*The Canadian Journal of Statistics*
Vol. xx, No. yy, 20??, Pages 1–25
*La revue canadienne de statistique*



# The Linear Lasso: A Location Model Resolution

D.A.S. FRASER[1] and Mylène BÉDARD[2]*

[1]*Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada*
[2]*Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec, Canada*

*Key words and phrases:* correlation; directed Lasso; inference; least squares; regression; variable selection

*MSC 2020:* Primary 62J07; secondary 62J20

*Abstract:* We use location model methodology to guide the least squares analysis in the Lasso problem of variable selection and inference. The nuisance parameter is taken to be an indicator for the selection of explanatory variables and the interest parameter is the response variable itself. Recent theory eliminates the nuisance parameter by marginalization on the data space and then uses the resulting distribution for inference concerning the interest parameter. We develop this approach and find: that primary inference is essentially one-dimensional rather than $n$-dimensional; that inference focuses on the response variable itself rather than the least squares estimate (as variables are removed); that computation is relatively easy; that a scalar marginal model is available; and that ineffective variables can be removed by distributional tilt or shift. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

*Résumé:* Nous utilisons la méthodologie des modèles de position dans le but de guider l'analyse des moindres carrés dans le cadre du problème de type Lasso, c'est-à-dire de sélection et d'inférence de variables. Le paramètre de nuisance est une variable indicatrice relative à la sélection des variables explicatives alors que le paramètre d'intérêt est la variable réponse. Selon des développements théoriques récents, le paramètre de nuisance est éliminé par marginalisation sur l'espace de données; la distribution résultante est alors utilisée pour effectuer une inférence sur le paramètre d'intérêt. Nous développons cette approche et constatons: que l'inférence primaire est essentiellement unidimensionnelle plutôt que $n$-dimensionnelle; que l'inférence se concentre sur la variable réponse elle-même plutôt que sur l'estimé des moindres carrés (à mesure que les variables sont supprimées); que les calculs sont relativement faciles; que le modèle marginal scalaire est disponible; et que les variables inefficaces peuvent être supprimées par inclinaison ou glissement distributionnel. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

## 1. INTRODUCTION

The Lasso (least absolute shrinkage and selection operator) approach is a regression method that simultaneously performs variable selection and parameter estimation. Introduced in the statistical literature by Tibshirani (1996), its goal is to enhance the accuracy of predictions while retaining the interpretability aspect of the resulting statistical model. The idea behind Lasso is to force the sum of the absolute regression coefficients to be smaller than a predetermined value, which consequently forces some coefficients to be null. It was initially introduced in the context of linear regression and least-squares estimation; its applicability is however much wider, including for instance generalized linear models and proportional hazards models.

The Lasso method essentially consists in solving a constrained minimization problem over the parameter space, i.e., over all possible regression coefficients. The objective function to minimize may vary in different contexts, but the constraint on the sum of absolute regression coeffi-

* *Author to whom correspondence may be addressed.*
*E-mail: mylene.bedard@umontreal.ca*

cients is generally present (although variations of this penalty may be used in different versions of Lasso). Geometrically, the Lasso is usually illustrated by comparing the shape of its constraint region to that arising from other penalty functions; this enhances the fact that Lasso's constraint region has several corners and edges over which one or several regression coefficients are null. As this region gets smaller, the contours of the objective density function become increasingly likely to hit one of those corners/edges, therefore leading to a regression model with fewer parameters (see Figure 2.2 of Hastie, Tibshirani, & Wainwright, 2015).

There is a very extensive literature about Lasso. Several variations of the method have been proposed, and various algorithms have been developed to solve the convex optimization problem it generates; Hastie, Tibshirani, & Wainwright (2015) and the references therein provide a detailed summary of these advances. One limitation of Lasso is that it becomes computationally intractable with large datasets, which are common in our current era of big data and sophisticated statistical models. Another one is that in presence of high pairwise correlations in a group of variables, Lasso tends to select only one variable and does not care which one it selects (see Zou & Hastie, 2005).

In this paper, we consider a linear regression context and address the variable selection and parameter estimation problems from a geometrical viewpoint. Given data on many variables, we identify one of particular importance (the response variable) and seek a small selection of others (the explanatory variables) that give good linear prediction of the interest variable. The vector containing the observed responses serves as the focal point of the interpretation, around which vectors containing observations from explanatory variables gravitate. The angles between the response and explanatory vectors, and among pairs of explanatory vectors, provide the basic input for a geometric analysis guided by location model theory.

To this end, the familiar location-scale standardization is applied to each variable; we also add sign standardization so that all explanatory variables be positively correlated with the response variable. This last modification is not required in the final implementation of the method, but does make the problem easier to visualize and, in turn, helps justifying the steps leading to our final approach. We then focus on normal linear models as a way to handle least squares and then rely on the above geometry to propose a simple resolution for the Lasso problem. The response variable is taken as the interest parameter and an indicator function for the selection of explanatory variables is used as the nuisance parameter. By projecting all pertinent information from explanatory variables on a single line in the space, we find that the available information about the response variable can be summarized in a one-dimensional distribution that is characterized by its prediction variance.

This allows fine-tuning the objective function of the standard Lasso to more closely agree with its intended purpose; this function however remains unaltered in the saturated cases. As a consequence, the elimination of ineffective variables becomes easier, avoiding the usual iteration procedure and making the problem largely dimension free. We eliminate seemingly underperforming explanatory variables by a tilt or moment generating type modification, and discard negative coefficients under the distributional shift. We obtain an iteration-free resolution of the Lasso that is essentially explicit, with out-of-sample prediction accuracy well exceeding that of the regular Lasso. We use the term Linear Lasso for our procedure to emphasize that the minimization trajectory above the parameter space is a straight line; this is in contrast to the regular Lasso, which has multiple segments of lines and curves. The end result is a selection approach that presents some similarities with that of Fan & Lv (2008), but in which variable selection is built-in and thus avoids the need for preliminary screening.

The resulting model may then be used for inference or prediction; we however emphasize that models obtained from data-driven variable selection procedures, such as the Lasso and Linear Lasso, should be handled with care. For instance, confidence intervals or statistical tests that

are naively performed on such models do not necessarily enjoy the advertised coverage/level. Indeed, the variables that are selected tend to be the significant ones and ignoring this can falsely amplify the apparent connection between variables. Statistical inference on models stemming from such methods should instead rely on recent developments about post-selection inference; see, for instance, Berk et al. (2013), Lee & Taylor (2014), Taylor & Tibshirani (2015), Lee et al. (2016), and Zhao, Witten, & Shojaie (2021).

In Section 2, we record background and notation. The stochastic framework is analyzed in Section 3 from a geometrical viewpoint, and is linked to location model theory. Section 4 introduces what is viewed as the latent or simulation model and shows that least squares is effectively equivalent to routine normal analysis, with a very simple example given in the subsequent section. Section 6 determines how much response distribution is hidden in a selection of explanatory variables and records the corresponding selection model. Sections 7 to 9 show how to construct a reduced set of explanatory variables, while Sections 10 and 11 illustrate the theory with two real data examples. We conclude with a discussion in Section 12.

This paper preserves the unique and original views of the late Professor D.A.S. Fraser. Appendix B is intended as an accompanying document containing section-by-section clarifications and details about the concepts introduced (this appendix should ideally be read side by side with the main document).

## 2. BACKGROUND AND NOTATION

In this work, we consider a scalar variable $y$ of particular interest and $r$ potential explanatory scalar variables $x_1, \ldots, x_r$, typically with $r$ large. We then seek a small sub-selection of the explanatory variables that provides acceptable or good prediction for the response $y$; we suppose that these predictors have subscripts in $J_s = \{j_1, \ldots, j_s\}$. To perform this task, we have $n$ observations on the $1 + r$ variables, providing full data as an $n \times (1 + r)$ array $(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_r)$, or as $1 + r$ vectors of length $n$. The location, scaling, and sign of the variables are typically conventional so we can widely apply standardizations. Accordingly, we hereafter assume that each column vector has been location-scale standardized so the average of the coordinates is zero and their standard deviation is one. To keep notation simple, we still refer to the modified data as $(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_r)$.

The usual Lasso procedure is to minimize, over choice of regression coefficients $\beta$, the expression

$$\sum_i (y_i - \mathbf{X}_i\beta)^2/2n + \gamma\Sigma_j|\beta_j| \,, \tag{1}$$

where $\mathbf{X}_i$ is the $i$-th row of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_r)$. The first term is a rescaled residual sum-of-squares from the linear model and the second term is a Lagrangian penalty to force fewer selected predictors, with $\gamma$ for tuning; see Hastie, Tibshirani, & Wainwright (2015). The few non-null regression coefficients $\beta$ retained by the Lasso can then be combined with observations from the selected explanatory variables to predict the associated response $y$.

Following the previous location-scale standardization, each data vector has length $n^{1/2}$ and the correlation between any pair of vectors is obtained by dividing the corresponding inner product by $n$. We view the pairwise correlations of the $1 + r$ vectors as the intrinsic data for the problem. Specifically, let $c = (c_j)$ be the correlations between $\mathbf{y}$ and the $\mathbf{x}_j$ vectors, and let $C = (c_{ij})$ be the correlations among the $\mathbf{x}_j$ vectors. We can then assemble these terms as a full

correlation matrix

$$\tilde{C} = \begin{pmatrix} 1 & c^t \\ c & C \end{pmatrix} . \tag{2}$$

The use of the letters $c$ and $C$ reminds us that the elements are just cosines of angles between unit data vectors, each conveniently obtained from a corresponding inner product. As with Lasso, these correlation terms are treated as constant.

As explanatory variables $x_j$ can be positively or negatively correlated with the response $y$, we apply a further standardization: any explanatory vector $\mathbf{x}_j$ that is negatively correlated with the response vector $\mathbf{y}$ has its sign reversed. Consequently, all explanatory vectors become positively correlated with the response vector. This modification is notational and for visual convenience only; it does not affect the substance and is in some agreement with the usual regression analysis.

Finally, in order to geometrically represent the vectors' directions relatively to each other, let $\mathbf{u}_y, \mathbf{u}_1, \ldots, \mathbf{u}_r$ be unit versions of the data vectors $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_r$ (i.e., $\mathbf{u}_j = \mathbf{x}_j/\sqrt{n}$). It is convenient to think of $\mathbf{u}_y$ as pointing upward; it is also natural to have the zero point of each vector $\mathbf{u}_y, \mathbf{u}_1, \ldots, \mathbf{u}_r$ placed directly on the origin of some underlying vector space. Under this framework, all $\mathbf{u}_j$ vectors are then directed into the upper half-space $\mathcal{L}^+\mathbf{y}$, above the plane $\mathcal{L}^\perp\mathbf{y}$ perpendicular to $\mathbf{y}$; see Figure 1.

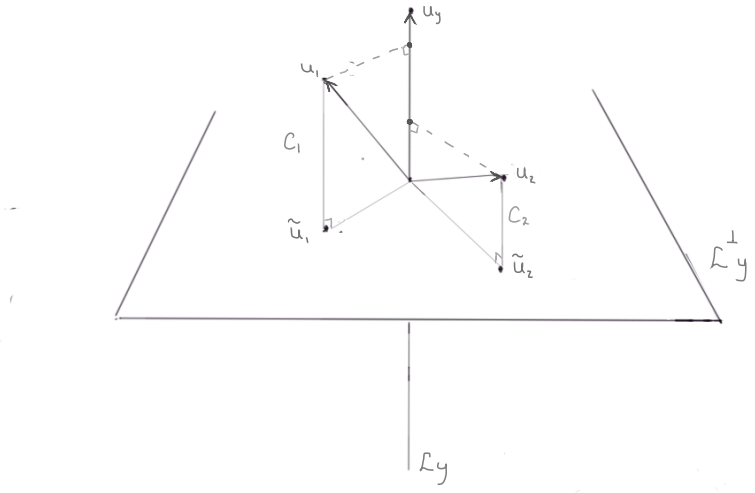


FIGURE 1: The unit response vector $\mathbf{u}_y$ and two (unit) predictor vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, along with their projections $c_j$ to $\mathcal{L}\mathbf{y}$ and residuals $\tilde{\mathbf{u}}_j$ on $\mathcal{L}^\perp\mathbf{y}$.

## 3. LATENT STOCHASTIC MODEL

Having geometrically represented the direction of each data vector, we now impose a distributional structure to pursue our analysis. The reference to correlated data indicates a common stochastic background for the $1 + r$ variables. In our linear regression context, each variable $y, x_1, \ldots, x_r$ can be viewed as a linear function of an $n$-dimensional latent normal distribution on an underlying vector space. The values of such a linear function can be recorded on the line

perpendicular to the contours of this linear function, that is, on the lines formed by the vectors $\mathbf{y}, \mathbf{x}_1, \cdots, \mathbf{x}_r$ and designated $\mathcal{L}\mathbf{y}, \mathcal{L}\mathbf{x}_1, \cdots, \mathcal{L}\mathbf{x}_r$. Conveniently, the use of least squares has algebraic equivalence to symmetric normal location model analysis, where the related distribution theory provides important guidance.

Specifically, for a linear model that corresponds to least squares, we use an $n$-dimensional latent variable space having a rotationally symmetric standard normal distribution; for convenience, we center this distribution at the origin. An observable variable (say, $x_j$) is then a linear function on that space, which can be viewed as a sort of "tap" on the latent stochastics. Each linear function has its own linear contours on the latent variable space, where a contour is formed of all $n$-dimensional latent observations leading to a common point on the linear function. These contours are parallel (hyper)planes that are perpendicular to the function; then, perpendicular to these contours, we find a line that goes through the origin (say, $\mathcal{L}\mathbf{x}_j$) and that indexes the contours of the linear function. Each such line thus records values for the corresponding variable (so values for the variable $x_j$ are recorded on $\mathcal{L}\mathbf{x}_j$) and presents a column of the given data array ($\mathbf{x}_j$ are observations from the variable $x_j$ on $\mathcal{L}\mathbf{x}_j$). The data can then be viewed as giving $n$ values on each observed line in the space, corresponding for instance to successive time points. As these $1 + r$ lines all go through the origin in the latent vector space $\mathbb{R}^n$, the model gives data on a rotationally symmetric normal latent model, and the observed lines provide a type of skewed coordinates that introduce the correlations $c$ and $C$ in the resulting distribution.

