International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2013), 81, 3, 361–387 doi:10.1111/insr.12023

A Survey of L_1 Regression

Diego Vidaurre[1], Concha Bielza[2] and Pedro Larrañaga[2]

[1]*Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, UK*
E-mail: diego.vidaurre@ohba.ox.ac.uk
[2]*Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain*

Summary

L_1 regularization, or regularization with an L_1 penalty, is a popular idea in statistics and machine learning. This paper reviews the concept and application of L_1 regularization for regression. It is not our aim to present a comprehensive list of the utilities of the L_1 penalty in the regression setting. Rather, we focus on what we believe is the set of most representative uses of this regularization technique, which we describe in some detail. Thus, we deal with a number of L_1-regularized methods for linear regression, generalized linear models, and time series analysis. Although this review targets practice rather than theory, we do give some theoretical details about L_1-penalized linear regression, usually referred to as the least absolute shrinkage and selection operator (lasso).

Key words: Regularization; lasso; L_1-regularized regression; sparsity.

1 Introduction

The use of L_1 regularization for statistical inference has become very popular over the last two decades. Although the concept of L_1 penalty dates back further, its application for regularization gained significant impetus after Tibshirani (1996) proposed the *least absolute shrinkage and selection operator* (or lasso) technique. The lasso framework in Tibshirani (1996) used an L_1-penalized likelihood for linear regression with independent Gaussian noise, so it involves minimizing the usual sum of squared error loss with L_1 regularization. The lasso has become the standard tool for sparse regression. Throughout the paper, sparse regression refers to situations where only a relatively small subset of the regression coefficients are non-zero. For example, sparse inverse covariance estimation means that the corresponding graphical model only has a subset of all possible edges.

The emergence of the *least-angle regression* (LARS) algorithm (Efron *et al.*, 2004) provided an efficient solution to the optimization problem underlying the lasso (for linear regression), and this was another key to the rapid spread of the lasso within the statistics and machine learning communities. See Hesterberg *et al.* (2008) for a careful description of the lasso and its connections to LARS by means of the LARS algorithm. More recently, pathwise coordinate descent methods have been proposed for solving the lasso problem efficiently (Friedman *et al.*, 2007; Wu & Lange, 2008).

L_1 regularization has attracted a lot of attention because of its ability to do variable/feature/model selection. As noted by Hesterberg *et al.* (2008), some researchers view this selection problem as one of the more important ones in modern statistics. Traditionally, model

selection was built on techniques such as forward stepwise regression, all subsets regression or prefiltering approaches. Some of these approaches are based on univariate measures and are thus known to be seriously biased. Others, such as all subset regression, are not computationally affordable in applications where the number of variables (or features and hence models) is moderate to large. The major advantage of the lasso and related methods is that they offer interpretable, stable models, and an efficient prediction at a reasonable cost (although they are not exempt from some bias). Specifically, the lasso/LARS approach has the same computational cost of least-squares estimation. Further, it provides a simple, data-driven approach for selecting the optimal level of model complexity, that is, how much the model should be regularized. A recent review by Fan & Lv (2010) discusses regularization for variable selection, placing special emphasis on ultra-high dimensionality. See also Bühlmann & van de Geer (2011) where the lasso for high-dimensional problems is extensively discussed, including the theory behind the models.

Although the lasso was proposed in 1996, regularization is a relatively old concept. It was devised by Tikhonov (1943) for approximating the solution of a set of unsolvable integral equations. This concept is the basis for ridge regression, which was formally introduced by Hoerl & Kennard (1970) almost 30 years later. Generally speaking, regularization introduces some constraint on the parameters to solve an inference problem, such as maximum likelihood estimation, that is unstable or cannot be solved by regular methods. In other words, regularization imposes trades some bias in exchange for a larger reduction in variance and hence avoids overfitting. The regularized solutions are more stable and typically less complex. See Bickel & Li (2006) for an excellent general review of regularization in statistics.

This paper considers selected areas that deal with L_1 regularization with a special attention on the lasso. The paper is organized as follows. Section 2 introduces the notation and describes the lasso for linear regression. Section 3 reviews a number of extensions that aim to improve the statistical properties of the lasso. Section 4 deals with L_1-regularized methods that addresses the specific problem configurations. Section 5 discusses the lasso for generalized linear models. Section 6 describes some approaches using L_1 regularization for time series analysis. Finally, Section 7 draws some conclusions.

2 The Lasso for Linear Regression

L_1 regularization for linear regression with independent Gaussian errors (and hence a likelihood with squared error loss) is probably the most popular incarnation of the lasso. This problem is explored in the current section, which is divided into several subsections. An introduction about linear regression and the lasso is presented in Subsection 2.1. We briefly discuss some theoretical details about the lasso in Subsection 2.2. Subsection 2.4 presents the Bayesian interpretation of the lasso. Subsection 2.3 discusses computational algorithms for solving the lasso optimization problem and related matters. Finally, Subsection 2.5 connects the lasso with boosting.

2.1 Notation and Main Concepts

The general linear regression problem can be formulated as follows. Denote the set of p input variables as $\{X_1, \ldots, X_p\}$ and the (scalar) response variable as Y. Let $\boldsymbol{D} = \{(x_{i1}, \ldots, x_{ip}, y_i),\ i = 1, \ldots, N\}$ be the N observations. Simple regression with one input variable corresponds to $\boldsymbol{x}_i \in \mathbb{R}^p$. We denote the vector of responses as $\boldsymbol{y}$, the $N \times p$ predictor data matrix as $\boldsymbol{X}$, and its columns as $\boldsymbol{X}_{\cdot j} \in \mathbb{R}^N$, that is

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_i' \\ \vdots \\ \boldsymbol{x}_N' \end{pmatrix} = (\boldsymbol{X}_{\cdot 1} \ \ldots \ \boldsymbol{X}_{\cdot j} \ \ldots \ \boldsymbol{X}_{\cdot p}) = \begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{N1} & \ldots & x_{Nj} & \ldots & x_{Np} \end{pmatrix}.$$

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Suppose the errors ε_i are independent and identically distributed Gaussian random variables with mean 0 and variance σ^2. The lasso (Tibshirani, 1996) estimation minimizes the residual sum of squares (equivalent to maximizing the likelihood) subject to an L_1 constraint:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \, ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \qquad \text{s.t.} \quad ||\boldsymbol{\beta}||_1 \leq s, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $s \geq 0$ and $||.||_q$ is the q-norm. Equivalently, the lasso can be defined in the Lagrangian form as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \, ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{2}$$

where the regularization parameter $\lambda \geq 0$ has a one-to-one correspondence with the parameter s of Equation (1). Thus, the lasso estimator substitutes the L_2 penalty of the ridge estimator (Hoerl & Kennard, 1970) with an L_1 penalty. This optimization problem is convex but, due to the L_1-penalty, not differentiable at zero. For a given data set with finite sample size, it is always possible to obtain a lower expected error with a biased (regularized) estimation (James & Stein, 1961).

