel to the coordinates axes a new coordinate system can be built in such way that the ellipsoid with this new coordinates have its axes parallel. What happens with the plane $\sum_{i=1}^{p} |\beta_i| = t$ in this new $\tilde{\beta}$ coordinates system? The normal vector change from $(1,\ldots,1)$ to $(b_1,\ldots,b_p)$ and it is easy to see that $\sum_{i=1}^{p} \left|b_i\tilde{\beta}_i\right| = t$.

Therefore we have

$$\left(\frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_1 - a_1\right)}{r_1^2}, \ldots, \frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_p - a_p\right)}{r_p^2} \right) = \alpha(t)\,(b_1,\ldots,b_p)\,,$$

and in this case we have again a straight line in this general situation. That is, we have the situation described in Figure 3.

As the tuning parameter t varies, the tangent point may move from a $p-1$ dimensional face of the hypercube to a lower dimensional face. In this case the normal vector changes, as an example, for a $(p-3)$-dimensional face the normal vector (1,1,0,0,1,...,1). And so,

$$\left(\frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_1 - a_1\right)}{r_1^2}, \ldots, \frac{\left(\left(\hat{\boldsymbol{\beta}}_{Lasso}(t)\right)_p - a_p\right)}{r_p^2} \right) = \alpha(t)\,(1,1,0,0,1,\ldots,1)\,.$$

Again we have a strait line but with a new direction. This fully describes the behavior of Lasso trace curves.

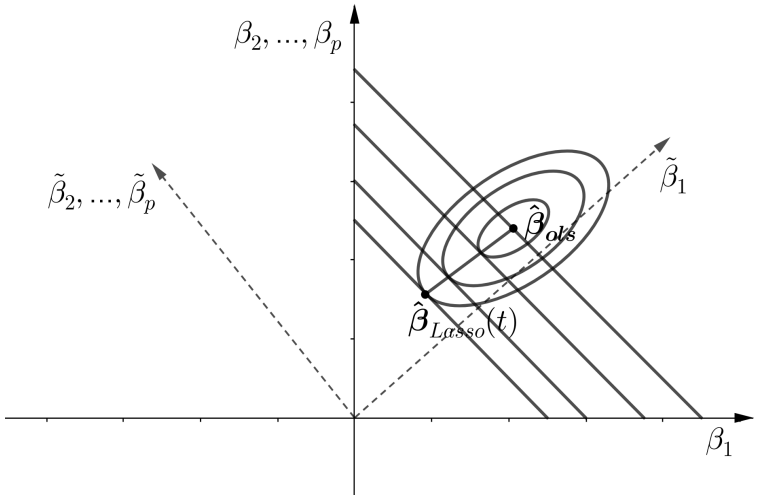

Figure 3 - Ellipsoids with principal axes in general position.

3 Degrees of freedom

If a model, for example, an ordinary linear regression, fits some data $\boldsymbol{y}$, producing an estimate $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$, $m : \mathbb{R}^n \to \mathbb{R}^n$, the question of how well $m(\boldsymbol{y})$ will predict a future dataset, independently generated from the same random mechanism that produced $\boldsymbol{y}$, is probably the main problem to be answered. This is the prediction error and it's the sum of expectation of the fitting error plus a penalty related to the covariance between the data $\boldsymbol{y}$ and the model $m(\boldsymbol{y})$ (EFRON, 2004). This drives us to the concept of degrees of freedom (df) as a covariance penalty.

Definition: The degrees of freedom of a model $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$ is defined as

$$\text{df} = \frac{\text{cov}(\boldsymbol{y}, m(\boldsymbol{y}))}{\sigma^2} = \sum_{i=1}^{n} \frac{\text{cov}(y_i, (m(y))_i)}{\sigma^2},$$

where σ^2 is the error variance.

In the linear case, $\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y}$, where $\boldsymbol{M}$ is a $n \times n$ matrix, the degrees of freedom is the trace of $\boldsymbol{M}$. If we are in the usual regression or analysis of variance (Anova), $\boldsymbol{M}$ is a projection matrix and, therefore, trace($\boldsymbol{M}$) $= p$, the dimension of the projection space, that is, the rank of $\boldsymbol{M}$, agreeing with the usual definition of degrees of freedom.