With the latent stochastic model, we are able to describe the $1 + r$ variables of interest in terms of their dependence on the $n$ latent variables. These variables of interest can then be presented as "taps" on the latent stochastics, or equivalently as functions on the latent variable space. This alternative modeling format offers advantages, including making explicit the continuity that is present among variables. Such linear functions on the latent space can be called "data generating" given their availability for simulations, or "structural" for their explicit presentation of the dependences. For the full set of variables, we use the data generating format and a choice of expressive but nonstandard notation: $\{y\mathcal{L}\mathbf{y}, x_1\mathcal{L}\mathbf{x}_1, \cdots, x_r\mathcal{L}\mathbf{x}_r\}$, where the lines record the directions of the stochastic "taps", and where the coefficients $y, x_1, \cdots, x_r$ each are standard normal on their respective line, but collectively have correlations recorded as $\tilde{C}$. Then, for the modeling to structure least squares, we use the lower case variables $y, x_1, \ldots, x_r$; these are jointly multivariate normal $(0; \tilde{C})$, that is,

$$y, x_1, \ \ldots, x_r \sim \mathcal{MN}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & c^t \\ c & C \end{pmatrix}\right) . \tag{3}$$

Before pursuing our analysis of the above model, we introduce a further coordinate standardization that makes the model's form and objectives more transparent. We explained above that the response variable $y$, represented as a linear function of the latent variables on the $n$-dimensional space, is recorded on the line $\mathcal{L}\mathbf{y}$; prediction is thus one-dimensional rather than $n$-dimensional, and the analysis involves scalar fitting rather than $n$-dimensional regression fitting. This then leads to the Linear Lasso procedure that allows selection of variables by minimum-number or maximum-variance viewpoints, where the elimination of unproductive $x$ variables is achieved by tilting (or equivalently by shifting) the response distribution along $\mathcal{L}\mathbf{y}$.

By opposition, we outline that the regular least squares modeling seeks a vector $\hat{\mathbf{y}}$ in the $n$-dimensional latent variable space that is as close as possible to $\mathbf{y}$ in terms of the residual sum-of-squares. Expressed differently, it looks for a linear function of the $r$ explanatory variables (which are themselves linear functions in the latent variable space) that will reach this goal. Such an approach thus involves $n$-dimensional regression fitting.

## 4. INFERENCE FROM A PARTICULAR SUBSET OF EXPLORATORY VARIABLES

Using the distributional structure outlined in the previous section, we would like to achieve inference on the response variable $y$ using a particular subset containing $s$ of the $r$ exploratory variables. We suppose that the predictors included in the subset are those with indices in $J_s = \{j_1, \cdots, j_s\}$, meaning that $j_1$ is the original subscript of the first coordinate selected, $j_2$ the original subscript of the second coordinate selected, and so on. Then, using the multivariate normal in (3), we find that the subset of variables labelled by $J_s$ have the joint distribution

$$y, x_{j_1}, \ \cdots \ , x_{j_s} \sim \mathcal{MN}\left( \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} ; \begin{pmatrix} 1 & c_s^t \\ c_s & C_s \end{pmatrix} \right) ,$$

where $c_s$ and $C_s$ designate the correlations restricted to the subset $J_s$.

For convenience, we assume that the full correlation matrix $\tilde{C}$ is nonsingular, and return to greater generality later. We then use the formulas of conditional probability to obtain the conditional distribution of the response variable $y$ given the subset of explanatory variables with subscripts in $J_s$. For a single $x$ variable, the familiar conditioning formula is $y|x \sim \mathcal{N}(\sigma_{y,x}\sigma_{x,x}^{-1}x \ ; \ \sigma_{y,y} - \sigma_{y,x}\sigma_{x,x}^{-1}\sigma_{x,y})$, where $\sigma_{x,x}$ and $\sigma_{y,y}$ are the variances of $x$ and $y$ respectively, and $\sigma_{x,y} = \sigma_{y,x}$ is the covariance term. Applying the corresponding vector version then gives

$$y|x_{j_1}, \ldots, x_{j_s} \sim \mathcal{N}\left( c_s^t C_s^{-1} \begin{pmatrix} x_{j_1} \\ \vdots \\ x_{j_s} \end{pmatrix} ; 1 - c_s^t C_s^{-1} c_s \right) \tag{4}$$

on the line $\mathcal{L}\mathbf{y}$. This represents the distribution of the response $y$ given that we have observed the explanatory variables $x_{j_1}, \ldots, x_{j_s}$, and this forecasts the value $c_s^t C_s^{-1}(x_{j_1}, \cdots, x_{j_s})^t$ on $\mathcal{L}\mathbf{y}$. We refer to the value $c_s^t C_s^{-1}(x_{j_1}, \cdots, x_{j_s})^t$ as the $y$-content of the subset $J_s$; it algebraically corresponds to the least squares prediction given the observed explanatory variables $x_{j_1}, \ldots, x_{j_s}$ and data vectors $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_s}$. This $y$-content, in turn, is normally distributed with mean 0 and standard deviation $\{c_s^t C_s^{-1} c_s\}^{1/2}$ on $\mathcal{L}\mathbf{y}$. The standard deviation can be interpreted as the fraction of $y$ variability inherent in the subset of explanatory variables; it represents the fraction of the marginal $y$ distribution that is captured by the subset of predictors $J_s$.

## 5. A VERY SIMPLE EXAMPLE

Consider a very simple example with $n = 3$ and $r = 2$, as indicated by Figure 1. In that context, the number of possible selected variables is either $s = 1$ or $s = 2$. Suppose the data array is

$$(\mathbf{y}\ \mathbf{X}) = (\mathbf{y}\ \mathbf{x}_1\ \mathbf{x}_2) = \begin{pmatrix} 1.000\,000 & 0.900\,000 & 0.600\,000 \\ 0.000\,000 & 0.435\,890 & 0.400\,000 \\ 0.000\,000 & 0.000\,000 & 0.692\,820 \end{pmatrix} ;$$

the vector $\mathbf{y}$ points upward and each of the data vectors is of unit length. If we place the zero points of these vectors on the origin of the latent vector space, then they correspond to the unit

TABLE 1: General expressions for individual predictions (second column) and their SD (third column), for each possible subset of selected predictors.

| Source | $y$-content | SD of content = fraction of explained $y$-variability |
|---|---|---|
| $\{x_1\}$ | $c_1\, x_1$ | $c_1$ |
| $\{x_2\}$ | $c_2\, x_2$ | $c_2$ |
| $\{x_1, x_2\}$ | $(c_1\ c_2) \begin{pmatrix} 1 & c_{12} \\ c_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ | $\left\{ (c_1\ c_2) \begin{pmatrix} 1 & c_{12} \\ c_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\}^{1/2}$ |

TABLE 2: Values of $y$-content (third column) and SD = fraction of explained $y$-variability (fourth column), for each possible subset of selected predictors.

| Source | $\mathbf{y}$ axis projection | $y$-content | SD of content |
|---|---|---|---|
| From $\{x_1\}$ | 0.9 | $0.9\, x_1$ | 0.9 |
| From $\{x_2\}$ | 0.6 | $0.6\, x_2$ | 0.6 |
| From $\{x_1, x_2\}$ | 0.902 | $0.963\, x_1 - 0.088\, x_2$ | 0.902 |

vectors $\mathbf{u}_y$, $\mathbf{u}_1$, and $\mathbf{u}_2$. The pairwise correlations between these vectors, all positive, are

$$\tilde{C} = \begin{pmatrix} 1 & c^t \\ c & C \end{pmatrix} = \begin{pmatrix} 1.000\,000 & 0.900\,000 & 0.600\,000 \\ 0.900\,000 & 1.000\,000 & 0.714\,356 \\ 0.600\,000 & 0.714\,356 & 1.000\,000 \end{pmatrix} .$$

After standardizing each data vector so that the means be equal to zero and the standard deviations to one, the latent stochastic model is expressed as $(y\mathcal{L}\mathbf{y}, x_1\mathcal{L}\mathbf{x}_1, x_2\mathcal{L}\mathbf{x}_2)$, where $(y, x_1, x_2)$ is a multivariate normal $(\mathbf{0}, \tilde{C})$ with $\mathbf{0}$ a vector of zeroes. In Figure 1, the latent model is a three-dimensional standard normal centered at the origin. The form of the model above $\mathcal{L}^{\perp}\mathbf{y}$ has a near-reflection through the origin, giving a near-duplicate below $\mathcal{L}^{\perp}\mathbf{y}$ (model has no cubic terms).

In this example, we only have three possibilities in terms of variable selection, namely $\{x_1\}$, $\{x_2\}$, and $\{x_1, x_2\}$. For each possible case, general expressions for the $y$-content and the fraction of explained response variability are provided in Table 1. The last line in the table uses results for the full set of explanatory variables. Using the data summarized in $\tilde{C}$ above, the general expressions in Table 1 lead to the models and values in Table 2. We then see that $\mathbf{x}_1$ has the largest projection on the $\mathbf{y}$ axis, and thus the variable $x_1$ is associated with the largest fraction of explained $y$-variability; including $x_2$ then adds very little to the fraction of $y$-variability captured by the model. This is unsurprising considering that $\mathbf{x}_2$ is more correlated with $\mathbf{x}_1$ than it is with $\mathbf{y}$.

Figure 2 illustrates the fraction of marginal $y$-density that is available from each explanatory variable separately. Figure 3 depicts the fraction of marginal $y$-density that is available from the best selection of size 1 and 2, respectively. In both cases, we included the full $y$ density for comparison. Although the example is simple, it does illustrate that for any subset of explanatory variables, there is an available formula that quantifies the $y$-information contained in that specific subset. Obviously, the number of non-trivial subsets in this example is very small ($r!/\{s!(r-s)!\} = 3!/\{2!1!\} = 3$), but this number grows exponentially with larger data arrays.

## 6. HOW MUCH $Y$ DISTRIBUTION IS HIDDEN IN SELECTED $X$ VARIABLES

In Section 4, we used the conditional distribution of $y$ given a subset $J_s$ of explanatory variables to make an inference about the response variable. Suppose now that we address the larger question of how much $y$-content distribution is accessible from an arbitrary selection $J_s = \{j_1, \cdots, j_s\}$ of explanatory variables. The distribution of such variables can be presented in data-generating form as $\{x_{j_1}\mathcal{L}\mathbf{x}_{j_1}, \cdots, x_{j_s}\mathcal{L}\mathbf{x}_{j_s}\}$, where the coefficients $x_{j_1}, \ldots, x_{j_s}$ are centered multivariate normal

$$x_{j_1}, \ldots, x_{j_s} \sim \mathcal{MN}\left( \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}; C_s = \begin{pmatrix} 1 & c_{j_1 j_2} & \cdots & c_{j_1 j_s} \\ c_{j_2 j_1} & 1 & & c_{j_2 j_s} \\ \vdots & & \ddots & \vdots \\ c_{j_s j_1} & \cdots & c_{j_{s-1} j_s} & 1 \end{pmatrix} \right) .$$

The variables in $J_s$ can then be linearly combined so as to yield a prediction $c_s^t C_s^{-1}(x_{j_1}, \ldots, x_{j_s})^t$ that is equivalent to the least squares prediction, as in Section 4. The distribution of this linear function is a centered normal on $\mathcal{L}\mathbf{y}$ with SD $\sigma = \{c_s^t C_s^{-1} c_s\}^{1/2}$. The fraction of $y$-variability explained by the $J_s = \{j_1, \cdots, j_s\}$ selection of explanatory variables is thus the fraction $\sigma(J_s) = \{c_s^t C_s^{-1} c_s\}^{1/2}$ of the marginal $\mathcal{N}(0,1)$ distribution for the response $y$; see Figure 2.

To account for the various possible subsets of predictors, we now introduce a parameter $\delta$ in the above distribution. Specifically, we let $\delta = (\delta_1, \cdots, \delta_r)$ be an indicator variable for the presence $(1)$ or absence $(0)$ of each of the $r$ available explanatory variables. We then define the vector $c_\delta = (c_{j\delta_j})$ and matrix $C_\delta = (c_{i\delta_i, j\delta_j})$, in which elements with a null subscript are simply excluded; this leads to a vector $c_\delta$ of length $\Sigma_j \delta_j$ and a matrix $C_\delta$ of dimension $\Sigma_j \delta_j \times \Sigma_j \delta_j$. Then, from the preceding paragraph, the fraction of $y$-density in our selection $\delta$ is expressed as a fraction $\sigma(\delta)$ of a standard normal distribution, where $\sigma(\delta) = \{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}$. The statistical density (with parameter $\delta$) of this $y$-information coming from the specified explanatory variables therefore is

$$\sigma(\delta) \text{ of } \phi(y) \equiv \sigma(\delta) \text{ of } \frac{1}{(2\pi)^{1/2}} \exp\{-y^2/2\} .$$

This presents a relative density that is recorded as a fraction of a standard normal distribution. It can also be recorded, more formally, in statistical model format:

$$f(y; \delta) = \sigma(\delta)\phi(y) .$$

The density arising from any component (or group of components) labeled $\delta$ will typically not integrate to one, as it is only recording the fraction of a hidden $y$ distribution that is accessible from the $x$ variables.