Typically, the data are standardized so that the penalty is invariant with regard to the scale of the variables. If the data are not centered, a non-penalized intercept should be included in the vector $\boldsymbol{\beta}$.

Unlike the ridge regression problem, the lasso solution cannot, in general, be given in a closed-form expression. An exception is the case of an orthonormal input matrix $\boldsymbol{X}$ where the lasso solution is

$$\hat{\beta}_j = \text{sign}\left(\hat{\beta}_j^{ls}\right)\left(\hat{\beta}_j^{ls} - \lambda\right)_+, \quad j \in \{1, \ldots, p\}, \tag{3}$$

where $\hat{\beta}_j^{ls}$ is the least squares estimator for the j-th variable and $(\cdot)_+$ indicates the positive part. This is called *soft-thresholding*. As a comparison, we can write the ridge solution in the orthonormal case as

$$\hat{\beta}_j = \hat{\beta}_j^{ls}/(1 + \lambda), \quad j \in \{1, \ldots, p\}. \tag{4}$$

Figure 1 compares the ridge and lasso estimators for $p = 1$. The X-axis represents the unrestricted coefficient $\hat{\beta}_1^{ls}$, and the Y-axis represents the corresponding regularized coefficient $\hat{\beta}_1$ (the ridge in the left panel and the lasso in the right panel). The dotted lines in each panel correspond to the unrestricted least squares estimation.

The lasso, which is particularly useful when the number of inputs is larger than the number of samples (or $p > N$), performs variable selection by driving a number of regression coefficients to be exactly zero, thanks to the non-differentiable nature of the L_1 penalty at the origin. A fairly small value of λ leads to a solution that is close to the least squares estimator. As we increase

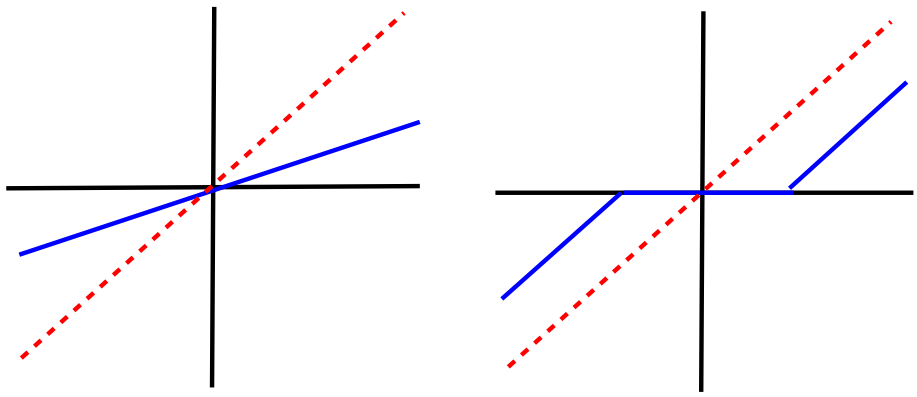

Figure 1. *Ridge (left) and lasso (right) estimation of a regression coefficient in the orthonormal case. The dotted line corresponds to the unrestricted least squares estimation.*

the value of λ, one coefficient at a time is made equal to zero, although some variables can sporadically exit the model in the presence of correlated inputs. In other words, the λ parameter controls the degrees of freedom of the estimation. Thus, by increasing the value of λ, we can control the number of variables that are included in the model. Note that different solutions to the lasso problem can be obtained when $p > N$ because, in this case, the problem is not strictly convex. However, it can be proved that the same sparsity pattern is shared among all these solutions.

Some discussion about the 'degrees of freedom' df is useful, for example, for tunning λ if we are to use some information criteria statistic, such as C_p, the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which include the number of degrees of freedom of the fitted model. For example, AIC is defined as

$$AIC = \log ||\hat{\boldsymbol{y}} - \boldsymbol{y}||_2^2 + 2\, df,$$

where $\hat{\boldsymbol{y}}$ is the predicted response. Although AIC focuses on maximizing the expected accuracy, the BIC criterion is typically recommended to achieve sparser models. The number of degrees of freedom in a Gaussian linear model with independent errors is given exactly by the trace of the hat matrix operator, which maps from the observed response to the predicted response. Because the lasso is a nonlinear fitting method (in the space spanned by $(\lambda, \boldsymbol{\beta})$), this formula cannot be applied, and an analytical expression is hard to derive. However, a simple unbiased estimator of the degrees of freedom of the lasso estimator can be given simply by the number of nonzero coefficients. It may seem that by shrinking the coefficients, we should be reducing the number of degrees of freedom, but this is compensated by the lasso's "freedom" for selecting certain variables and discarding others. A complete discussion of the estimation of the degrees of freedom for the lasso in the framework of Stein's unbiased risk estimation is given by Zou *et al.* (2007). Unfortunately, this nice simple rule has been proved only for the $p < N$ case, and it is not known if the result holds when $\boldsymbol{X}$ is not full-rank. In this case, if the computations are affordable, one may resort to bootstrapping techniques for estimating the number of degrees of freedom Efron (2004).

Although the L_1 and the L_2 penalties are the most popular, other L_q penalties are possible. The bridge regression (Frank & Friedman, 1993) generalizes lasso and ridge to $q > 0$. In this case, the penalty in Equation (2) becomes $\lambda\, ||\boldsymbol{\beta}||_q^q$. By using $q < 1$, we can reduce the bias of the lasso estimation, although the computation becomes more challenging because of non-convexity of the penalty function.