The degrees of freedom is a population parameter and has to be estimated. For this, we have to use the multidimensional version of classical Stein's lemma. Under very reasonable mathematical conditions on $m(\boldsymbol{y})$ we have:

Stein's lemma: If

$$\sum_{i=1}^{n} \frac{\operatorname{cov}\left(y_i,(m\left(\boldsymbol{y}\right))_i\right)}{\sigma^2} = E\left[\operatorname{div}\left(m\left(\boldsymbol{y}\right)\right)\right] = E\left[\sum_{i=1}^{n} \frac{\partial(m\left(\boldsymbol{y}\right))_i}{\partial y_i}\right]$$

then $\operatorname{div}\left(m\left(\boldsymbol{y}\right)\right)$ is an unbiased estimator of degrees of freedom.

3.1 Degrees of freedom for the Lasso

It is well know that the nonzero number of covariates select by Lasso is an unbiased estimator of the degree of freedom for the Lasso. Thus, the model $\hat{\boldsymbol{\mu}} = m\left(\boldsymbol{y}\right)$ is given by $m\left(\boldsymbol{y}\right) = P_{K_p}\left(\boldsymbol{y}\right) = \boldsymbol{y_p}$, where P_{K_p} is the minimum projection distance of the data $\boldsymbol{y}$ on the convex set K_p. To calculate $\operatorname{div}(\mathrm{P_{K_p}}\left(\mathrm{y}\right))$ we will follow Kato (2009).

Let $P_K : \mathbb{R}^p \to K$ be the minimum Mahalanobis distance projection on the convex set K. Therefore,

$$m\left(\boldsymbol{y}\right) = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{Lasso}\left(t\right) = \boldsymbol{X}P_K\hat{\boldsymbol{\beta}}_{ols}\left(\boldsymbol{y}\right) = \boldsymbol{X}P_K\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'y}.$$

To compute the divergence of $m\left(\boldsymbol{y}\right)$ we will have to use the chain rule. If $f : \mathbb{R}^n \to \mathbb{R}^n$, then $df\left(x\right)$ is a linear transformation $df : \mathbb{R}^n \to \mathbb{R}^n$. The divergence definition does not depend on coordinates and is given by $\operatorname{div} f\left(x\right) = \operatorname{tr}\left(\operatorname{df}\left(x\right)\right)$. If f is the linear transformation $f\left(\boldsymbol{y}\right) = \boldsymbol{My}$, then $\operatorname{df}\left(\boldsymbol{y}\right) = \boldsymbol{M}$ and $\operatorname{div}\left(\boldsymbol{y}\right) = \operatorname{tr}\left(\boldsymbol{M}\right)$. In the case of composition, $g \circ f : \mathbb{R}^n \overset{f}{\to} \mathbb{R}^p \overset{g}{\to} \mathbb{R}^n$, a derivative is given by composition of linear transformation $d\left(g \circ f\right) = dg\left(f\left(x\right)\right) \circ df\left(x\right)$.

Thus, $d\,m(\mathbf{y}) = \boldsymbol{X} \cdot d\,P_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right)\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'}$. It follows that,

$$\begin{aligned}
\operatorname{div} m(\mathbf{y}) &= \operatorname{tr}\left(\boldsymbol{X}\, dP_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right)\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'}\right) \\
&= \operatorname{tr}\left(d\,P_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right)\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'X}\right) \\
&= \operatorname{tr}\left(d\,P_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right)\right).
\end{aligned}$$

Therefore, the divergence of $m\left(\boldsymbol{y}\right)$ is equal to the divergence the projection P_K, in relation to the variable $\boldsymbol{\beta}$, applied to the point $\hat{\beta}_{ols}\left(\boldsymbol{y}\right)$.