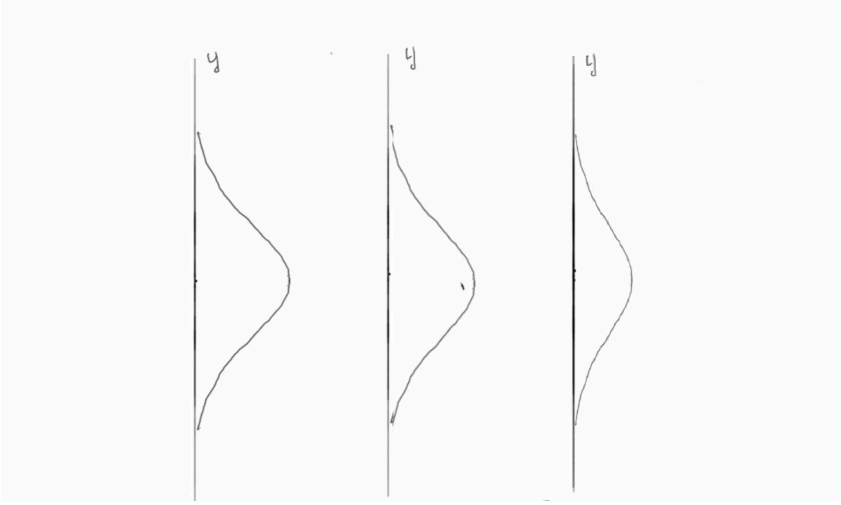


FIGURE 2: For the simple example in Section 5, we record the fraction $0.6$ of $y$-density inherent in $x_2$ (right), $0.9$ in $x_1$ (middle), and finally $1$ in the target variable $y$ (left); percentages are illustrated by fractioning the height of the standard normal.

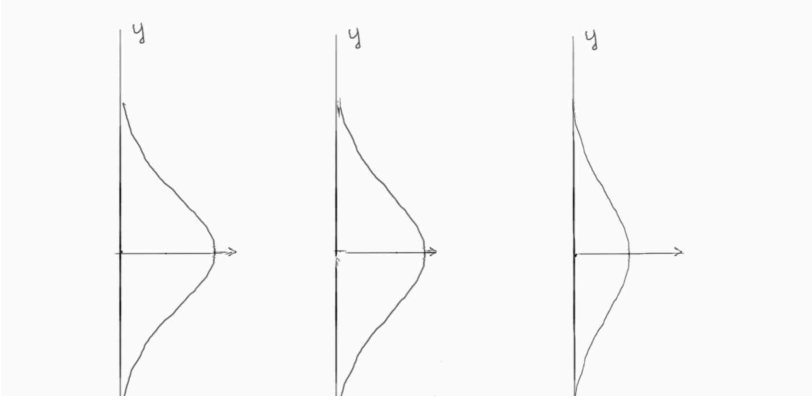


FIGURE 3: For the simple example in Section 5, we record the fraction $0.9$ of $y$-density coming from $x_1$ alone (right), $0.902$ from $x_1$ and $x_2$ (middle), and finally $1$ for the target variable $y$ (left).

## 7. SUBSTANTIAL REDUCTION IN THE SET $J_S$ OF EXPLANATORY VARIABLES

The objective of the Lasso is to minimize the residual sum-of-squares with respect to $\beta$, subject to a penalty on the total magnitude of the regression coefficients. Having now projected the pertinent $y$-information contained in the explanatory variables on the line $\mathcal{L}\mathbf{y}$, we can recast the objective as a model search along the response line. In light of the statistical context discussed in the previous section, our primary objective is now to find a small selection $J_s$ of explanatory variables that collectively give large $y$ variance (and thus high distribution) for the $y$-content. This involves a trade-off between the variance $\sigma^2(\delta) = c_\delta^t C_\delta^{-1} c_\delta$ and the cardinality $\Sigma\delta_j$ of the selection.

The variance formula for a single $x_j$ is straightforward; it gives the corresponding squared correlation $c_j^2$, and is thus immediately available. However, if we seek an additional explanatory variable, the variance is typically not the sum $c_1^2 + c_2^2$ of the individual variances; it generally

includes weights from the inverse of the correlation matrix $C$. The variance for the $y$-content in the variables $x_1$ and $x_2$ is $c_1^2c^{11} + 2c_1c_2c^{12} + c_2^2c^{22}$, where the $c$'s with two raised indices are elements of $C^{-1}$; hence, the maximum variance given a cardinality $\Sigma\delta_j$ does not seem easily available. A direct search for this problem has an exponential order of computation and could be viewed as unfeasible with large data.

We now briefly discuss a simpler computational route, appropriate with any number of predictors, that drops components by walking along $\mathcal{L}\mathbf{y}$. The method has the potential of removing a large batch of underperforming explanatory variables (Section 8) and it can also be fine-tuned so as to obtain a one-by-one procedure; the latter is discussed in Section 9. Our marginal model $f(y;\delta)$ is a centered normal with variance $\sigma^2(\delta) = c_\delta^t C_\delta^{-1} c_\delta$; this model depends on the selection $\delta$ entirely through the scaling (or spread) $\sigma$ of its distribution. We then propose to shift this distribution on the positive $y$-axis by applying an exponential tilt $\exp\{\gamma y\}$ to the distribution; see Appendix A. In doing so, the center of the distribution goes from zero to $\gamma$ and we then retain only positive regression coefficients. This eliminates underperforming $x_j$ and provides a distributional analog of the penalty function approach in Tibshirani (1996). The $x_j$ eliminated by this process are those with small $c_j$ values. This substantial reduction is easy, entirely based on small correlations, and does not require iterations. It is available here because of our direct search for $y$-content rather than the focused use of fitted regression.

## 8. MANY EXPLANATORY VARIABLES

The Linear Lasso uses location model methodology to obtain an ordering on sets of explanatory variables $J_s$. The development uses a latent normal space, projects relevant information on the response line, and discards predictors on the basis of their correlations $\{c_j\}$ with the response. The simplicity evolves from the focus on the one-dimensional $y$-content distribution associated to sets $J_s$. The ordering is used to drop least contributing explanatory variables as part of a stepwise regression, and is consistent when combining or reducing such sets of variables.

Conveniently, in contexts where there is a very large number of explanatory variables ($r >> n$ for instance), the consistent procedure would still be to drop variables in accord with the correlations. Therefore, for stepwise regression, we can quite generally drop the small $c_j$ variables as part of the reduction process. In our approach, making use of correlations is therefore built-in and replaces the need for preliminary screening as used in Fan & Lv (2008) in the Lasso context.

## 9. ONE-BY-ONE

When few variables are left in the model, we might want to account for the correlations $C$ between predictors by eliminating the variable $x_j$ that leads to the smallest decrease in the variance term $\sigma^2(\delta) = c_\delta^t C_\delta^{-1} c_\delta$. We thus apply backward regression with an exclusion criterion based on the variance of the $y$-content distribution.

To this end, suppose that $m$ variables have been eliminated on the basis of their correlations with the response, and that the original subscripts of these variables are listed in $\mathcal{M}$; here, $m$ is a value selected by the user. There are thus $r - m$ variables left to order; iteratively, we proceed as follows.

1. Initialize $\delta^{(0)}$ such that $\delta_j = 0$ for $j \in \mathcal{M}$ and $\delta_j = 1$ for $j \notin \mathcal{M}$, $j = 1, \ldots, r$. The cardinality $\Sigma\delta_j$ of the selection $\delta^{(0)}$ is then $r - m$.
2. Let $\delta^{(1)} = \delta^{(0)}$ and suppose that $k_1 \notin \mathcal{M}$. Set $\delta_{k_1} = 0$, where $k_1$ minimizes

$$\sigma^2(\delta^{(0)}) - \sigma^2(\delta^{(1)}) = c_{\delta^{(0)}}^t C_{\delta^{(0)}}^{-1} c_{\delta^{(0)}} - c_{\delta^{(1)}}^t C_{\delta^{(1)}}^{-1} c_{\delta^{(1)}} \ ;$$

include $k_1$ in $\mathcal{M}$ and the cardinality $\Sigma\delta_j$ becomes $r-m-1$.

3. Let $\delta^{(2)} = \delta^{(1)}$ and suppose that $k_2 \notin \mathcal{M}$. Set $\delta_{k_2} = 0$, where $k_2$ minimizes

$$\sigma^2(\delta^{(1)}) - \sigma^2(\delta^{(2)}) = c^t_{\delta^{(1)}} C^{-1}_{\delta^{(1)}} c_{\delta^{(1)}} - c^t_{\delta^{(2)}} C^{-1}_{\delta^{(2)}} c_{\delta^{(2)}} \ ;$$

include $k_2$ in $\mathcal{M}$ and the cardinality $\Sigma\delta_j$ becomes $r-m-2$.

4. Repeat these steps until all variables have been removed; the resulting ordering provides a progressive selection of variables for various cardinalities $\Sigma\delta_j$.

Combining the substantial reduction of Sections 7 and 8 with the present one-by-one procedure gives the process for the Linear Lasso; this new approach works directly on the response line, after the data have been sign standardized. Neither the Linear nor the regular Lasso can be expected to fully achieve its desired optimization, but the Linear Lasso uses $y$-change directly as the desired target and has the substantial property of avoiding iterative steps. The tuning parameter for the Linear Lasso is the parameter $\gamma$ in the exponential tilt or shift.

## 10. EXAMPLE: CRIME DATA

To illustrate the use of the Linear Lasso, we study the small example on page 10 of Hastie, Tibshirani, & Wainwright (2015). The data, originally taken from Thomas (1990), reports the crime rate per million residents in $n = 50$ U.S. cities. There are $r = 5$ explanatory variables: annual police funding (dollars/resident), people age $\geq 25$ with four years of high school (%), people age 16 to 19 neither in high school nor high school graduates (%), people age 18 to 24 in college (%), and people age $\geq 25$ with $\geq 4$ years of college (%). The resulting data array is $50 \times 6$; the crime rate in the first column of the array is the outcome vector and columns 2 to 6 are the five potential explanatory variables.

A first step consists of standardizing the data so that each column has an average of zero and a standard deviation of one. The standardized data array is $(\mathbf{y}\ \mathbf{x}_1\ \ldots\ \mathbf{x}_5)$. Inner products $c_j = \mathbf{y} \cdot \mathbf{x}_j / n$ $(j = 1, \ldots, 5)$ represent the correlations between the outcome $\mathbf{y}$ and the explanatory variables ($\mathbf{x}$'s). The variables $\mathbf{x}_2$, $\mathbf{x}_4$, and $\mathbf{x}_5$ represent different measures of the population's education level and are all negatively correlated with the crime rate. If we picture the outcome $\mathbf{y}$ as a vector pointing upwards, this implies that vectors $\mathbf{x}_2$, $\mathbf{x}_4$, and $\mathbf{x}_5$ lie in the lower half-space. We invert the sign of these three explanatory variables and use $\mathbf{x}^*_j = -\mathbf{x}_j$ (for $j = 2, 4, 5$) and $\mathbf{x}^*_j = \mathbf{x}_j$ (for $j = 1, 3$); we thus work with vectors that are in the upper half-space $\mathcal{L}^+\mathbf{y}$. The resulting vector of correlations $c$ is then formed of the elements $|c_j| = \mathbf{y} \cdot \mathbf{x}^*_j / n$ $(j = 1, \ldots, 5)$. Similarly, inner products $c_{jk} = \mathbf{x}^*_j \cdot \mathbf{x}^*_k / n$ $(j, k = 1, \ldots, 5)$ form the $5 \times 5$ matrix $C$ of correlations between the $\mathbf{x}^*$ vectors.

From (3), $(y, x^*_1, \ldots, x^*_5)$ is jointly distributed according to a $\mathcal{MN}(\mathbf{0}, \tilde{C})$. Using (4), the predictive distribution for the full model is

$$y | x^*_1, \ldots, x^*_5 \sim \mathcal{N}\left(0.516x^*_1 + 0.206x^*_2 + 0.112x^*_3 - 0.019x^*_4 - 0.097x^*_5; 0.666\right).$$

Of interest is to use a predictive distribution that features a variance as small as possible, while simultaneously relying on a relatively simple model (parsimony). We thus look for a small selection of explanatory variables $J_s$ whose $y$-content $c^t_s C^{-1}_s (x^*_{j_1}, \ldots, x^*_{j_s})^t$ has a high variance; the associated standard deviation then represents the fraction of $y$-variability that is explained by the linear model.

We proceed as expounded in Section 9, starting with the full model and iteratively removing variables. In this example, the vector of correlations between the outcome $\mathbf{y}$ and explanatory variables $\mathbf{x}^*_j$ is $c = (0.533, 0.135, 0.323, 0.175, 0.026)^t$. Let us suppose that the first $m$ variables are eliminated from the model on the basis of having small $c_j$ (recall that $m$ is selected by the

TABLE 3: Selection of subsets $J_s$ obtained with different $m$ values in the Linear Lasso, along with the corresponding fractions of $y$-content distribution ($\{c_s^t C_s^{-1} c_s\}^{1/2}$, in %); this is based on all observations. Mean-squared prediction errors (CV-MSE) and their standard deviations (SD), obtained with 50 repetitions of 10-fold cross-validation, are also provided.