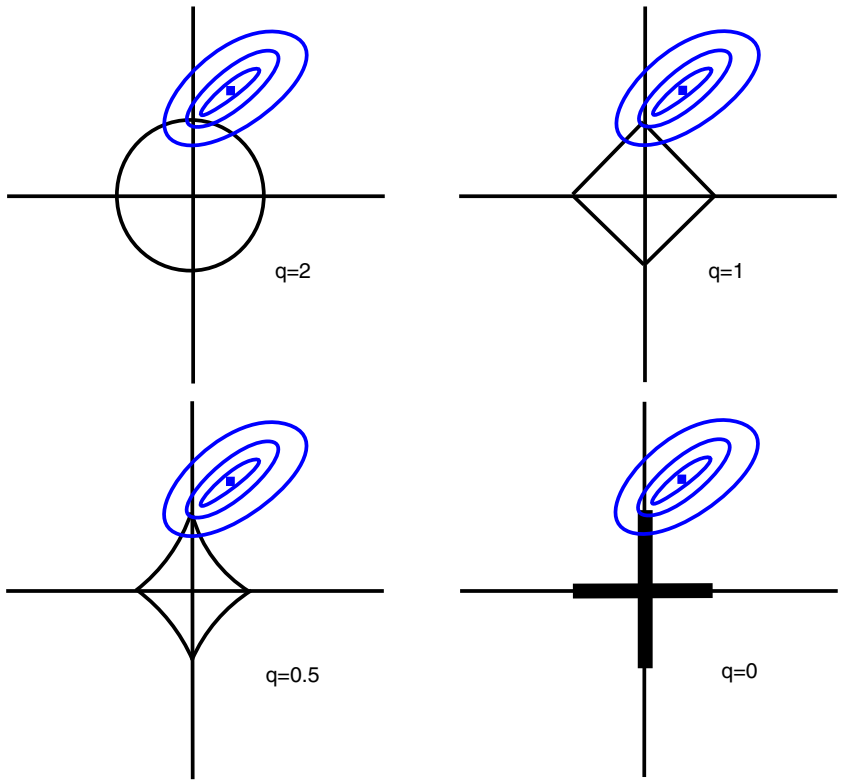

Figure 2. *Concentric contours of the least squares function and contour of the constraint region for some penalties.*

Figure 2 illustrates the optimization problem for two variables with various penalties. The solution of the penalized problem is given at the point where the elliptical contour of the least squares function hits the constraint region centered at the origin. In this example, variable selection will occur (depending on the magnitude of λ) for $q \in \{0, 0.5, 1\}$ but not for $q = 2$. The reason is rather intuitive: when the penalty function is non-differentiable at points where $\beta_j = 0$, it forms a corner in the axes, which promotes that the loss and the penalty functions precisely meet in the axis. This means that the corresponding β_j is exactly zero. With ridge regression ($q = 2$), there are no corners in the axes, and the probability of exactly $\beta_j = 0$ is infinitely close to zero. Note that when $q = 0$, the penalty is just the number of free parameters. Note also that the operator $||\cdot||_q$ is a proper norm only for $q > 1$.

Therefore, the L_1 penalty's biggest advantage lies in its compromise: variable selection only takes place when $q \leq 1$ (non-differentiability of the penalty function), whereas the related optimization problem is convex only when $q \geq 1$. The lasso is the intersection of both conditions so it achieves variable selection without surrendering the computational advantages of convexity. In other words, the L_1 penalty is the closest convex $||\cdot||_q^q$ operator to the L_0 penalty and the only one that is convex and non-differentiable at the origin.

Another advantage of the lasso over ridge is that predictions are less biased under some conditions. Ridge pushes all the coefficients towards zero with a force that depends on the regularization parameter and is proportional to the magnitude of the coefficient. In the orthonormal case, for example, the ridge solution is given by Equation (4). The lasso, on the other hand, further shrinks coefficients that are effectively discarded. This results in a weaker shrinkage of

the coefficients of the variables that remain within the model. These are supposed to be the most valuable predictors for the regression problem, that is, they have a bigger influence on the response. There are, however, other situations where ridge outperforms the lasso. For example, ridge dominates the lasso when the predictors are highly correlated (Tibshirani, 1996; Zou & Hastie, 2005). In this case, ridge shrinks the coefficients of all redundant variables so that the total contribution is well-balanced, whereas the lasso tends to drop all but one of the redundant variables. In general, the lasso will perform considerably better if the underlying model is indeed sparse. If the underlying model is dense (and there are sufficient data, which is often not the case), ridge regression is the right choice. Some details about the concrete properties of the lasso are given in Subsection 2.2.

2.2 Statistical Properties

Statistical performance can be measured in two different ways: how accurately can the lasso predict a new response from new data and how well does the lasso estimate the true regression coefficients. In the second case, assuming that the true parameter vector is indeed sparse, the main interest is to recover the true sparsity pattern. Next, we present some basic results concerning these goals. Although both objectives are related, they require different conditions on the input design. For example, variable selection optimality is more ambitious than prediction optimality and needs stronger assumptions. We discuss a property that is usually more realistic than exact variable selection: variable screening. We assume throughout the exposition that the true regression function is linear. For further results and their mathematical justification, see Bühlmann & van de Geer (2011).

Greenshtein & Ritov (2004) studied the prediction consistency of the lasso, proving that, under mild regularity conditions and if the true parameter vector is sparse enough, the expected squared prediction error approximates the irreducible error (the Bayes error) for an adequate choice of λ. This value depends on N and p. A considerably faster convergence rate can be obtained, for example, if there are no zero eigenvalues in the Gram matrix $\boldsymbol{X}'\boldsymbol{X}/N$. In this case, the lasso achieves a squared error that is not far from what could be achieved if the true sparsity pattern were known. Results under more refined (weaker) conditions have been developed in the statistical literature. These include the compatibility condition (van de Geer & Bühlmann, 2009), the restricted eigenvalue conditions (Bickel *et al.*, 2009), the coherence conditions (Bunea *et al.*, 2007), or the restricted isometry conditions (Candès & Tao, 2007). These conditions, however, are difficult to check in practice. In a yet unpublished paper, Chatterjee (2013) showed that these conditions can be further reduced to a minimum if we restrict ourselves to the sum of squared error lost function.

Besides prediction performance, the interpretability of the model is often a primary goal. The focus is typically on the identification of a simple enough model rather than on the best prediction accuracy. A substantial amount of research has looked at the lasso's ability to recover the true sparsity pattern (i.e., to discard the irrelevant variables and only those) and give a consistent estimation of the non-zero coefficients.

Zhao & Yu (2006) discussed the *irrepresentable condition* concept, already outlined by Meinshausen & Bühlmann (2006). Briefly, Zhao & Yu (2006) showed that the true model can be recovered only if there are no high correlations between relevant predictors and irrelevant predictors. This is a sufficient and necessary condition for the lasso to achieve consistent variable selection. The irrepresentable condition depends on the Gram matrix and the sign of the true parameters and can be formalized as follows:

$$\max \left(\left(\boldsymbol{X}'_{\vartheta} \boldsymbol{X}_{\vartheta} \right)^{-1} \boldsymbol{X}'_{\vartheta} \boldsymbol{X}_{\vartheta^c} \right) < 1 - \epsilon,$$

where $\boldsymbol{X}_{\vartheta}$ includes the variables whose coefficients are substantially different from zero in the true model, $\boldsymbol{X}_{\vartheta^c}$ includes the variables with zero coefficients in the true model, and ϵ is some positive constant. Hence, we further assume that the true non-zero parameters are sufficiently large in absolute value. This is called the *beta-min condition*.