In the orthogonal case, $\boldsymbol{X'X} = \boldsymbol{I}$, the Mahalanobis metric is the Euclidian metric and an explicit formula for the projection P_K is possible, given by:

$$P_K\left(\boldsymbol{\beta}\right) = \left(\operatorname{signal}(\beta)_1(|(\beta)_1 - \gamma|)^+, \ldots, \operatorname{signal}(\beta)_p\left(\left|(\beta)_p - \gamma\right|\right)^+\right),$$

where γ is a constant.

In this case the Lasso estimator is

$$\left(\hat{\boldsymbol{\beta}}_{Lasso}\left(t\right)\right)_i = \text{signal}\left(\hat{\boldsymbol{\beta}}_{ols}\left(\boldsymbol{y}\right)\right)_i \left(\left|\left(\hat{\boldsymbol{\beta}}_{ols}\left(\boldsymbol{y}\right)\right)_i - \gamma\right|\right)^+,$$

where γ depends on the value of t.

With this formula is possible to calculate $\text{div}P_K\left(\hat{\beta}_{ols}\left(\boldsymbol{y}\right)\right)$

$$\frac{\partial}{\partial\beta_i}\text{signal}(\beta)_i(|(\beta)_i| - \gamma)^+ = \begin{cases} \text{signal}(\beta)_i\text{signal}(\beta)_i = 1, & if\,|(\beta)_i| > \gamma \\ 0, & if\,|(\beta)_i| \leq \gamma \end{cases}.$$

Then, $tr\left(dP_K\left(\hat{\boldsymbol{\beta}}_{ols}\left(\boldsymbol{y}\right)\right)\right)$ is the number of nonzero coordinates on Lasso estimative, that is, the number of selected covariates.

For the general case, the proofs of the degrees of freedom for the Lasso (KATO, 2009; TIBSHIRANI and TAYLOR, 2012) are quite of complex since the diamond shaped set K has faces with dimensions $0, 1, \ldots, p-1$ and it is necessary to find the intersections of ellipsoids with these low dimensional faces. Such a situation requires measure theory. Here, we will present a mathematical semi-complete proof, however with a much more intuitive and useful application for a broad statistical audience.

The only thing that we have to intuitively accept is that each face has a domain of attraction. That is, for almost every $\boldsymbol{\beta}$ that projects on a face L there is an open subset around $\boldsymbol{\beta}$ that also projects in the same face. Let us give an example: for the orthogonal case, on $\mathbb{R}^2$, the singular face with only one point (0,t) has the domain of attraction as Figure 4.

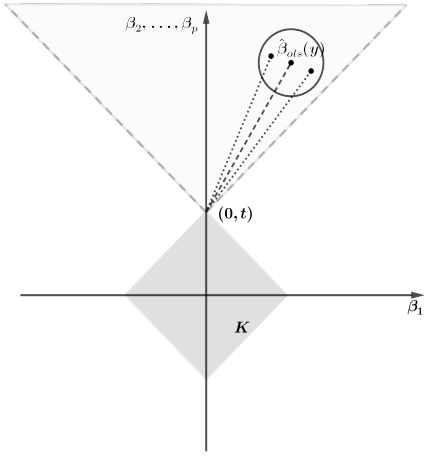

Figure 4 - Attraction domain of a singular face in the orthogonal case.

What changes for the non orthogonal case? The Mahalanobis metric preserves straight lines. The only thing that changes is the angles. Therefore, the orthogonal

projection in the Mahalanobis distance is the same as a oblique projection in Euclidean metric as seen in the previous section (see Figure 5).

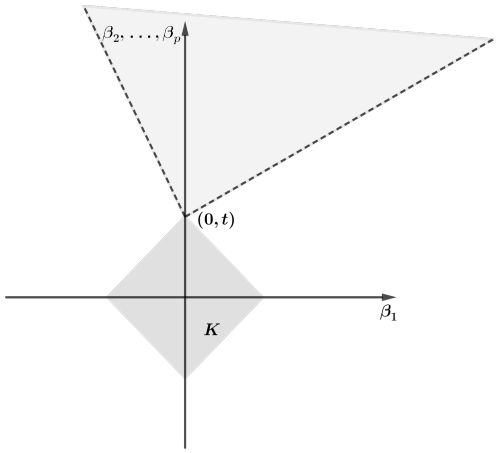

Figure 5 - Attraction domain of a face with Mahalanobis metric.