**Linear Lasso**

| | $s$ | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|---|
| | $J_s$ | {1,2,3,4,5} | {1,2,3,5} | {1,2,5} | {1,2} | {1} |
| $m = 0$ | % $y$-cont. | 57.758 | 57.749 | 57.277 | 56.984 | 53.320 |
| | CV-MSE | 0.8524 | 0.8476 | 0.8434 | 0.7864 | **0.7784** |
| | SD | 0.0397 | 0.0393 | 0.0372 | 0.0273 | 0.0452 |
| | $J_s$ | {1,2,3,4,5} | {1,2,3,4} | {1,2,3} | {1,2} | {1} |
| $m = 1$ | % $y$-cont. | 57.758 | 57.548 | 57.231 | 56.984 | 53.320 |
| | CV-MSE | 0.8556 | 0.8305 | 0.8398 | 0.7919 | **0.7776** |
| | SD | 0.0381 | 0.0331 | 0.0342 | 0.0281 | 0.0420 |
| | $J_s$ | {1,2,3,4,5} | {1,2,3,4} | {1,3,4} | {1,3} | {1} |
| $m = 3$ | % $y$-cont. | 57.758 | 57.548 | 56.457 | 55.802 | 53.320 |
| | CV-MSE | 0.8582 | 0.8288 | 0.7883 | **0.7679** | 0.7757 |
| | SD | 0.0498 | 0.0380 | 0.0362 | 0.0424 | 0.0389 |
| | $J_s$ | {1,2,3,4,5} | {1,2,3,4} | {1,3,4} | {1,3} | {1} |
| $m = 5$ | % $y$-cont. | 57.758 | 57.548 | 56.457 | 55.802 | 53.320 |
| | CV-MSE | 0.8554 | 0.8312 | 0.7870 | **0.7756** | 0.7837 |
| | SD | 0.0420 | 0.0373 | 0.0291 | 0.0571 | 0.0613 |

user); the remaining $r - m$ variables then have the highest correlations and are discarded according to a backward elimination based on the variance criterion. This elimination procedure can be viewed as a one-sided version of the two-sided penalty function in Tibshirani (1996). It uses a type of moment generating penalty (the tilt $\exp\{\gamma y\}$) which, combined to the standard normal latent distribution, just provides a shift of that distribution in the direction of positive $y$. The "other side" of the penalty (usually managed using absolute values, as in Lasso) is handled here by having only non-negative regression coefficients. At each step, this eliminates the smallest contributor to estimable $y$-distribution.

The resulting sequence of models is detailed in Table 3 for different choices of $m$, along with the corresponding percentage of $y$-content distribution for each model. For comparison, Table 4 provides the models obtained using Lasso with a continuum of $\gamma$ values. Note that these models and percentages (first two lines of Tables 3 and 4) are obtained using all available observations in the dataset. Lasso and Linear Lasso do not propose the same sequences of models; in fact, Linear Lasso with $m = 3$ and $m = 5$ are the only instances with identical sequences. In all cases, the first variable to be eliminated is always either the fifth or fourth one.

TABLE 4: Selection of subsets $J_s$ obtained with different $\gamma$ values in Lasso, along with the corresponding fractions of $y$-content distribution ($\{c_s^t C_s^{-1} c_s\}^{1/2}$, in %); this is based on all observations. Mean-squared prediction errors (CV-MSE) and their standard deviations (SD), obtained with 50 repetitions of 10-fold cross-validation, are also provided.

**Lasso**

| $\gamma$ | 0.00 | 0.03 | 0.06 | 0.10 | 0.14 |
|---|---|---|---|---|---|
| $s$ | 5 | 4 | 3 | 3 | 3 |
| $J_s$ | {1,2,3,4,5} | {1,2,3,5} | {1,2,3} | {1,2,3} | {1,2,3} |
| % $y$-cont. | 57.758 | 57.749 | 57.231 | 57.231 | 57.231 |
| CV-MSE | 0.8673 | 0.8213 | 0.8037 | 0.8158 | 0.8377 |
| SD | 0.04488 | 0.0401 | 0.0362 | 0.0383 | 0.0440 |
| $\gamma$ | 0.18 | 0.22 | 0.25 | 0.30 | |
| $s$ | 2 | 2 | 1 | 1 | |
| $J_s$ | {1,3} | {1,3} | {1} | {1} | |
| % $y$-cont. | 55.802 | 55.802 | 53.320 | 53.320 | |
| CV-MSE | 0.8591 | 0.8935 | 0.9252 | 0.9492 | |
| SD | 0.0445 | 0.0439 | 0.0265 | 0.0254 | |

To find the optimal number of explanatory variables ($s$) and the optimal value of $m$ in terms of prediction, we use a repeated 10-fold cross-validation approach. The 50 observations are randomly divided into 10 groups of size five. One of these groups is taken as the test set, while the nine remaining groups form the training set. The Linear Lasso is then applied to the training set to obtain a sequence of nested models, as well as coefficient estimates for these models. Each of the five fitted models (size five to size one) is then used to predict responses in the test set; for each model, we record the mean-squared prediction error. These steps are repeated 10 times, each time selecting a different group as the test set. The mean-squared prediction errors are averaged separately for each of the five models. This process is then repeated 50 times, every time with a new random partitioning of the observations into 10 groups. For each model size ($s$) and each choice of $m$, the output is thus a 50-dimensional vector of mean-squared prediction errors. The last two lines of Table 3 report the means and standard deviations of these vectors (one vector for each pair $m, s$). The same steps are repeated with the standard Lasso, using several values of the tuning parameter $\gamma$; results are reported in the last two lines of Table 4. We note that in the cross-validation process, models are fitted using portions of the initial dataset; such models may vary from one training set to another, and in particular may differ from the models reported in Tables 3 and 4 (obtained using all observations). This nonetheless allows identifying the optimal pair $m, s$ for prediction.

According to Table 3, the Linear Lasso favours the model with two explanatory variables ($x_1^*$ and $x_3^*$) as this is the selection that minimizes the mean-squared prediction error. The model with a single explanatory variable ($x_1^*$) however offers a comparable performance. The standard Lasso rather selects the model with $x_1^*$, $x_2^*$, and $x_3^*$ ($\gamma = 0.06$ minimizes the mean-squared prediction error). The prediction errors are, on average, smaller when using the Linear Lasso model with

TABLE 5: Least squares estimates and their standard errors for each subset $J_s$ selected by the Linear Lasso with $m = 3$ and $m = 5$.

| | $s = 5$ | | $s = 4$ | | $s = 3$ | | $s = 2$ | | $s = 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}$ | SE | $\hat{\beta}$ | SE | $\hat{\beta}$ | SE | $\hat{\beta}$ | SE | $\hat{\beta}$ | SE |
| $\mathbf{x}_1^*$ | 0.516 | 0.143 | 0.533 | 0.136 | 0.489 | 0.128 | 0.479 | 0.126 | 0.533 | 0.121 |
| $\mathbf{x}_2^*$ | 0.206 | 0.219 | 0.145 | 0.157 | – | – | – | – | – | – |
| $\mathbf{x}_3^*$ | 0.112 | 0.204 | 0.129 | 0.198 | 0.240 | 0.157 | 0.173 | 0.126 | – | – |
| $\mathbf{x}_4^*$ | -0.019 | 0.220 | -0.080 | 0.159 | -0.111 | 0.156 | – | – | – | – |
| $\mathbf{x}_5^*$ | -0.097 | 0.239 | – | – | – | – | – | – | – | – |

one or two variables than when using the three-variable Lasso model. If one had reason to prefer a three-variable model, the best options are the Linear Lasso with $m = 3$ or $m = 5$.

Table 5 presents the least squares estimates $c_s^t C_s^{-1}$ for each model proposed by the Linear Lasso with $m = 3$. The standard errors of the estimates are obtained as the square root of $\sigma_s^2 C_s^{-1}$, where the estimate of $\sigma_s^2$ is the residual sum-of-squares, divided by $n - s$. Overall, both methodologies seem to agree that the first explanatory variable (police funding) has a large effect, while the other variables (all related to the population's education level) have small or moderate effects. This indicates that more police resources are allocated in cities with higher crime rates.

## 11. EXAMPLE: MATHEMATICS GRADES

As a second example, we study student performance in secondary institutions using the dataset in Cortez & Silva (2008). This dataset records the final mathematics grades of $n = 395$ students along with 32 potential explanatory variables; these variables include the students' past grades, as well as other factors including demographic, social, and education-related features (age, family status, absences, etc). Nominal variables, such as the field of the mother's job, were converted into binary variables; the total number of variables is thus $r = 41$.

We study three different scenarios: in Scenario A, the first- and second-period grades are available ($r = 41$); in Scenario B, the first-period grades are available, but the second-period grades are not ($r = 40$); in Scenario C, the first- and second-period grades are not available ($r = 39$).

To find the optimal model in terms of prediction, we run a repeated 5-fold cross-validation algorithm similar to that described in the previous section. The 395 observations are thus randomly divided into five groups of size 79; once each of these five groups has acted as the test set (the other four groups being combined into a training set), new groups are formed and the approach is repeated for a total of 50 times. This repeated cross-validation approach is applied on the Linear Lasso with $m = r$, $m = 37$, and $m = 0$, each generating a nested sequence of models ranging from size $s = 41$ to size $s = 1$; the choice $m = 37$ corresponds to the number of $c_j$ elements $\leq 0.2$. The approach is then repeated on the standard Lasso, using a $\gamma$-vector of length 400 in order to find the best possible model.

The cross-validation method described above generates 50 mean-squared prediction errors for each pair $(s, m)$ tested; we then compute the mean and standard deviation of each such 50-dimensional vector. Table 6 provides some information about the models that minimize the mean-squared prediction error for each method implemented (Linear Lasso and standard Lasso) and each scenario studied (A, B, and C).

TABLE 6: Cross-validation mean-squared prediction errors (CV-MSE) and their standard deviations (SD) for the best models of the Linear and standard Lasso in each of Scenarios A, B, and C. The number of parameters in each model is also specified; for the standard Lasso, $s$ is the number of regression coefficients greater or equal to 0.01.

| | **A** | | | **B** | | | **C** | | |
|---|---|---|---|---|---|---|---|---|---|
| | CV-MSE | SD | $s$ | CV-MSE | SD | $s$ | CV-MSE | SD | $s$ |
| Lin. Lasso ($m = r$) | 0.1792 | 0.0008 | 2 | 0.3563 | 0.0026 | 3 | 0.8708 | 0.0092 | 2 |
| Lin. Lasso ($m = 37$) | 0.1804 | 0.0015 | 3 | 0.3573 | 0.0028 | 3 | 0.8714 | 0.0103 | 2 |
| Lin. Lasso ($m = 0$) | 0.1794 | 0.0043 | 5 | 0.3580 | 0.0101 | 4 | 0.8760 | 0.0043 | 1 |
| Stand. Lasso | 0.1785 | 0.0019 | 6 | 0.3514 | 0.0052 | 9 | 0.8777 | 0.0134 | 21 |

When past grades are available (first and/or second period), the models obtained show good prediction potential. When past grades are excluded from the model, it becomes quite difficult to predict final grades, which is in line with the conclusions of Cortez & Silva (2008). In that case, there are nonetheless a few variables that are kept in the model, such as the number of past failures.

Prediction errors are similar under the Linear and standard Lasso approaches. The standard Lasso however systematically keeps a large number of explanatory variables in the model, paradoxically offering a fit that is no better than that of the Linear Lasso in terms of prediction. Linear Lasso, in contrast, offers parsimonious fits; this agrees with Cortez & Silva (2008)'s claim about the high number of irrelevant variables in the dataset.

Linear Lasso keeps first- and second-period grades as explanatory variables in Scenario A, first-period grades and number of failures in Scenario B, and number of failures and mother's education in Scenario C. In Scenario A, Lasso keeps past grades, age, number of failures, quality of family relationships, and number of absences. In Scenario B, it keeps first-period grades, age, number of failures, and number of absences, but it replaces the quality of family relationships (which was included in the Scenario A model) by the existence of a romantic relationship, reason for choosing the school, and a few other variables. In Scenario C, Lasso keeps 21 variables, that is, too many variables to enumerate all of them; we however note that the number of failures is still there and has the largest coefficient, followed by gender.

## 12. DISCUSSION

The Linear Lasso uses sign-adjusted explanatory vectors so as to work with predictors that are positively correlated with the interest variable; this is entirely notational, but means that the recorded explanatory vectors all point into the positive half-space $\mathcal{L}^+\mathbf{y}$, "above" the plane $\mathcal{L}^\perp\mathbf{y}$. This allows certain characteristics to be more easily described in geometric terms; this also argues that the Lasso objective itself should be recast as the scalar change $y$ along the line $\mathcal{L}\mathbf{y}$ rather than the vector change $y\mathcal{L}\mathbf{y}$ in the vector space.

The modified objective means that the maximum likelihood value for the response is now on the line $\mathcal{L}\mathbf{y}$, and all explanatory vectors intersect that line at the origin. A penalty function then becomes the one-sided moment generating function $\gamma\Sigma\beta_j$ with the "other side" being handled

by the usual positive regression coefficient requirement. As a result, computation is strictly on the line $\mathcal{L}\mathbf{y}$; then, as $\gamma$ is increased, the $\mathbf{x}_j$ vectors are shifted in the $-\mathcal{L}\mathbf{y}$ direction, and dropped from the lower end as determined.

When a particular $\mathbf{x}_j$ is dropped in computation, there is a minimum reduction in the variance of the accessible $y$-information. However, when a group of $\mathbf{x}_j$ is dropped, there is no assurance that the composite change results in a minimum variance reduction. This is the same for the usual Lasso as it is here for the Linear Lasso, and would be as expected from the exponential ordering in the possible selection of subsets.

The Linear Lasso handles cases with $r >> n$ in a straight-forward manner, by simultaneously dropping several variables featuring the smallest $c_j$'s. It also largely works with singular matrices $C$, due to the simplicity of the minimizing procedure. Empirical evidence from the real dataset examples of Sections 10 and 11 show that the performance of Linear Lasso is in accordance with the theoretical results developed in earlier sections; in these examples, the Linear Lasso finds models that are comparable to those found by the usual Lasso in terms of prediction accuracy, yet it consistently proposes more parsimonious models. The main advantage of Linear Lasso stems from its simplicity and ease of application, translating into a computational problem that is basically independent of the dimension once correlations are obtained.