More recently, Meinshausen & Yu (2009) examined the behavior of the lasso when only a relaxed version of the irrepresentable condition is met. Specifically, although the true sparsity pattern cannot be exactly recovered, the estimation of the coefficients can still be consistent (in the L_2 sense) if both the number of relevant variables and the minimum eigenvalue of the design matrix (restricted to the relevant variables) are bounded.

Despite its theoretical interest, the irrepresentable condition may hold only under restricted situations. Although it is commonly used in theoretical demonstrations, the beta-min condition is not always met either. Often, a more reasonable goal is to select the relevant variables but allow for a moderate number of false positives. This is typically referred to as variable screening or variable filtering. Variable screening is very useful for dimension reduction in high-dimensional settings.

A practical problem with exact variable selection is the choice of the regularization parameter λ. It is well known that the optimal value of λ for prediction, usually selected by cross-validation, is higher than the value required for variable selection. Because the performance of statistics such as AIC and BIC is not theoretically proved for variable selection, it is difficult to select the adequate value of λ for this purpose. The value of λ obtained by cross-validation is, however, typically acceptable for variable screening if we allow the inclusion of some irrelevant variables in the model.

So far, we have assumed a fixed design, that is, that there is no uncertainty in the covariates. There are mirroring results for random covariates, that is, when these are subject to noise. Also, there are extensions of the aforementioned results for the case when the true regression function is not linear. See Bühlmann & van de Geer (2011) for an overview.

Finally, Xu *et al.* (2010) shows the connection between the lasso formulation and the robust regression min–max problem:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \{\max_{\boldsymbol{Z} \in \Theta} ||\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{Z})\boldsymbol{\beta}||_2\},$$

where $\boldsymbol{Z}$ is a disturbance of the design matrix and Θ is called the uncertainty set. The connection between the lasso and robust regression holds for a certain type of uncertainty set, which bounds the L_2 norm of the columns of its elements. This indicates that the lasso has a valuable robustness property. They use this robustness property to give a novel prediction-consistency proof of the lasso.

2.3 Computational Algorithms

The LARS algorithm (Efron *et al.*, 2004) has boosted the practical use of the lasso enormously. LARS computes exactly the regularization path for the L_1-penalized least squares problem if the input covariates are orthogonal. The *regularization path* is the entire set of solutions for each λ value. With a slight modification of the basic algorithm, LARS is able to compute the entire lasso regularization path for the general case at the cost of a single ordinary least square fit.

To compute the regularization path efficiently, the LARS algorithm takes advantage of its linearity. More specifically, because the regularization path is piecewise linear for the lasso, we need to compute only the solution for a finite number of λ values (knots of the regularization path). These values represent a variable's removal from or addition to the model. Furthermore,

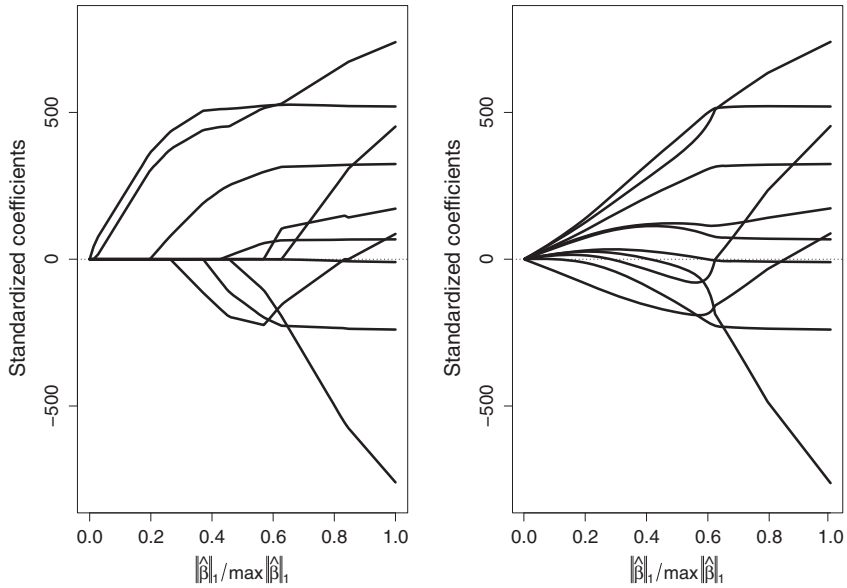

Figure 3. *Regularization path for the Diabetes data set for the lasso (left) and ridge (right) regression.*

the minimum value of λ that makes all regression estimates vanish is given analytically by $\lambda_{\max} = \max_j |2\boldsymbol{X}_{.j}\boldsymbol{Y}|/N$.

We use a simple version of the *Diabetes* data set (see Efron *et al.* (2004)) to compare the results. The *Diabetes* data set has $N = 442$ subjects and $p = 10$ binary or continuous covariates, which have been standardized for convenience. The response to predict is some measure of the disease progression. This data set is well suited to be used as a benchmark for penalized regression in the $N > p$ case. Figure 3 shows the regularization path for for the lasso (left) and ridge (right) regression. The Y-axis represents the magnitude for the regression coefficients and the X-axis represents the scaled L_1 norm of the estimated vector of coefficients. Each coefficient is represented by a different line. All the coefficients are zero at the start of the regularization path, where λ takes a high value (say, $\lambda_{\max}$). As λ decreases (moving right in the figure), the coefficients evolve towards the least squares solution. Note that for any $\lambda < \lambda_{\max}$, all coefficients are different from zero in the ridge solution.

The LARS algorithm sequentially builds a regularization path by adding a predictor to the model at each step. It starts with no predictors; then it includes the predictor that is most correlated with the response, say $\mathcal{A}$, into the active set of predictors. At each iteration, the predictors in $\mathcal{A}$ are regressed on the current residual, being their coefficients pushed in the direction of the least squares solution. If we progressively updated the coefficients, we would see that the absolute correlation between the current residual and the predictors in $\mathcal{A}$ smoothly decreases. The coefficients are thus not updated all the way until the least squares solution but only until a new predictor not in $\mathcal{A}$ reaches the same absolute correlation with the residual than the predictors in $\mathcal{A}$. Then this predictor is included in $\mathcal{A}$, and the next iteration begins. The essence of the LARS algorithm is that, to save computation time, it uses standard algebra to exactly calculate the updates of $\mathcal{A}$ and to select the predictor to be added at each step. For the case $N \geq p$, this procedure is repeated until all predictors are into the model. Otherwise, after $N - 1$ steps, the residuals are zero, and the algorithm terminates. The LARS modification for computing the exact regularization path of the lasso problem is based on detecting when a non-zero coefficient hits zero.