Consider now a l-dimensional face F such that $\hat{\boldsymbol{\beta}}_{Lasso}(t) = P_K\left(\hat{\boldsymbol{\beta}}_{ols}(y)\right) \in F$. Any small enough open ball centered on $\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})$ is necessarily mapped on the face F. It's possible to get, in this ball, a small l-dimensional subspace, parallel, on the Mahalanobis metric, to the face F. As the projection P_K preserves distance (see Figure 6), the derivative of the projection P_K on the point $\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})$ can be given by the following matrix:

$$dP_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right) = \begin{bmatrix} \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}_{l \times l} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, $\operatorname{div}\left[dP_K\left(\hat{\boldsymbol{\beta}}_{ols}(\boldsymbol{y})\right)\right] = l$. But the dimension of the face l is exactly the number of selected covariates and the results follows.

4 The case $p > n$

Zou and Hastie (2005) proposed the elastic net estimator as an alternative to the Lasso. They pointed out that if the number of covariates p is greater than the number of observations, the Lasso selects at most n covariates. It implies that Lasso is not a very satisfactory variable selection method if $p > n$. They claim that this deficiency comes from the nature of the convex optimization problem that defines the estimator, but they do not present any other explanation or proof for

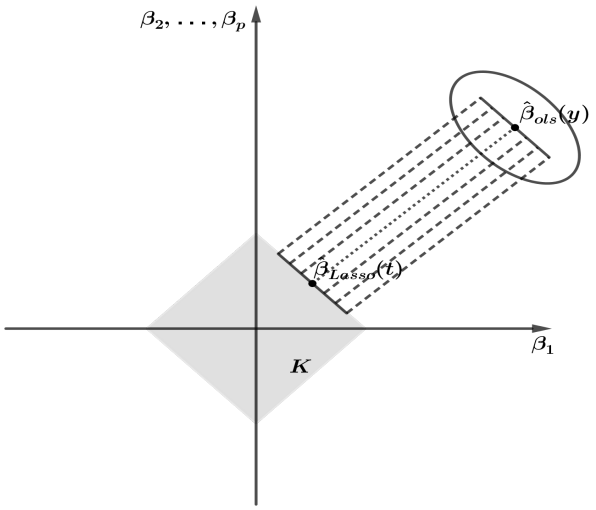

Figure 6 - Projection over a face.

this fact. There is a great source of confusion here. Lasso is defined based on the minimum squares estimator and if the tunning parameter t is such that, for example, $t = \left\|\hat{\boldsymbol{\beta}}_{ols}\right\|_1$ then $\hat{\boldsymbol{\beta}}_{Lasso} = \hat{\boldsymbol{\beta}}_{ols}$ and certainly the number of covariates selected is p. Such confusion is recurrent and quite frequent is statistics forums, and the answers posted are somewhat incomplete. We develop a more in-depth discussion of this issue.

The linear transformation $\boldsymbol{X}$ can't be injective if the dimension of the image of $\boldsymbol{X}$ is $k < n$. Then $\text{Ker}\boldsymbol{X}$ has dimension $n-k$. The geometric construction developed in the previous sections remains fully valid, that is, after defined the value for the tunning parameter t we want to find the vector in K_p as close as possible to the data vector. Using the orthogonal projection of $\boldsymbol{y}$ in $\text{Im}(\boldsymbol{X})$ and any generalized inverse of $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}_{ols}$ is obtained. Again it is possible to build the family of ellipsoids $\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right)' \boldsymbol{X}'\boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right) = c$. The difference here is that these ellipsoids are singular in the sense that they are in some subspace of dimension k. If the value of c is changed until the ellipsoid get a tangent point with the hyperface $\|\boldsymbol{\beta}\|_1 = t$ we have a Lasso estimate, which we will call momentarily $\hat{\boldsymbol{\beta}}^p_{Lasso}$. This estimative solves the variational take of problem $\min \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|$ restricted to $\|\boldsymbol{\beta}\|_1 \leq t$. As $\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}^p_{Lasso} + \boldsymbol{z}\right) = \boldsymbol{X}\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}_{Lasso}\right)$, we have that $\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}_{Lasso} + \boldsymbol{z}$ is also a solution of the minimization problem for all $\boldsymbol{z}$ belonging to $\ker \boldsymbol{X}$.