The theory and objective function of the Linear Lasso have been proposed in a context of multiple linear regression. Resolution algorithms for the usual Lasso offer a numerical solution for a wide range of regression models. It will be interesting to find out how the geometric arguments of the Linear Lasso can be adapted to suit other contexts that are of interest for the regular Lasso.

## APPENDIX A

### A. NORMAL LOCATION – A TILT IS A SHIFT

For a standard normal $c\exp\{-z^2/2\}$ on the real line let $\exp\{\gamma z\}$ be a factor that gives an exponential tilt or boost to the right:

$$c\exp\{-z^2/2\}\exp\{\gamma z\} = c\exp\{-(z-\gamma)^2/2\}.$$

We thus see that a $\gamma$ tilt to the right can be viewed as a $\gamma$ shift of the distribution to the right. Now consider a standard normal $c\exp\{-\sum z_i^2/2\}$ on a vector space space coupled with a $\gamma$ tilt in some direction $\mathbf{x}$:

$$c\exp\{-\Sigma z_i^2/2\}\exp\{\gamma\Sigma z_i x_i\} = \exp\{-\Sigma(z_i-\gamma x_i)^2/2\}.$$

Then similarly we see that a $\gamma$ tilt in the direction $\mathbf{x}$ can be viewed as a $\gamma$ shift in the direction $\mathbf{x}$.

If only $s$ of some explanatory variables are being considered, say, those with subscripts in the set $J_s = \{j_1, \cdots, j_s\}$, we can use the corresponding correlation arrays as, say, $c_s$ and $C_s$, and then have distributional results analogous to the two preceding equations but in the appropriate subspace.

## ACKNOWLEDGEMENTS

We are grateful to the Associate Editor and referees, whose comments have contributed to significantly improving the paper. This work has been supported by the Natural Sciences and Engineering Research Council of Canada.

## BIBLIOGRAPHY

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2), 802–837.

Cortez, P. & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Future Business Technology Conference*, pp. 5–12.

Fan, J. & Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Royal Statistical Society*, B 70, 849–911.

Hastie, T., Tibshirani, R., & Wainwright, N. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3), 907–927.

Lee, J. D. & Taylor, J. E. (2014). Exact post model selection inference for marginal screening. *arXiv preprint arXiv:1402.5596*.

Taylor, J. & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25), 7629–7634.

Thomas, G. S. (1990). *The Rating Guide to Life in America's Small Cities*. ERIC.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Zhao, S., Witten, D., & Shojaie, A. (2021). In defense of the indefensible: A very naive approach to high-dimensional inference. *Statistical Science*, 36(4), 562–577.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

---

*Received 15 April 2019*
*Accepted 8 July 2010*

*The Canadian Journal of Statistics*
Vol. xx, No. yy, 20??, Pages 1–25
*La revue canadienne de statistique*



# Appendix B: More on the Linear Lasso

D.A.S. FRASER and Mylène BÉDARD

*Abstract:* This paper is the last manuscript of Professor D.A.S. Fraser. The main document preserves his unique and original writing style. It presents an approach that simultaneously performs model selection and estimation in the context of linear regression. This goal is achieved by analyzing the standard Lasso from a geometrical viewpoint and then summarizing all available information about the response variable on a single line in the space; this leads to what we call the Linear Lasso approach.

This appendix aims at clarifying and detailing the ideas covered in the paper. As Professor Fraser's co-author and from our numerous discussions on the subject, my goal is to cover our motivations and explain the concepts in my own words, so as to provide a different perspective and hopefully shed some light on sections that might be more obscure. These pages may thus be seen as an accompanying document that provides section-by-section clarifications. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

*Résumé:* Cet article est le dernier manuscrit du regretté professeur D.A.S. Fraser. Le document principal préserve son style d'écriture unique et original; on y retrouve une approche qui effectue simultanément la sélection et l'estimation de modèle dans le contexte de la régression linéaire. Cet objectif est atteint en analysant le Lasso standard d'un point de vue géométrique, puis en projetant toute l'information disponible concernant la variable réponse sur une unique droite dans l'espace. Cette approche mène à ce que nous appelons le Lasso linéaire.

Cette annexe vise à clarifier et à détailler les idées abordées dans le document principal. À titre de co-auteure du professeur Fraser et en m'inspirant de nos nombreuses discussions sur le sujet, mon objectif est de couvrir nos motivations et d'expliquer les concepts dans mes propres mots, de sorte à fournir une perspective différente et, ultimement, à faire de la lumière sur certaines sections qui pourraient paraître plus obscures. Ces pages devraient donc être vues comme un document d'accompagnement procurant des précisions, section par section. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

## 1. INTRODUCTION

In a linear regression context, Lasso simultaneously performs variable selection and coefficient estimation. It consists in a constrained optimization problem where we minimize the residual sum-of-squares arising from a linear model (the objective function), subject to a constraint on the sum of absolute regression coefficients. This is intuitive and elegant, but requires the assistance of algorithms to reach a solution. The larger is the dimensionality of the problem studied, the more demanding is the associated Lasso algorithm in terms of computations.

In the steepest descent algorithm for instance, the process goes down in the steepest possible direction given the current value of the process, until a minimum of the objective function (or a barrier) is reached. The algorithm then re-evaluates its direction in order to pursue its descent (it again chooses the direction of steepest descent given the current point of the function to minimize). When the process hits a barrier (corner/edge of the constraint region), one of the regression coefficients is set to 0; the algorithm therefore consecutively discards coefficients, until a minimum of the function is reached. The number of discarded coefficients eventually depends on a tuning parameter $\gamma$ that acts as a weight on the Lasso constraint. The larger is $\gamma$, the heavier is the penalty compared to the objective function and the more intent we are on discarding explanatory variables (equivalently, the smaller is the constraint region in Figure 2.2 of Hastie, Tibshirani, & Wainwright, 2015). Conversely, if $\gamma = 0$, then there is no constraint and we are back to full model estimation; the regression coefficients are then the usual least squares

estimates. Gradually increasing $\gamma$ in the Lasso thus leads to an ordering of models from full to empty.

The efficiency of the Lasso method of course depends on the dimensionality of the context studied. The algorithm selected to solve the constrained optimization problem has to explore a multidimensional space, which turns out to be an intense exercise in certain situations. Considering this drawback, we wish to make use of geometrical arguments to make the variable selection problem dimension-free. Instead of considering data $(y_i, x_{i,1}, \ldots, x_{i,r})$ on $n$ separate subjects $(i = 1, \ldots, n)$, and then using these $n$ subjects to produce a statistical model (fixed $\gamma$ value) or an ordering of models (continuum of $\gamma$ values), we turn the problem around. We visualize the $n$ observed responses $(y_1, \ldots, y_n)$ as a single $n$-dimensional response vector $\mathbf{y}$, and similarly group the $n$ measures from a given explanatory variable $(x_{1,j}, \ldots, x_{n,j})$ into an $n$-dimensional explanatory vector $\mathbf{x}_j$. By extracting the information (about the response) contained in each of the $r$ explanatory vectors, and then projecting these pieces of information onto the response vector, we transform the initial $r$-dimensional convex optimization problem into a single-dimensional one.

These geometrical inputs are then combined with location model theory to produce the Linear Lasso method. In particular, we use a latent standard normal model that is conveniently equivalent (algebraically) to the use of least squares, and then obtain a predictive distribution lying on the response line. This then allows characterizing a (one-dimensional) distribution for the $y$-content hidden in the explanatory variables. By varying the location parameter of this $y$-content distribution, it gradually shifts and allows discarding some predictors. Since every pertinent piece of information about $y$ (in a linear regression context, that is) is projected on the response vector/line, we refer to the new method as Linear Lasso; this emphasizes that our constrained optimization problem has a linear trajectory (in opposition to the bits and pieces that are produced by the traditional Lasso).

## 2. BACKGROUND AND NOTATION

In this section, a number of simplifying assumptions are introduced to make the problem easier to visualize; these do not affect the applicability of the results.

The regular Lasso looks for a small selection of explanatory variables that provides good prediction for the response variable. Typically, the number of available explanatory variables $r$ is large and the number of variables $s$ kept by the Lasso is much smaller. Indeed, Lasso is on a budget and has to limit the total magnitude of the $r$ coefficients in $\beta$; it thus sets a number of these coefficients to 0, keeping only the most useful ones in terms of response prediction. Naturally, the tighter is the budget, the greater is the compromise in terms of data fitting.

Typically, explanatory variables are standardized; this means that each column vector in $\mathbf{X}$ has mean 0 and unit variance. This keeps predictors from depending on the unit with respect to which they were measured (feet versus meters, for instance). We apply the same location-scale standardization to the response vector $\mathbf{y}$. Centering variables is convenient as it allows avoiding the use of a model intercept $\beta_0$.

Using the rescaled residual sum-of-squares as the objective function, we are left minimizing $\|\mathbf{y} - \mathbf{X}\beta\|_2^2/2n$ with respect to $\beta$ and subject to $\|\beta\|_1 \leq t$, with $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively denoting the $\ell_1$ and Euclidean norms. This optimization problem is then equivalent to its Lagrangian form in (1). The tuning parameter $\gamma$ in that equation is usually specified through cross-validation procedures, to be discussed in later sections.

Following the above-mentioned standardization steps, each explanatory vector $\mathbf{x}_j$ $(j = 1, \ldots, r)$ has unit variance $\sum_{i=1}^n x_{ij}^2/n = 1$, implying that each vector is of length $\|\mathbf{x}_j\|_2 = \{\sum_{i=1}^n x_{ij}^2\}^{1/2} = n^{1/2}$. Generally speaking, pairwise correlations between vectors are conve-

niently obtained by computing inner products between unit versions of these vectors. For example, the correlation between $\mathbf{y}$ and $\mathbf{x}_1$ simply is $(\mathbf{y}/n^{1/2}) \cdot (\mathbf{x}_1/n^{1/2}) = \mathbf{y} \cdot \mathbf{x}_1/n$. In our specific context, pairwise correlations are thus inner products between two standardized vectors, divided by the sample size $n$. These correlations, obtained directly from the data, act as the input for our regression problem.

Standardizing the location and scale of data vectors produces regression models with desirable properties. In terms of geometrical representation however, it is more convenient to work with unit versions of these data vectors, i.e., $\mathbf{u}_y = \mathbf{y}/\sqrt{n}$ and $\mathbf{u}_j = \mathbf{x}_j/\sqrt{n}$ for $j = 1, \ldots, r$. We then imagine $\mathbf{u}_y$ as pointing upwards, with its zero point on the origin, and assume that the zero point of every other unit vector $\mathbf{u}_j$ is also on the origin. Since correlations are cosines of angles between corresponding unit data vectors, we can thus view each term $c_j$ $(j = 1, \ldots, r)$ as the projection of $\mathbf{u}_j$ on $\mathbf{u}_y$ (or on the response line $\mathcal{L}\mathbf{y}$, which coincides with $\mathbf{u}_y$).

Now, according to this geometrical representation, the vectors $\mathbf{u}_j$ that are positively correlated with $\mathbf{u}_y$ are directed above the plane $\mathcal{L}^{\perp}\mathbf{y}$ that is perpendicular to $\mathcal{L}\mathbf{y}$ at 0. Similarly, vectors that are negatively correlated with $\mathbf{u}_y$ are directed below that plane. To make the problem one-sided, we choose to reverse the sign of those predictors that are negatively correlated with $\mathbf{y}$; for $j \in \{1, \ldots, r\}$ such that $-1 \leq \mathbf{y} \cdot \mathbf{x}_j/n < 0$, we use $-\mathbf{x}_j$ instead of $\mathbf{x}_j$ ($-\mathbf{u}_j$ instead of $\mathbf{u}_j$) and exclusively work with predictors whose unit vectors $\mathbf{u}_j$ lie in the upper half-space $\mathcal{L}^{+}\mathbf{y}$.

Visually, the unit response vector $\mathbf{u}_y$ is thus the focal point (vector pointing upwards) and explanatory vectors gravitate around it. The smaller is the angle between $\mathbf{u}_y$ and $\mathbf{u}_j$, the more correlated are $\mathbf{y}$ and $\mathbf{x}_j$, and the more information about the response is carried by that predictor. Our analysis will also need to take account of the angles among explanatory vectors. Indeed, two explanatory vectors that are strongly correlated with $\mathbf{y}$ likely contain more information about the response if their own pairwise correlation is low; if they are independent for instance, there is no redundancy in the response information they carry. Our geometrical analysis will thus use, as basic input, the vector $c$ of correlations between the response vector $\mathbf{y}$ and each explanatory vector, as well as the matrix $C$ of correlations among explanatory vectors.

This geometrical setting will eventually allow us to visualize a plane perpendicular to $\mathbf{u}_y$ that elevates itself, gradually setting regression coefficients to 0 when the corresponding unit explanatory vector $\mathbf{u}_j$ finds itself completely under the plane, or equivalently when the plane finds itself over $c_j$, the projection of the vector $\mathbf{u}_j$ on $\mathbf{u}_y$. The height of the perpendicular plane will then play the role of $\gamma$ in the regular Lasso and act as a tuning parameter for the budget (but, as we will see, this parameter needs not be tuned explicitly). From now on, we then assume that the above standardization steps have been applied when referring to the data vectors $(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_r)$ and their unit versions $(\mathbf{u}_y, \mathbf{u}_1, \ldots, \mathbf{u}_r)$.