Khan *et al.* (2007) proposed a robust version of LARS, although they did not explicitly state which optimization problem is solved by this approach (certainly, not the usual lasso problem (2)). Fraley & Hesterberg (2009) devised a variation for dealing with large data sets. Rosset & Zhu (2007) analyzed the L_1-penalized problems with general loss functions and ascertain the conditions under which a LARS procedure can be applied. Assuming a fixed design, for the regularization path to be piecewise linear, there are two sufficient conditions: (a) the loss function is quadratic as a function of $\boldsymbol{\beta}$ and (b) the penalty function is piecewise linear as a function of $\boldsymbol{\beta}$ along the regularization path (as is the L_1 penalty). As another example, the authors designed a LARS algorithm for an L_1-penalized Huber's loss function. Wang & Leng (2007) transformed different types of loss functions into quadratic approximations that can be computed by the LARS algorithm.

Although LARS is considerably fast, there exist pathwise coordinate descent optimization algorithms that can be more efficient than LARS in high-dimensional settings. The theory under pathwise coordinate optimization is developed by Tseng (2001) and Tseng & Yun (2009). Friedman *et al.* (2007) and Wu & Lange (2008) exploited it for solving convex statistical problems such as the lasso. In a nutshell, coordinate descent optimization updates one coordinate (or block of dependent coordinates) at a time in an iterative fashion, leaving the others fixed, until convergence is reached. In some cases, this update can be analytically carried out, yielding an extremely fast and simple algorithm. For the lasso, for example, the update corresponds, after reparametrizing, to the soft-thresholding operator (3). This way, the lasso problem (2) can be solved for a certain λ value. For a lower value of λ, the former solution can be used as a warm start. Instead of obtaining the solutions at each knot in the regularization path (as LARS does), this procedure is typically used for obtaining the solutions on an equispaced grid (on the log scale) of λ values. We must stress that coordinate descent optimization is computationally efficient to the extent that the coordinate updates are efficient.

Unlike LARS, coordinate descent optimization can deal, for example, with non-squared losses and can thus be used in a wider range of problems. Generally speaking, this methodology is applicable when the loss of function is differentiable and convex and the penalty is convex and separable—although some tricks are possible when it is not separable; see, for example, Friedman *et al.* (2007) for the application of coordinate descent optimization to the fused lasso (Tibshirani *et al.*, 2005), which will be described in Section 4.

2.4 A Bayesian Interpretation

We can interpret the least squares solution as the maximum a posteriori (MAP) estimate with a non-informative prior on the coefficients. A shrinkage prior centered at zero for the parameters leads to more stable estimates. The double exponential or Laplace prior is a relevant example because it can lead to sparsity. Although some authors had resorted in earlier papers that the Laplace prior leads to sparse models (Williams, 1995; Figueiredo, 2003; Yuan & Lin, 2005), the explicit characterization of the Bayesian version of the lasso was carried out in Park & Casella (2008). The Laplace prior on $\boldsymbol{\beta}$ has the form

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp \frac{-\lambda|\beta_j|}{\sigma}.$$

This prior has attracted the attention of the researchers in Bayesian inference partly because it can be formulated as a scale mixture of normal distributions with independent exponentially distributed variances (Andrews & Mallows, 1974). This property allows the usage of easy-to-implement Gibbs samplers and EM algorithms for making inference.

Assuming centered data, the lasso's Bayesian hierarchy is thus as follows:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N\right), \\
&\boldsymbol{\beta}|\sigma^2, \varsigma_1^2, \ldots, \varsigma_p^2 \sim \mathcal{N}\left(\mathbf{0}_N, \sigma^2 \text{diag}\left(\varsigma_1^2, \ldots, \varsigma_p^2\right)\right), \\
&\varsigma_j^2 \sim \mathcal{E}(0, \lambda),
\end{aligned}$$

where $\boldsymbol{I}_N$ and $\mathbf{0}_N$ are $N \times N$ identity and zero matrices, respectively. Also, we can use an improper prior $\pi\left(\sigma^2\right) = 1/\sigma^2$ or any inverse-gamma prior. For the regularization parameter λ, Park & Casella (2008) suggested either taking an empirical Bayes approach (Bernardo & Smith, 1994) or including λ in the hierarchy with a gamma hyperprior.

From the aforementioned Bayesian hierarchy, it is not hard to derive the full conditional distributions of the parameters, so that, for example, a Gibbs sampler (Geman & Geman, 1984) can be used for inference. Park & Casella (2008) reported that the convergence of this Gibbs sampler is reasonably fast.

It is worth noting that, as shown by Park & Casella (2008), conditioning on σ^2 is mandatory for obtaining a unimodal full posterior. Note also that, because the Bayesian approach considers uncertainty in the parameter estimates, the full posterior is only sparse in the limit of infinite data even though the prior distribution encourages sparse estimation. Hence, only the MAP estimation is sparse. Alternative choices, which do not achieve sparsity, are the posterior median and means. The posterior median estimate, for example, appears to be a compromise between the ridge regression and the frequentist lasso estimates.

A primary advantage of the Bayesian lasso is that it also provides reliable standard errors for the regression coefficients. Although bootstrapping techniques are always feasible for assessing the standard errors, it is not always ideal. Knight & Fu (2000) show that bootstrapping inevitably introduces a bias in the estimation and can be sometimes unstable, making inference complicated for the frequentist lasso. In fact, it can be shown that bootstrap standard errors are inconsistent when $\beta_j = 0$.

Another advantage of the Bayesian lasso is that it can reveal some information about relevant dependence between predictors through the joint posterior distribution of such coefficients. Whereas the lasso behaves rather arbitrarily in the presence of strong dependence, some extensions of the lasso admittedly cope with this issue within the frequentist paradigm (see Section 4). Although not discussed in this paper, many of these extensions can also be formulated in a Bayesian fashion (Kyung *et al.*, 2010). Finally, note that, in general, the Bayesian lasso is more computationally expensive than its frequentist counterpart.