Consistent with the Lasso estimation philosophy of shrinkage and covariate selection, it is reasonable to choose among the solutions $\hat{\boldsymbol{\beta}}^p_{Lasso} + \boldsymbol{z}$ one of minimum norm, that is, we have to solve a new minimization problem

$$\min \|\boldsymbol{\beta}\|_1, \text{ restricted to } \boldsymbol{\beta} = \boldsymbol{\beta}^p_{lasso} + \boldsymbol{z},\ \boldsymbol{z} \in \ker \boldsymbol{X}. \tag{2}$$

The solution to this problem is simple and follows from the proposition.

Proposition (Boyd-Vandenberghe, page 141, 2004) The convex optimization problem

$$\begin{array}{ll} \min & f \\ \text{restricted} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{array},$$

has solution $\boldsymbol{x}^*$ with $\nabla f(\boldsymbol{x}^*)$ orthogonal to the $\ker \boldsymbol{A}$.

Proof:

Assuming that all coordinates of $\boldsymbol{\beta} = \boldsymbol{\beta}^p_{lasso} + z$ are positive, we have $\nabla\|\boldsymbol{\beta}\|_1 = (1,1,\ldots,1)$. If some coordinate is negative, simply place -1 in the corresponding position. Since $\ker \boldsymbol{X}$ has dimension $p-k$ this subspace for dimensionality reasons can't have empty intersection with all coordinate subspaces of dimension n because $\dim \ker \boldsymbol{X} = p-k$ and $\dim \ker \boldsymbol{X} + n = p-k+n > p$. Therefore there is a coordinate plane of dimension less than or equal to n that intercepts the subspace $\{\hat{\boldsymbol{\beta}}^p_{Lasso} + \boldsymbol{z}\,;\ \boldsymbol{z} \in \ker \boldsymbol{X}\}$. We can suppose, without loss of generality, that $\ker \boldsymbol{X}$ is not parallel to a hyperface $\|\boldsymbol{\beta}\|_1 = t$. In this case a vector normal to the subspace $\{\hat{\boldsymbol{\beta}}^{\boldsymbol{p}}_{Lasso} + \boldsymbol{z}\,;\ \boldsymbol{z} \in \ker \boldsymbol{X}\}$ can't be parallel to the vector $(1,1,\ldots,1)$. Therefore the solution of the minimization problem (2) can only occurs in the intercession of $\{\hat{\boldsymbol{\beta}}^p_{Lasso} + \boldsymbol{z}\,;\ \boldsymbol{z} \in \ker \boldsymbol{X}\}$ with a coordinate subspace. The vectors of this intersection are then candidates for to be a Lasso estimative and they have at most n non-zero covariates (Figure 7).

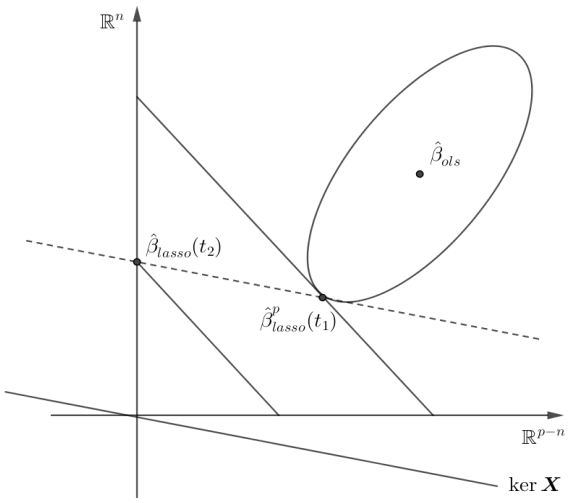

Figure 7 - Lasso selects at most n covariates.