## 3. LATENT STOCHASTIC MODEL

We started with $n$ observations for each of the $1 + r$ variables (one response variable and $r$ explanatory variables). Now that we have standardized the data vectors $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_r$ and geometrically represented their directions in an underlying vector space using $1 + r$ ($n$-dimensional) unit vectors $\mathbf{u}_y, \mathbf{u}_1, \ldots, \mathbf{u}_r$, we need to provide some distributional structure on that space. To this end, we assume that the directions of the data vectors are fixed (we thus implicitly condition on these directions through the use of data vectors). Using correlations as input for our approach obviously indicates a common stochastic background for the $1 + r$ variables $y, x_1, \ldots, x_r$. We use the normal distribution to represent this stochastic background; this choice is based on large sample theory (first-order asymptotics) and, as we will see, it is algebraically equivalent to least squares analysis; this kind of validates our choice of latent normal model.

Consider, as an underlying stochastic model, $n$ latent variables $Z_1, \ldots, Z_n$ jointly distributed according to a standard multivariate normal $\mathcal{MN}(\mathbf{0}, I_{n\times n})$ on $\mathbb{R}^n$, with $\mathbf{0}$ a vector (of length $n$ here) and $I_{n\times n}$ the identity matrix. We express each of our $1+r$ scalar variables $y, x_1, \ldots, x_r$ as a linear function of this latent model (in other words, our first-order—or linear—approximation for a variable is a linear combination of $Z_1, \ldots, Z_n$). Specifically, $x_1 = (\mathbf{x}_1^t/\sqrt{n}) \cdot (Z_1, \ldots, Z_n)^t$ is a scalar variable with distribution

$$\begin{aligned} x_1 \sim \mathbf{x}_1^t/\sqrt{n}\, \mathcal{MN}(\mathbf{0}, I_{n\times n}) &= \mathcal{N}(\mathbf{u}_1^t \cdot \mathbf{0},\ \mathbf{u}_1^t I_{n\times n}\mathbf{u}_1) \\ &= \mathcal{N}(0,1)\ . \end{aligned}$$

Now, the variable $x_1$ takes values on a line that is perpendicular to the contours of the linear function $(\mathbf{x}_1^t/\sqrt{n}) \cdot (Z_1, \ldots, Z_n)^t$. One contour of a function consists of all points $(z_1, \ldots, z_n)$ that lead to a common value of that function. Geometrically, the contours of any linear function are perpendicular to the gradient of this function; the contours are thus (hyper)planes that are perpendicular to the vector $\mathbf{x}_1/\sqrt{n} = \mathbf{u}_1$. The scalar variable $x_1$ thus takes values on $\mathcal{L}\mathbf{x}_1$, the line that coincides with the vector $\mathbf{u}_1$ (and hence goes through the origin).

Similar conclusions are reached for each of the $r$ remaining variables. This means that the $1+r$ observed lines $\mathcal{L}\mathbf{y}, \mathcal{L}\mathbf{x}_1, \ldots, \mathcal{L}\mathbf{x}_r$ fix the directions of the variables $y, x_1, \ldots, x_r$ in the space. From the latent standard multivariate normal model, each of these variables is (marginally) normally distributed with mean 0 and variance 1. Hence, $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})$, $j = 1, \ldots, r$, each represents a set of $n$ observed values on one of the $1+r$ distinct lines; we note that the mean and variance of each set of observations indeed correspond to those of the marginal distributions.

Now, since each scalar variable is normally distributed, then the collection of $1+r$ variables (or of any subset of these) is also jointly normally distributed. Furthermore, since each scalar variable is a linear function of the latent standard normal model on the underlying $n$-dimensional space, then the individual variables in this joint distribution lie on the observed lines $\mathcal{L}\mathbf{y}, \mathcal{L}\mathbf{x}_1, \ldots, \mathcal{L}\mathbf{x}_r$. For instance, let us consider the bivariate distribution of the variables $x_1 = (\mathbf{x}_1^t/\sqrt{n}) \cdot (Z_1, \ldots, Z_n)^t$ and $x_2 = (\mathbf{x}_2^t/\sqrt{n}) \cdot (Z_1, \ldots, Z_n)^t$:

$$\begin{aligned} \begin{pmatrix} n^{-1/2}\mathbf{x}_1^t \\ n^{-1/2}\mathbf{x}_2^t \end{pmatrix} \mathcal{MN}\left(\mathbf{0}, I_{n\times n}\right) &= \mathcal{MN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, n^{-1}\begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \end{pmatrix} I_{n\times n}(\mathbf{x}_1, \mathbf{x}_2)\right) \\ &= \mathcal{MN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{u}_1^t\mathbf{u}_1 & \mathbf{u}_1^t\mathbf{u}_2 \\ \mathbf{u}_2^t\mathbf{u}_1 & \mathbf{u}_2^t\mathbf{u}_2 \end{pmatrix}\right) \\ &= \mathcal{MN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & c_{12} \\ c_{21} & 1 \end{pmatrix}\right)\ . \end{aligned}$$

The component $x_1$ in this bivariate normal takes values on the line $\mathcal{L}\mathbf{x}_1$, while the component $x_2$ takes values on $\mathcal{L}\mathbf{x}_2$. This holds generally, for any subset of the $1+r$ variables $y, x_1, \ldots, x_r$.

As it turns out, the implications of the above modeling are quite important. The marginal distribution of the response variable $y$, obtained as a linear function of the latent $Z_1, \ldots Z_n$, is a $\mathcal{N}(0,1)$ along the line $\mathcal{L}\mathbf{y}$. This means that values of $y$ can be recorded directly on $\mathcal{L}\mathbf{y}$, i.e., the dataset provides the direction $\mathcal{L}\mathbf{y}$ used for prediction. Consequently, instead of working in an $n$-dimensional space for achieving predictions, we only work along $\mathcal{L}\mathbf{y}$; prediction is thus one-dimensional rather than $n$-dimensional. This is where the expression "Linear Lasso" originates: instead of exploring a space, we need only move along a line. Eventually, we will include parameters regulating the presence/absence of each predictor in the model; more will be said later

about this.

By contrast, the standard least squares modeling uses a proportion $\hat{\beta}_j$ of each vector $\mathbf{x}_j$ to form a vector $\hat{\mathbf{y}}$ that is as close as possible to $\mathbf{y}$ in terms of the residual sum-of-squares. We can view this as opening the "taps" of the different lines on the latent stochastics more or less freely, so as to approach $\mathcal{L}\mathbf{y}$ as much as possible. If the model is saturated ($n = r$ and $\mathbf{u}_y$ is in the linear span of $\mathbf{u}_1, \ldots, \mathbf{u}_r$), then the least squares line $\mathcal{L}\hat{\mathbf{y}}$ coincides with $\mathcal{L}\mathbf{y}$. In the subsaturated case ($r < n$), the line $\mathcal{L}\hat{\mathbf{y}}$ will be in the linear span of the predictor lines, but distinct from $\mathcal{L}\mathbf{y}$. In the supersaturated case, exact solutions are not unique. According to this least squares approach, prediction is obviously $n$-dimensional as we search for the line that is closest to $\mathcal{L}\mathbf{y}$ in the $n$-dimensional space.

To summarize, we started with a rotationally symmetric normal latent model; after observing data (the training sample in a cross-validation context, for instance), we obtain one vector (or line) in $\mathbb{R}^n$ for each variable. These $1 + r$ lines, which all go through the origin, are assumed to be fixed and turn out to be correlated; they can thus be regarded as a kind of $(1 + r)$-dimensional skewed coordinate system in a variable space with an underlying standard normal model. The resulting distribution on the $1 + r$ observed lines $\mathcal{L}\mathbf{y}, \mathcal{L}\mathbf{x}_1, \ldots, \mathcal{L}\mathbf{x}_r$ is thus the correlated normal mentioned in (3). Given observations for the $i$-th subject $\mathbf{X}_i = (y_i, x_{i,1}, \ldots, x_{i,r})$ (the test sample in a cross-validation context, say), we then ignore the scalar $y$ and hope that we can provide a good predictive distribution for $y$ on $\mathcal{L}\mathbf{y}$, using $x_1, \ldots, x_r$ on the lines $\mathcal{L}\mathbf{x}_1, \ldots, \mathcal{L}\mathbf{x}_r$ (by conditioning on these variables, for instance).

The latent stochastics we use allow one to picture each of the $1 + r$ variables as a linear function of the latent $Z_1, \ldots, Z_n$. This clearly illustrates the continuity that is present among variables, as they all depend on the same latent variables. In what follows, the notation $\{y\mathcal{L}\mathbf{y}, x_1\mathcal{L}\mathbf{x}_1, \cdots, x_r\mathcal{L}\mathbf{x}_r\}$ refers to $1 + r$ correlated scalar variables $y, x_1, \ldots, x_r$ that have a joint normal distribution as in (3), and that take values on the fixed lines $\mathcal{L}\mathbf{y}, \mathcal{L}\mathbf{x}_1, \ldots, \mathcal{L}\mathbf{x}_r$; these are the data generating equations.

## 4. INFERENCE FROM A PARTICULAR SUBSET OF EXPLANATORY VARIABLES

Using the distributional structure established in Section 3 of the paper, we want to perform inference on the response variable $y$ using a subset $J_s$ of $s$ explanatory variables $x_{j_1}, \ldots, x_{j_s}$. For the moment, we do not worry about how this specific subset is chosen.

From the properties of the multivariate normal distribution, any subset of variables in $y, x_1, \ldots, x_r$ is also normally distributed; the resulting distribution is still centered at $\mathbf{0}$ and its covariance matrix is obtained by removing, from $\tilde{C}$ in (2), the rows and columns corresponding to the discarded variables. Using the resulting $(1 + s)$-dimensional joint distribution for $(y, x_{j_1}, \ldots, x_{j_s})$, it is then easy to obtain the conditional distribution of the response $y$ given the $s$ explanatory variables; this distribution on $\mathcal{L}\mathbf{y}$ is provided in (4).

The forecasted value, or $y$-content, is taken as the mean of the conditional distribution $c_s^t C_s^{-1}(x_{j_1}, \ldots, x_{j_s})^t = (x_{j_1}, \ldots, x_{j_s}) \cdot C_s^{-1} c_s$. We thus find ourselves in the familiar situation where prediction is computed using a model that is algebraically equivalent to least squares; indeed, we have $\hat{\beta}_s = C_s^{-1} c_s = (\mathbf{X}_s^t \mathbf{X}_s)^{-1} \mathbf{X}_s^t \mathbf{y}$, where the matrix $\mathbf{X}_s$ is an $n \times s$ matrix whose columns are the $n$-dimensional explanatory vectors $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_s}$. The variance $1 - c_s^t C_s^{-1} c_s$ of the conditional distribution is a variability around the forecasted value that is unexplained by the linear model.

Now, the $y$-content $c_s^t C_s^{-1}(x_{j_1}, \ldots, x_{j_s})^t$ itself possesses a distribution on $\mathcal{L}\mathbf{y}$; indeed, the correlations are taken as fixed, but the explanatory variables $x_{j_1}, \ldots, x_{j_s}$ are jointly normally distributed. The forecast is thus a linear function of a normal with mean $\mathbf{0}$ and covariance matrix

$C_s$, which yields a normal distribution with mean 0 and variance

$$\begin{aligned}\mathbb{V}ar(c_s^t C_s^{-1}(x_{j_1},\cdots,x_{j_s})^t) &= c_s^t C_s^{-1}\mathbb{V}ar((x_{j_1},\cdots,x_{j_s})^t)C_s^{-1}c_s \\ &= c_s^t C_s^{-1}C_s C_s^{-1}c_s \\ &= c_s^t C_s^{-1}c_s\ .\end{aligned}$$

This $\mathcal{N}(0, c_s^t C_s^{-1}c_s)$ represents the distribution of possible forecasted values using the subset $J_s$ of explanatory variables. In this case, the residual/error terms in our linear model have a variance of $1 - c_s^t C_s^{-1}c_s$ (see (4)) and are independent of the forecasted values. We therefore see that these two variances add to unity, leading to the marginal $\mathcal{N}(0,1)$ distribution for $y$.

As mentioned, the term $c_s^t C_s^{-1}c_s$ represents the variability of forecasted values obtained using the predictors in $J_s$. Hereafter, we however prefer to use the standard deviation $\{c_s^t C_s^{-1}c_s\}^{1/2}$ and express the distribution of forecasted values, or $y$-content distribution, as $\{c_s^t C_s^{-1}c_s\}^{1/2}\mathcal{N}(0,1)$. We then interpret the standard deviation $\{c_s^t C_s^{-1}c_s\}^{1/2}$ as the fraction of $y$ variability, or rather the fraction of the marginal response distribution, that is contained in the subset $J_s$ of explanatory variables.

## 5. A VERY SIMPLE EXAMPLE

From this simple example with only two predictors, we can easily visualize what happens. The data vectors are of unit length and conveniently refer to the vectors $\mathbf{u}_y, \mathbf{u}_1, and \mathbf{u}_2$. By placing the zero point of these vectors on the origin as in Figure 1, we easily imagine their projections on $\mathcal{L}\mathbf{y}$. We then realize that the stronger is the correlation between variables $x$ and $y$, the larger is the projection of the vector $\mathbf{u}_x$ on $\mathbf{u}_y$ and the greater is the fraction of $y$-variability captured by $x$.

Whether or not a second variable is useful in predicting the response does not only depend on its correlation with $y$, but also depends on its correlation with the other predictor. In this example, the vector $\mathbf{x}_2$ is significantly correlated with $\mathbf{y}$, but is even more strongly correlated with $\mathbf{x}_1$; the extra $y$-variability it explains in the model and projects on the line $\mathcal{L}\mathbf{y}$ is therefore very small in the end.