2.5 Connection to Boosting

Boosting was born as an ensemble method for classification and regression that sequentially evolves a weighted set of base learners, where the weights are updated at each iteration (Hastie *et al.*, 2008). The idea is powerful and yields flexible algorithms by choosing among different loss functions and base learners. Alternatively, boosting can be formulated as a gradient descent method in a functional space. This consideration has important computational implications.

The connection of the lasso to boosting and forward stagewise regression fitting has been studied in the literature (Efron *et al.*, 2004; Rosset *et al.*, 2004; Zhao & Yu, 2007; Hastie *et al.*, 2007). Boosting regression with a squared loss (using infinitesimal steps) and the lasso produce arbitrarily close regularization paths when the positive cone condition is satisfied (Efron *et al.*, 2004). Loosely speaking, this occurs when the correlation among the input covariates is low. When the correlation is high, the solution computed by boosting and the lasso are typically similar only at early stages of the regularization paths. Still, boosting and forward stagewise

regression are attractive, easy-to-implement alternatives when the direct computation of the L_1-penalized regression is hard, which is often the case for general convex functions. Note that, for parametric least squares problems, there exist very efficient algorithms that exactly recover the regularization path (see previous text) and boosting may be less interesting.

Zhao & Yu (2007) went one step further by introducing a stagewise algorithm that, by considering also backwards steps, computes the exact regularization path for L_1-penalized convex loss problems and a finite number of base learners under some conditions (weaker than the positive cone condition). In particular, only strong convexity of the loss function and bounded second derivatives are required. similar to classical boosting methods, the so-called *blasso* does not need to compute any derivatives. Zhao & Yu (2007) empirically showed that if the underlying model is sparse, the obtained solutions are sparser and slightly more accurate in terms of prediction than regular boosting for regression. This way, the lasso idea can be generalized to any convex loss function, and computation remains efficient. Note that, for non-differentiable loss functions, the algorithm can be stuck in a non-differentiable point. Of course, for non-convex loss functions, the blasso is not guaranteed to find the optimal solution.

3 Improving the Lasso's Properties

Several modifications have been introduced to the lasso to either improve its statistical properties or to adapt to a specific problem configuration. This section deals with the former, whereas Section 4 is devoted to the latter. We focus on methods that still use an L_1 penalty, omitting methods such as *Smoothly Clipped Absolute Deviation* (Fan & Li, 2001) that change the type of penalty.

Any form of regularization introduces a bias in the estimation in exchange for a (hopefully) larger reduction in variance. In addition, when the number of true nonzero coefficients is small relative to p, the lasso introduces considerable bias in the correct variable coefficients while discarding the irrelevant variables,. To minimize the bias, the *relaxed lasso* (Meinshausen, 2007) introduced two-stage estimation. First, the lasso with an appropriately tuned regularization parameter is used to discover the sparsity pattern. Then it is used with a lower regularization parameter to estimate the coefficients on the selected variables. Hence, we have two regularization parameters, one per stage, that need to be estimated, for example by cross-validation. A similar idea using ordinary least squares in the second phase was already proposed by Efron *et al.* (2004). This is, however, possible only when $N \geq p$.

In the spirit of the relaxed lasso, the *Variable Inclusion and Shrinkage Algorithm* (VISA) (Radchenko & James, 2008) performs a two-stage estimation with two different regularization parameters. In the second stage, the VISA approach does not definitely discard the variables dropped in the first stage, although it gives a higher priority to the previously selected variables. The paper by Radchenko & James (2008) provides a convincing theoretical justification of the method. Although it is slightly more complicated, the VISA method can outperform the relaxed lasso.

The *adaptive lasso* (Zou, 2006) penalizes each variable according to its importance. The adaptive lasso estimator is

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where, in the $N \geq p$ case, the weights $\boldsymbol{w} = (w_1, \ldots, w_p)$ can be computed as $w_j = 1/\left|\hat{\beta}_j^{ls}\right|^{\gamma}$ $(\gamma > 0)$. If $N < p$, the weights can be computed by using estimates with minimal

regularization instead of $\hat{\boldsymbol{\beta}}^{ls}$. Zou (2006) shows that the adaptive lasso properties are superior to those of the lasso, specially for variable selection purposes. In particular, the adaptive lasso meets the so-called oracle properties Fan & Li (2001): (i) identify the true sparsity pattern; and (ii) have an optimal estimation rate of the coefficients. It is fair to mention, however, that the adaptive lasso may suffer more for strong collinearity than the lasso itself.

A simple variation of this scheme is to use cross-validation or some C_p-like statistic for selecting λ in the first stage, that is, for computing $\boldsymbol{w}$. Variables that have been discarded now will not be used in the second phase. Hence, sparser models are eventually obtained and accuracy typically does not decrease much. One can even repeat this procedure a number of times so that the set of candidate covariates is (probably) reduced at each step, giving rise to even sparser models.

The adaptive lasso can be considered a convex approximation of the L_q penalties, with $0 < q < 1$, which have been proved to have the oracle properties (Knight & Fu, 2000). Further theoretical analysis of the adaptive lasso is performed by Huang *et al.* (2008) and also by Pötscher & Schneider (2009), who study the distribution of the adaptive lasso estimator.

The LARS algorithm can be used to compute the adaptive lasso regularization path. Figure 4 shows the adaptive lasso regularization paths for the *Diabetes* data set, using ordinary least squares (left) and lasso with cross-validation (right) for computing $\boldsymbol{w}$; in the latter case, we are entirely pruning one covariate from the entire regularization path. Note the differences with Figure 3.

Alternatively to the lasso, the *Dantzig selector* (Candès & Tao, 2007) substitutes the sum of squared error in Equation (2) by an L_∞ norm, that is, the maximum absolute value of the components of the argument. Thus, the Dantzig selector yields

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \, ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_\infty + \lambda ||\boldsymbol{\beta}||_1.$$