5 Conclusion

The theory of lasso estimators is strongly based on geometric constructions, although it is presented as a convex optimization problem. In this paper it is shown that using basic linear algebra and geometric arguments gives a greater intuitive understanding of the basic facts of the theory.

Acknowledgements

We would like to thank referees and editors for their valuable suggestions.

REIS, C. J.; CARVALHO, L. D.; CHAVES, L. M.; SOUZA, D. J. Três provas heurísticas matemáticas simples da teoria de LASSO. *Rev. Bras. Biom.,* Lavras, v.38, n.2, p.243-256, 2020.

- *RESUMO: Neste artigo são estudado três fatos relevantes sobre a teoria de Lasso (Least absolute shrinkage and select operator): As estimativas seguem curvas lineares por partes em relação ao parâmetro de ajuste; o número de covariáveis selecionadas diferentes de zero é um estimador imparcial de seus graus de liberdade e quando o número de covariáveis p é maior que o número de observações n, no máximo n covariáveis são selecionadas. Esses resultados são bem conhecidos e descritos na literatura, mas sem demonstrações simples. Apresentamos, com base em uma abordagem geométrica, provas heurísticas simples e intuitivas para esses resultados.*

- *PALAVRAS-CHAVE: Seleção de covariáveis; graus de liberdade; Lema de Stein*

References

BELLONI, A.; CHERNOZHUKOV, V.; WANG, L. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, v.42, n.2, p.757-788, 2014.

BOYD, S; VANDENBERGHE, L. *Convex optimization*. Cambridge: Cambridge University Press, 2004.

CAVALCANTE, L.; BESSA, R. J.; REIS, M.; BROWELL, J. LASSO vector autoregression structures for very short-term wind power forecasting. *Wind Energy*, v.20, n.1, p.657–675, 2017.

EFRON, B . The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, v.99, n.467, p.619-642, 2004.

EFRON, B.; HASTIE, T.; JOHNSTONE, I.; TIBSHIRANI, R. Least angle regression. *The Annals of Statistics*, v.32, n.2, p.407-499, 2004.

HASTIE, T.; TIBSHIRANI, R. *Statistical learning with sparsity: The Lasso and generalizations*. Atlanta: Chapman and Hall, 2015.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Thecnometrics*, v.12, n.1, p.51-67, 1970.

KATO, K. On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, v.100, n.1, p.1338-1352, 2009.

LEE, T.-F.; CHAO, P. J.; TING, H. M.; CHANG, L.; HUANGY, Y. J.; WU, J. M.; WANG, H. Y.; HORNG, M. F.; CHANG, C. M.; LAN, J. H.; HUANG, Y. Y., FANG, F. M. LEUNG, S. W. Using Multivariate Regression Model with Least Absolute Shrinkage and Selection Operator (LASSO) to Predict the Incidence of Xerostomia after Intensity-Modulated Radiotherapy for Head and Neck Cancer. *PLOS ONE*, v.9, n.2, p.1-11, 2014.

LEE, J. D.; SUN, D. L.; SUN, Y.; TAYLOR, J. E. Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, v.44, n.3, p.907-927, 2016.

MANGALATHU, S.; JEON, J.-S.; DESROCHES, R. Critical uncertainty parameters influencing seismic performance of bridges using Lasso regression. *Earthquake Engineering & Structural Dynamics*, v.47, n.1, p.784–801, 2018.

MUSORO, J. Z.; ZWINDERMAN, A. H.; PUHAN, M. A.; RIET, G. T.; GESKUS, R. B. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Medical Research Methodology*, v.14, n.116, p.1-13, 2014.

TIBSHIRANI, R. Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*, v.58, n.1, p.267-288, 1996.

TIBSHIRANI, R.; TAYLOR, J. Degrees of freedom in LASSO problems. *The Annals of Statistics*, v.40, n.2, p.1198-1232, 2012.

ZOU, H; HASTIE, Y. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, v.67, n.2, p.301-320, 2005.

Received on 10.07.2019.

Approved after revised on 21.11.2019.