This explains why the regression coefficient for $x_2$ is negative in the full model with $s = 2$, despite the fact that both predictors are positively correlated with $y$. Indeed, a large portion of the $y$-content captured by $x_2$ is redundant with that from $x_1$; the exclusive $y$-content coming from $x_2$ is thus modest. Due to the pairwise correlation between predictors, the extra information from $x_2$ can be modeled by increasing the regression coefficient associated to $x_1$, and then fine-tuning the adjustment with a coefficient for $x_2$ that is slightly below 0.

In this example, we did not explicitly standardize the data vectors (mean of 0 and SD of 1) since we performed no prediction and only spoke in general terms. It is however implicitly assumed that the forecasted values ($y$-content) in Tables 1 and 2 are computed using observations (from the explanatory variables $x_1$ and $x_2$) that were previously standardized.

## 6. HOW MUCH $Y$ DISTRIBUTION IS HIDDEN IN SELECTED VARIABLES

In Section 4 (of the paper and of Appendix B), we obtained the conditional distribution of the response variable $y$ given a specific subset $J_s$ of explanatory variables. This distribution is a normal on the line $\mathcal{L}\mathbf{y}$, with a mean $c_s^t C_s^{-1}(x_{j_1},\ldots,x_{j_s})^t$ that simply corresponds to the least squares prediction and that we interpret as the forecasted value (or $y$-content). This $y$-content itself is normally distributed on the line $\mathcal{L}\mathbf{y}$, with mean 0 and variance $c_s^t C_s^{-1}c_s$.

Our ultimate objective being to perform variable selection, we wish to have more flexibility in our model; we thus introduce a parameter $\delta$, as defined in the paper, that provides a selection of predictors. The statistical distribution of the forecasted value is then expressed as $c_\delta^t C_\delta^{-1}\mathbf{x}_\delta \sim \mathcal{N}(0, c_\delta^t C_\delta^{-1} c_\delta)$ on $\mathcal{L}\mathbf{y}$ with $\mathbf{x}_\delta = (x_{1\delta_1}, \ldots, x_{r\delta_r})^t$, where a null subscript indicates that the variable is excluded from the model. The density of $c_\delta^t C_\delta^{-1}\mathbf{x}_\delta$ naturally integrates to 1 as we have access to the complete distribution of forecasted values given the selection of explanatory variables $\delta$.

Factoring the standard deviation in the previous distribution leads to

$$\{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}\mathcal{N}(0,1) = \sigma(\delta)\mathcal{N}(0,1)$$

on $\mathcal{L}\mathbf{y}$; the selection $\delta$ thus captures a fraction $\sigma(\delta)$ of the marginal $y$-distribution. Since we are working with a fraction of a the response distribution, the term $\{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}\mathcal{N}(0,1)$ similarly corresponds to a fraction of the response density, which does not integrate to 1. This implies that the remaining fraction $(1 - \{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2})$ *of the same marginal $y$-distribution* has been lost in the modeling process (that is, the subset of explanatory variables possesses limited information about the response variable). The statistical distribution of the response contained in a selection $\delta$ of predictors is thus expressed as $f(y;\delta) = \sigma(\delta)\phi(y)$, where $\phi(\cdot)$ is the density of a standard normal. Note that this explanation is not to be confused with the claim $c_\delta^t C_\delta^{-1}\mathbf{x}_\delta = \sigma(\delta)y$, which simply does not hold here; if it were true, it would mean that we could recover, from the $y$-content $c_\delta^t C_\delta^{-1}\mathbf{x}_\delta$ and the fraction $\{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}$, the exact value of the response $y$.

We could now view this variable selection situation as an inferential problem involving interest and nuisance parameters. The statistical model $f(y;\delta)$ can be interpreted as a likelihood function in which $y$ is the location parameter of interest and $\delta$ is the nuisance parameter. The interpretation of $y$ as a location parameter is possible thanks to the symmetry of the normal density in the variable and mean (i.e., $f(\mu; y, \sigma) = f(y; \mu, \sigma)$). The likelihood function is then $L(y, \delta) = \sigma(\delta)\phi(y)$, which is just a fraction of the marginal likelihood for $y$. The goal thus obviously becomes to maximize $\sigma(\delta)$ with respect to $\delta$ so as to eliminate the nuisance parameter while simultaneously having access to the greatest possible fraction of the marginal likelihood for $y$. We may of course include, in our inferential problem, a potential constraint on the total number of explanatory variables $\Sigma_j \delta_j$. The variable selection problem can thus been seen as being related to location model theory.

## 7. SUBSTANTIAL REDUCTION IN THE SET $J_S$ OF EXPLANATORY VARIABLES

In this section, we wish to propose a variable selection approach based on the $y$-content distribution in a selection of predictors $\delta$, with density $f(y;\delta) = \sigma(\delta)\phi(y)$. Using this statistical model, our goal is to drop the predictors that do not contribute much in terms of $y$-content. We thus wish to restrict our search for a solution along the line $\mathcal{L}\mathbf{y}$, on which we previously projected all pertinent information about the response.

In the regular Lasso, the goal is to minimize the residual sum-of-squares, subject to a contraint on the total magnitude of the regression coefficients. Lasso does not achieve an exact solution, but uses algorithms to approximate a solution in the multidimensional space. This is exactly what we do in this section, but we restrict our search along $\mathcal{L}\mathbf{y}$. For this, recall that the unit vector $\mathbf{u}_y$ points upwards and that all unit explanatory vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are positively correlated with $\mathbf{u}_y$, lying in the upper half space over $\mathcal{L}^\perp\mathbf{y}$.

### 7.1. Independent predictors

To ease the discussion, we first suppose that all covariates are independent from each other ($C$ is the identity matrix). Because of the predictors' positive correlation with $\mathbf{y}$, the regression coef-

ficients are such that $\beta_j \geq 0$; the problem then becomes one-sided and the function to minimize is $\|\mathbf{y} - \mathbf{X}\beta\|_2^2/2n + \gamma \sum_j \beta_j$ (we drop the absolute values in the constraint). In the independent case, there is no redundancy in the $y$-information carried by the predictors; each explanatory variable thus contributes totally new information about $y$.

When there is no constraint ($\gamma = 0$), the regression coefficient estimates that minimize $\|\mathbf{y} - \mathbf{X}\beta\|_2^2$ satisfy $\beta_j = c_j$. The predictor with the smallest $\beta_j$ is thus the least correlated with $y$ and contributes a smaller fraction to the $y$-distribution than the other predictors. In our standardized context, this means that one unit of predictor $j$ brings less information about $y$ than one unit of any other predictor. As $\gamma$ grows, the budget becomes tighter and the regression coefficients gradually shrink. In particular, the smallest regression coefficient gradually decreases until it reaches 0. Since the information carried by the predictors is mutually exclusive, the information loss from decreasing $\beta_1$, say, cannot be partially recovered by rebalancing the coefficient of another predictor. Therefore, as $\gamma$ increases, each coefficient successively shrinks toward 0, starting with the predictor associated to the smallest $c_j$ and ending with the greatest one.

Now, in the specific context outlined in Section 6, we know that the $y$-content distribution is a $\mathcal{N}(0, c_\delta^t c_\delta)$ on $\mathcal{L}\mathbf{y}$. Furthermore, coefficient estimates are of the form $\hat{\beta}_\delta = c_\delta = (c_{1\delta_1}, \ldots, c_{r\delta_r})$, where a null subscript $j * \delta_j$ indicates that the variable $j$ is excluded from the vector. When computing the residual sum-of-squares using estimates $\hat{\beta}_\delta = c_\delta$, we find that

$$\begin{aligned}
\|\mathbf{y} - \mathbf{X}_\delta \hat{\beta}_\delta\|_2^2 &= \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{X}_\delta\hat{\beta}_\delta + (\mathbf{X}_\delta\hat{\beta}_\delta)^t\mathbf{X}_\delta\hat{\beta}_\delta \\
&= \mathbf{y}^t\mathbf{y} - 2(\mathbf{X}_\delta^t\mathbf{y})^t\hat{\beta}_\delta + \hat{\beta}_\delta^t(\mathbf{X}_\delta^t\mathbf{X}_\delta)\hat{\beta}_\delta \\
&= n - 2nc_\delta^t c_\delta + nc_\delta^t I_{(\Sigma\delta_j \times \Sigma\delta_j)} c_\delta \\
&= n(1 - c_\delta^t c_\delta) \ ,
\end{aligned}$$

where $\mathbf{X}_\delta = (\mathbf{x}_{1\delta_1}, \ldots, \mathbf{x}_{r\delta_r})$ is the matrix $\mathbf{X}$ from which columns with null subscripts have been removed. Based on estimates $\hat{\beta}_\delta$, minimizing the residual sum-of-squares with respect to the selection $\delta$ is thus exactly equivalent to maximizing the variance $c_\delta^t c_\delta$ with respect to $\delta$.

Instead of minimizing $\|\mathbf{y} - \mathbf{X}\beta\|_2^2/2n + \gamma \sum_j \beta_j$ with respect to $\beta$, we could thus maximize $\sigma^2(\delta) - \gamma\|\hat{\beta}_\delta\|_1 = c_\delta^t c_\delta - \gamma \sum_j c_{j\delta_j}$ with respect to $\delta$ (these approaches are in agreement, but are not exactly equivalent since Lasso is generally not restricted to estimates of the form $\hat{\beta}_\delta = c_\delta$). When $\gamma = 0$, there is no constraint and the maximum variance is $c^t c$; as $\gamma$ increases, we start dropping predictors. For this, we temporarily assume that $c_1 > c_2 > \ldots > c_r$; the $(r - k + 1)$-th predictor to be dropped is then $x_k$ and this happens when

$$\sum_{j=1}^{k} c_j^2 - \gamma \sum_{j=1}^{k} c_j \ < \ \sum_{j=1}^{k-1} c_j^2 - \gamma \sum_{j=1}^{k-1} c_j \ ,$$

which implies $\gamma > c_k$. In other words, every time $\gamma$ becomes larger than a new $\hat{\beta}_k = c_k$ value, then the corresponding predictor is discarded. This thus gives rise to an ordering of models from full to empty, as in Lasso.

For a neat visual representation of this selection process on the line $\mathcal{L}\mathbf{y}$, we may think of our line as pointing upwards, along with the fraction of $y$ density $f(y;\delta) = \sigma(\delta)\phi(y) = \{c_\delta^t c_\delta\}^{1/2}\phi(y)$ lying on it. We then apply a sort of upwards tilt, or moment generating type

penalty, on the density:

$$\begin{aligned}\exp\{\gamma y\} f(y;\delta) &= \sigma(\delta)\phi(y)\exp\{\gamma y\} \\ &\propto \sigma(\delta)\exp\{-y^2/2\}\exp\{\gamma y\} \\ &\propto \sigma(\delta)\exp\{(y-\gamma)^2/2\}\,,\end{aligned}$$

and note that this tilt $\exp\{\gamma y\}$ simply corresponds to a shift of the distribution in the direction of positive $\mathcal{L}\mathbf{y}$. Instead of being centered at 0, the $y$-content distribution is now centered at $\gamma$ and as the latter increases, the distribution $\mathcal{N}(\gamma, c_\delta^t c_\delta)$ shifts upwards. When the imaginary horizontal line at $\gamma$ goes over the vector $\mathbf{u}_j$ (or equivalently when $\gamma > \hat{\beta}_j = c_j$), then the variable $x_j$ is excluded from the model. We note here that the regression coefficient $\beta_j$ can be interpreted as the projection, on $\mathcal{L}\mathbf{y}$, of admissible $y$-information coming from $\mathbf{x}_j$; for instance, $\hat{\beta}_j = c_j$ means that the unit vector $\mathbf{u}_j$ is entirely admissible in the regression model, and so we use its projection $\beta_j = c_j$ on $\mathcal{L}\mathbf{y}$.

We refer to this method as the Linear Lasso since it describes the process of variable selection and coefficient estimation through a constrained maximization problem on a line (instead of a multidimensional space). The output of the Linear Lasso is thus an ordering of models (with a decreasing number of coefficients) whose forecasted values $c_\delta^t \mathbf{x}_\delta = c_\delta^t (x_{1\delta_1}, \ldots, x_{r\delta_r})^t$ algebraically correspond to least squares predictions. Of course, the above situation was extra simple as covariates were assumed to be independent from each others.

### 7.2. Correlated predictors

When correlation between pairs of covariates is present, the matrix $C$ is no longer diagonal. Even though covariates are positively correlated with $y$, some terms in $\hat{\beta} = C^{-1}c$ may be negative; this is however the exception rather than the rule. In fact, this only happens when covariates carry too much redundancy in their $y$-content; in that case, the coefficient of one of these variables may go slightly negative to avoid accounting for the same information several times.

As before, we aim at minimizing $\|\mathbf{y} - \mathbf{X}\beta\|_2^2/2n + \gamma \sum_j |\beta_j|$. As the constraint $\gamma$ on the total magnitude of coefficients grows, some regression coefficients are shrunken and eventually one of them is dropped (likely the smallest one, but not necessarily). Because of the correlation between predictors, the loss of information from shrinking or dropping one covariate can sometimes be partially recovered by adjusting the coefficient of one (or several) correlated covariate(s). It is thus difficult to know exactly in which order predictors are dropped as $\gamma$ increases. Lasso uses algorithms to find an approximate solution to this minimization problem in the multidimensional space.

In this paper, we use the fact that all available information about the response is projected onto the line $\mathcal{L}\mathbf{y}$ and propose an algorithm that searches for a solution along this line only. We then wish to minimize $\|\mathbf{y} - \mathbf{X}\beta\|_2^2/2n + \gamma \sum_j |\beta_j|$ with respect to $\beta$, but also wish to restrict our search along $\mathcal{L}\mathbf{y}$. We have already established that for a specific subset of predictors $\delta$, the forecasted value is $\mathbb{E}[y|\mathbf{x}_\delta] = c_\delta^t C_\delta^{-1} \mathbf{x}_\delta$ on $\mathcal{L}\mathbf{y}$. This algebraically corresponds to the prediction based on least squares estimates for a model $\delta$, which minimizes the residual sum-of-squares $\|\mathbf{y} - \mathbf{X}_\delta \beta_\delta\|_2^2$ for this model.