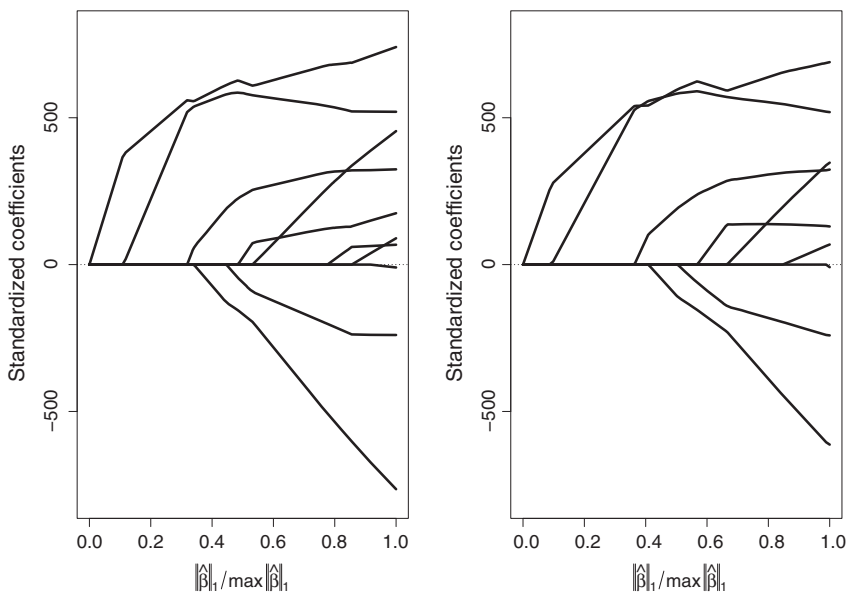

Figure 4. *Regularization paths of adaptive lasso for the Diabetes data set, where* $\boldsymbol{w}$ *is computed by ordinary least squares (left) and the lasso with cross-validation (right).*

The Dantzig selector shares some statistical properties with the lasso, particularly regarding the recovery of the true sparsity pattern. However, the Dantzig selector estimator might be less stable than the lasso. Bickel *et al.* (2009) examined the theoretical properties of the Dantzig selector in comparison with the lasso. James *et al.* (2009) proposed an algorithm to find the entire regularization path for the Dantzig selector.

To achieve further robustness, Wang *et al.* (2007a) propose the *LAD-lasso*, whose loss of function is regarding the L_1 norm:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_1 + \lambda||\boldsymbol{\beta}||_1.$$

As mentioned earlier, when the data set contains strong correlations among the predictors, the ridge's prediction performance is better than the lasso's. Motivated by this, Zou & Hastie (2005) propose the *elastic net*, a popular method that mixes the lasso and the ridge penalties:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \left(\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2\right), \tag{5}$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$. Ridge regression and lasso regression are particular cases corresponding, respectively, to $\alpha = 0$ and $\alpha = 1$.

Besides performing better in the presence of correlated predictors (even for α very close to 1), the elastic net has other interesting properties. For the $p > N$ case, in particular, whereas the lasso can select at most N predictors (then the solution saturates), the elastic net can include more than N predictors into the model. Furthermore, assuming that there are some group of relevant and redundant variables, the lasso tends to discard all but one from this group. The elastic net instead selects the entire group of redundant variables, giving a balanced estimation of their coefficients. We often want to keep all redundant (if relevant) variables for interpretation's sake. In microarray analysis, for example, one usually wants to identify all the genes involved in a particular process, even when their expression levels are very similar. Bunea (2008) analyzes the variable selection capability of the elastic net for generalized linear models (see Section 5) compared with the lasso.

Figure 5 shows the elastic net regularization path (computed with the LARS-EN algorithm, devised by Zou & Hastie (2005)) for the *Diabetes* data set and two different values of α; note the differences from Figure 3. The regularization path (for a given α) can be computed by a version of LARS (Zou & Hastie, 2005) or by pathwise coordinate descent optimization. Li & Li (2008) modify the elastic net penalty to accommodate prior biological knowledge. Lorbert *et al.* (2010) extend the elastic net to encourage similar variables to have similar coefficients, improving the interpretability of the model.

Finally, some authors, such as Bach (2008a) or Chatterjee & Lahiri (2011), have considered subsampling to improve stability and model selection accuracy. In particular, the so-called *bolasso* uses bootstrapping. In terms of stability, a related approach is the *randomized lasso* (Meinshausen & Bühlmann, 2010) which, similar to the adaptive lasso, uses weights within the penalty. In this case, the weights are randomly generated (for example, uniformly within the range $(0.2, 0.8)$), and this procedure is repeated a number of times by using any subsampling method. The randomized lasso has interesting theoretical properties with regard to stability selection. Wang *et al.* (2011) present a bit more involved methodology where bootstrapping is performed in two stages. In both stages, the lasso is repeated a number of times by using only a selected subset of the covariates. Whereas in the first stage, this selection is purely random, in the second stage, this is influenced by the results of the first stage. This approach can perform better than the elastic net for selecting groups of (very) correlated variables in a coherent manner. A representative example of this scenario is when there are variables strongly correlated

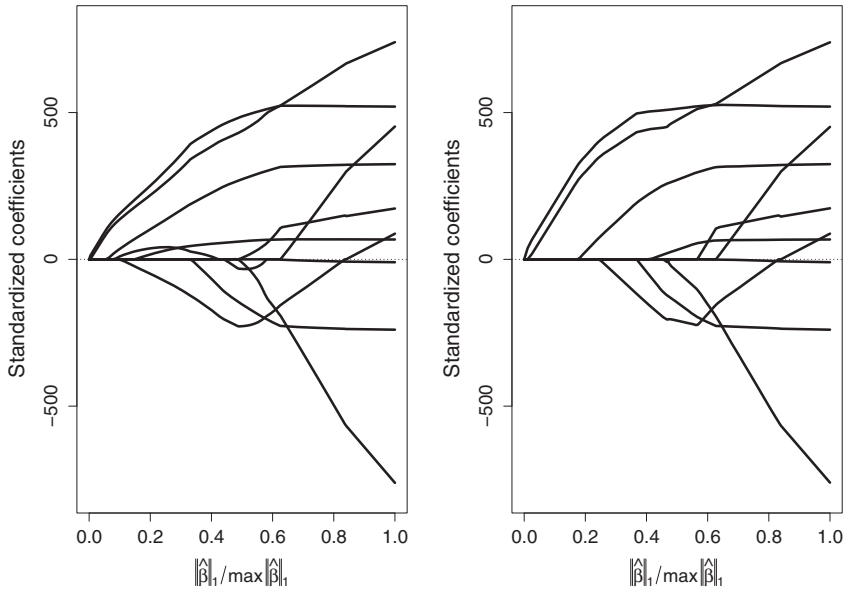

Figure 5. *Regularization path of the elastic net for the Diabetes data set for* $\alpha = 0.2$ *(left) and* $\alpha = 0.8$ *(right).*

that have different signs. All these algorithms trade stability and improved variable selection in exchange of computational cost.

4 Adapting the Lasso to Particular Problems

Until now, we have described a number of methods that either improve the properties of the lasso or attempt to give an alternative. In the following, we describe several modifications of the lasso for tailoring to particular problem settings. Note that the elastic net could also be included in this section.