Having proposed a solution for the "estimation" part of the problem, we now tackle the "selection" part. The only parameter left to tune in the residual sum-of-squares is the model selection $\delta$; in fact, when trying to minimize $\|\mathbf{y} - \mathbf{X}_\delta \hat{\beta}_\delta\|_2^2 = \|\mathbf{y} - \mathbf{X}_\delta C_\delta^{-1} c_\delta\|_2^2$ with respect to

$\delta$, we find

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}_\delta C_\delta^{-1} c_\delta\|_2^2 &= \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{X}_\delta C_\delta^{-1} c_\delta + c_\delta^t C_\delta^{-1}\mathbf{X}_\delta^t\mathbf{X}_\delta C_\delta^{-1} c_\delta \\ &= n - 2nc_\delta^t C_\delta^{-1} c_\delta + nc_\delta^t C_\delta^{-1} C_\delta C_\delta^{-1} c_\delta \\ &= n(1 - c_\delta^t C_\delta^{-1} c_\delta)\ .\end{aligned}$$

Since the distribution of forecasted values satisfies $c_\delta^t C_\delta^{-1}\mathbf{x}_\delta \sim \mathcal{N}(0, c_\delta^t C_\delta^{-1} c_\delta)$ on $\mathcal{L}\mathbf{y}$, our search thus consists in finding the model that has the greatest variance $c_\delta^t C_\delta^{-1} c_\delta$, subject to a penalty $\gamma\sum_j |\hat\beta_{j\delta_j}| = \gamma\sum |C_\delta^{-1} c_\delta|$. Because estimates $C_\delta^{-1} c_\delta$ are a function of $\delta$ only, our objective may be recast as a trade-off between a high variance $c_\delta^t C_\delta^{-1} c_\delta$ and a small cardinality $\Sigma_j\delta_j$. Naturally, solving this problem is computationally intensive. We propose, in Section 9, an approximate solution when there are only a few predictors left in the model; we however need a simpler route for general cases.

We assume, as in Hastie, Tibshirani, & Wainwright (2015), that models with a very large number of covariates are sparse. This is a realistic assumption in our big data era, where only a small number of predictors are usually found to be significant. We may choose to reformulate the constrained optimization problem as the maximization of the variance $c_\delta^t C_\delta^{-1} c_\delta$ (with respect to $\delta$) subject to a penalty $\gamma\sum_j \hat\beta_{j\delta_j} = \gamma\sum C_\delta^{-1} c_\delta$ that automatically discards negative regression coefficients. In our sign-standardized framework where all predictors are positively correlated with the response, negative regression coefficients are an indication that the corresponding predictor is of marginal interest; it is thus appropriate to set coefficients to 0 when they try to go negative.

Now, one way to approximate a solution for this one-sided variable selection problem is to use the fraction of $y$ density $f(y;\delta) = \{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}\exp\{-y^2/2\}$ on $\mathcal{L}\mathbf{y}$, to which we then apply an upward tilt $\exp\{\gamma y\}$ (as in the independent framework). Conveniently, the moment generating type penalty $\exp\{\gamma y\}$ is one-sided and so the other side of this constraint is simply managed by discarding negative coefficients, which is consistent with the approach described in the previous paragraph. This penalty function therefore leads to the shifted model

$$f(y;\delta)\exp\{\gamma y\} = \{c_\delta^t C_\delta^{-1} c_\delta\}^{1/2}\exp\{-(y-\gamma)^2/2\}\ .$$

Visually, the selection approach is represented using an horizontal line at $\gamma$. As the tuning parameter $\gamma$ grows, the distribution moves upwards on $\mathcal{L}\mathbf{y}$; when $\gamma > \beta_j$, the corresponding predictor $x_j$ is dropped from the model and the variance $c_\delta^t C_\delta^{-1} c_\delta$ decreases accordingly.

We however realize that dropping predictors according to the size of their regression coefficients $C_\delta^{-1} c_\delta$ is not an option in practice. It involves estimating the coefficients and, therefore, inverting $C_\delta$ at every step. Recall that in our context, $\beta_j$ may be interpreted as the projection, on $\mathcal{L}\mathbf{y}$, of some portion $\hat{\mathbf{u}}_j$ (of the vector $\mathbf{u}_j$) that is admissible in the modeling of the response; as for the coefficients $\beta_j$ however, the vectors $\hat{\mathbf{u}}_j$ are not easily computed. We thus go to the next order and discard regression coefficients according to the slope of the associated $\hat{\mathbf{u}}_j$ on $\mathcal{L}\mathbf{x}_j$. Indeed, the closer are $\mathcal{L}\mathbf{y}$ and $\mathcal{L}\mathbf{x}_j$, the more likely is the projection of $\hat{\mathbf{u}}_j$ on $\mathcal{L}\mathbf{y}$ to be large compared to another vector that is less correlated with $\mathbf{u}_y$ (despite the fact that the lengths of the vectors $\hat{\mathbf{u}}_j$ might differ). In terms of our selection process, this means that we can focus on unit vectors only; when $\gamma$ goes over the vector $\mathbf{u}_j$ (or its projection $c_j$ on $\mathcal{L}\mathbf{y}$), we drop the associated explanatory variable. This is immensely convenient as it leads to an iteration-free approach: we end up dropping predictors according to their projection $c_j$ on $\mathcal{L}\mathbf{y}$, as before, and these projections are constant across iterations! We eventually obtain an ordering of models from moving up on the response line, referred to as the Linear Lasso.

## 8. MANY EXPLANATORY VARIABLES

The main advantage of the Linear Lasso is that it leads to an ordering of nested models, without requiring any iteration. It also allows to drop either a single predictor, or several explanatory variables at once, on the basis of the vector $c$. Because elements in $c$ are fixed, proceeding forward or backward does not make any difference and leads to a consistent ordering of models.

Accordingly, since the projections $c_j$ are constant across iterations, this variable selection approach easily manages the case $r >> n$. It suffices to drop $r - n$ predictors simultaneously according to the $c_j$'s; once we are left with $n$ explanatory variables or less, we can obtain coefficient estimates and forecast values. It is however more likely that we wish to further reduce the model size, until we reach a parsimonious fit. In that case, we just drop more predictors on the basis of their $c_j$'s, until we have a model of the desired cardinality (or we can perform cross-validation to find the best model, as will be done in the examples of Sections 10 and 11). When there are few predictors left, we may also choose to fine-tune the model search by taking the correlation between predictors into account, as discussed in Section 9.

## 9. ONE-BY-ONE

One of the strengths of the above procedure based on the $c_j$'s is that the ordering of models does not require iterations, and is consistent in a forward or backward selection approach. Naturally, the resulting sequence of models does not necessarily perfectly agree with the ordering that would result from maximizing the variance $c_\delta^t C_\delta^{-1} c_\delta$ subject to a certain budget $\Sigma_j \delta_j$. In particular, the approach does not account for the fact that two correlated predictors may carry some redundancy in their $y$-content. To overcome this problem, we may want to fine-tune the above procedure when the number of remaining predictors is manageable.

To fine-tune the method and account for correlation between predictors, we can use a dual approach in the Linear Lasso: first drop $m$ explanatory variables according to $c$; then, when there are only a few explanatory variables left, perform a more thorough search and discard the variable that leads to the smallest drop in the variance $c_\delta^t C_\delta^{-1} c_\delta$. The fine-tuning part is then equivalent to a simple backward selection in which the criterion relies on the variance $c_\delta^t C_\delta^{-1} c_\delta$.

The number $m$ of predictors that are discarded on the basis of $c$ is user-selected. We could choose to fine-tune when there are only 5 predictors left, for instance. We could also let $m$ be the number of explanatory variables with $c_j < 0.2$; we would then drop the first $m$ variables according to $c_j$, and then fine-tune the sequence using the above backward approach on the remaining $r - m$ variables. In the next two sections, we implement the Linear Lasso on two different datasets; we then use cross-validation to find the pair $(\Sigma\delta_j, m)$ that leads to the best model.

## 10. EXAMPLE: CRIME DATA

In this section, we study a simple example with $r = 5$ predictors; this allows visualizing the Linear Lasso process, and also comparing it to the regular Lasso. The Linear Lasso approach selects a more parsimonious model than the regular Lasso; based on the mean-squared errors (MSEs) obtained through cross-validation, it chooses a model containing only 2 explanatory variables ($x_1^*$ and $x_3^*$) while Lasso selects a 3-variable model ($x_1^*$, $x_2^*$, and $x_3^*$). The model selected by the Linear Lasso has a smaller MSE than that of Lasso; this is not overly surprising given that our approach does not rely on shrunken regression coefficient estimates as Lasso does. A look at Tables 3 and 4 tells us that this holds quite generally: Lasso produces higher MSEs than Linear Lasso.

According to our explorations, it appears like a reasonable approach to let $m = \{\#j \in \{1, \ldots, r\} : c_j < 0.2\}$. Indeed, in practice, we seldom have several variables that are highly correlated with the response. Ordering models according to the correlations $c$ while the $c_j$ elements are "small", and then using a backward, variance-based approach to put the finishing touch on the selection, is thus an interesting avenue in general contexts.

The 2-variable, cross-validation model selected by the Linear Lasso (which, as it turns out, corresponds to $m = \{\#j \in \{1, \ldots, r\} : c_j < 0.2\} = 3$) does not really come as a surprise when we look at the available predictors. The variables $x_2$, $x_4$, and $x_5$ present different measures of the population's education level, while the variable $x_3$ measures the extent to which the population is uneducated. These predictors are thus highly correlated, which explains why the Linear Lasso only keeps one of them in its optimal model. Contrarily to the standard Lasso, which does not care which variable is selected in a group of highly correlated predictors, the Linear Lasso uses a coherent approach to drop explanatory variables in presence of high correlation.

## 11. EXAMPLE: MATHEMATICS GRADES

In this example, the correlations between predictors are quite weak: 79% of elements in the matrix $C$ are $\leq 0.1$, and 94% are $\leq 0.2$. One of the few cases where correlation is highly significant happens to be between first- and second-period grades; despite this correlation, Linear Lasso and standard Lasso keep both predictors in the model. Other predictors featuring significant pairwise correlations are mother's education and mother's job, school attended and home address, etc. Given the fact that these predictors are only slightly correlated with the response, these pairwise correlations between predictors do not appear to have an impact on the final model selection.

## 12. DISCUSSION

In the introduction, we identified two limitations of the standard Lasso: its computational complexity and inconsistent approach in dropping correlated explanatory variables. The proposed Linear Lasso addresses these two limitations; we now summarize the foundations of this new approach.

Using the latent normal model, the conditional distribution of $y$ given the predictors included in a selection $\delta$ is on the line $\mathcal{L}\mathbf{y}$; the maximum likelihood estimate (MLE) for the response $y$ is found to be the forecasted value $\mathbf{x}_\delta C_\delta^{-1} c_\delta$ on $\mathcal{L}\mathbf{y}$. The function to maximize then becomes the variance of this forecasted value, $c_\delta^t C_\delta^{-1} c_\delta$. In our sign-standardized context, the penalty on the function to maximize may be re-expressed as $\exp\{\gamma y\}$; according to this penalty, we thus automatically discard the regression coefficients that are negative since they indicate that their $y$-content is largely redundant with that of other predictors.

When $\gamma = 0$, there is no penalty on the magnitude of the coefficients. The forecasted value $c^t C^{-1}(x_1, \ldots, x_r)^t$ is thus the MLE of the predictive distribution when all explanatory variables are present. Now, it may help to picture each regression coefficient $\beta_j$ as the projection, on $\mathcal{L}\mathbf{y}$, of the model-admissible portion of the vector $\mathbf{u}_j$ on $\mathcal{L}\mathbf{x}_j$. This portion of the vector points in $\mathcal{L}^+\mathbf{y}$ from the origin when the coefficient is positive and points below $\mathcal{L}^\perp\mathbf{y}$ when it is negative; its length is directly related to the magnitude of the associated regression coefficient.

As a selection rule, we thus start by discarding explanatory variables whose coefficients completely find themselves in the lower half-space. Now, as $\gamma$ grows, we may picture the whole system of regression coefficient vectors being pulled down along $-\mathcal{L}\mathbf{y}$ (this is an alternative visualization to imagining a horizontal line at $\gamma$ that moves up $\mathcal{L}\mathbf{y}$). As these coefficient vectors find themselves totally below 0 (i.e., as the projection $\beta_j$ on $\mathcal{L}\mathbf{y}$ is below 0), they are discarded. This leads to an ordering of models according to the projection of coefficient vectors on $\mathcal{L}\mathbf{y}$. Naturally, to propose a solution (or algorithm) that is manageable in practice, we have to approximate

this process using the $c_j$'s, just like Lasso provides an approximate solution to its constrained minimization problem.

Before concluding, we note that the goal of the sign standardization in the development of the Linear Lasso was to make the problem one-sided and thus easier to visualize. In theory, the selection process based on the penalty $\exp\{\gamma y\}$ discards negative regression coefficients. In practice, we cannot afford to compute these coefficients at every iteration and thus propose an approximate solution based on the slope of the vectors associated to these coefficients. The procedure detailed in Section 9, which relies on this approximate solution, may thus be easily implemented without having to worry about the sign standardization step; we however still need to standardize the vectors' mean and standard deviation in order to implement the Linear Lasso.

## BIBLIOGRAPHY

Hastie, T., Tibshirani, R., & Wainwright, N. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.