In some situations, the variables are naturally grouped and we are interested in including only entire groups in the model. For this setting, Yuan & Lin (2006) propose the *group lasso* that is defined as follows:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{\boldsymbol{W}_j}, \tag{6}$$

where the set of variables is partitioned into J groups and $\boldsymbol{\beta}_j$ are the parameters of the j-th group. The penalty is defined as $||\boldsymbol{\beta}_j||_{\boldsymbol{W}_j} = \left(\boldsymbol{\beta}_j' \boldsymbol{W}_j \boldsymbol{\beta}_j\right)^{1/2}$, where $\boldsymbol{W}_j$ is typically chosen to be the identity matrix. This penalty can be considered a generalization of the L_2 penalty. A trivial application of the group lasso is the case of categorical variables, where each level is codified as a set of indicator dummy variables. Yuan & Lin (2006) also introduce an efficient LARS-type algorithm to approximate the group lasso solution, giving an exact solution when the design matrix $\boldsymbol{X}$ is orthogonal. Note that it is still possible to use a pathwise coordinate descent method to ascertain the exact regularization path in the general case.

The consistency of the group lasso estimator was proved by Bach (2008b) under some assumptions. Wang & Leng (2008) proposed an adaptive version of the group lasso. Jacob

et al. (2009) extended the group lasso for overlapping groups. Huang *et al.* (2009) introduced the so-called *group bridge* approach, which, besides entire groups, can discard individual variables within the groups.

The *Composite Absolute Penalties* (CAP) approach (Zhao *et al.*, 2009) generalizes the group lasso by using a specific L_{γ_j} penalty for each group, plus some L_{γ_0} penalty to combine the groups:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{J} \left(||\boldsymbol{\beta}_j||_{\gamma_j}\right)^{\gamma_0}. \tag{7}$$

Different objectives can be pursued by this flexible approach. For example, when L_∞ penalties are used for each group, the coefficients within each group are driven towards equality. Also, the CAP approach can account for hierarchical relationships between the predictors by defining groups with specific overlapping patterns. Zhao *et al.* (2009) develop an algorithm to compute the entire regularization path for a CAP problem with $\gamma_0 = 1$ and $\gamma_j = \infty$ for all j.

In other problems, variables are ordered significantly and (spatially) close variables should have similar coefficients. The *fused lasso* (Tibshirani *et al.*, 2005) penalizes both the coefficients and the difference between adjacent coefficients:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

with $\lambda_1, \lambda_2 \geq 0$. The fused lasso is motivated by the problem of analyzing protein mass spectroscopy data, where spatially closer variables (sites) are known to be jointly relevant or irrelevant. The solution of the fused lasso problem can be obtained, for example, by pathwise coordinate optimization.

Finally, the LARS/lasso approach can also be extended for multivariate (multiresponse) regression. Similä & Tikka (2006) propose an extension of the LARS algorithm by modifying the correlation criterion between the predictors and the current residual (which depends on multiple outputs). Unfortunately, the exact regularization path can be recovered only when $\boldsymbol{X}$ is orthonormal. Similä & Tikka (2007) suggest a useful approximation for the general case. A related algorithm is proposed by Vidaurre *et al.* (2013).

5 Generalized Linear Models

In this section, we extend the discussion on L_1-penalized regression to a more general setting, referred to as the generalized linear models (GLMs) framework (McCullagh & Nelder, 1989). Again, we have variables $\{X_1, \ldots, X_p, Y\}$, where Y is the response variable. We also have a data set $\boldsymbol{D} = \{(x_{i1}, \ldots, x_{ip}, y_i), i = 1, \ldots, N\}$. Now, the linear contribution of the inputs defines the expectation of the output through some *link function* $g(\cdot)$,

$$g(E[Y|\boldsymbol{x}]) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta},$$

so that the linear case with squared loss corresponds to the special case where $g(\cdot)$ is the identity function. The inverse of the link function is usually called the *mean function*.

Depending on the form of $g(\cdot)$ and the nature of Y, we obtain different flavours of GLM. Typically, the distribution of Y belongs to the exponential family distribution, which includes the Gaussian, the Bernoulli, and the Poisson distributions. Also, $g(\cdot)$ is chosen to be monotonic and differentiable.

Probably, the most popular member of the family of GLMs is the logistic regression model, where Y is categorical (Bernoulli distributed in the two-class, simplest classification problem), and the link function corresponds to the logit function, $g(\rho) = \log \frac{\rho}{1-\rho}$. Although the logit function is the canonical link function for the Bernoulli distribution, other link functions are also applicable to Bernoulli distributed data, such as the probit function, the log-log function, and the complementary log-log function. These are not discussed in this paper.

Next, we deal with some relevant contributions on L_1-penalized logistic regression, including the case when Y can take more than two values. We also discuss the use of the L_1 penalty in another two popular GLMs: the Poisson regression model and the Cox's proportional hazards model. Other examples of GLMs, which are left out for space constraints, are gamma regression and inverse-Gaussian regression, both used for modeling positive continuous data. In all cases, the linear configuration can be straightforwardly extended for nonlinearity.

Most of the theoretical analysis of the lasso for linear regression is valid for GLM regression. The algorithms to solve the resulting optimization problem, however, might be sometimes a bit more involved.

5.1 Logistic Regression

Logistic regression aims to model the posterior probability of the response or class variable Y, $Pr(Y = k|\boldsymbol{x})$ as a transformation of a linear combination of the inputs. For a K-classes problem, $K-1$ logit functions are defined as

$$\log \frac{Pr(Y = k|\boldsymbol{x})}{Pr(Y = K|\boldsymbol{x})} = \beta_0^{(k)} + \boldsymbol{x}'\boldsymbol{\beta}^{(k)}, \quad k \in \{1, \ldots, K-1\}, \tag{8}$$

where $\left\{\beta_0^{(k)}, \boldsymbol{\beta}^{(k)}\right\}$ are the logistic linear regression parameters for the k-th class value. The denominator is set to be the K-th class value, but it could be any. This yields

$$Pr(Y = k|\boldsymbol{x}) = \frac{e^{\left(\beta_0^{(k)} + \boldsymbol{x}'\boldsymbol{\beta}^{(k)}\right)}}{1 + \sum_{l=1}^{K-1} e^{\left(\beta_0^{(l)} + \boldsymbol{x}'\boldsymbol{\beta}^{(l)}\right)}}, k \in \{1, \ldots, K-1\},$$

and, hence, $Pr(Y = K|\boldsymbol{x}) = 1 - \sum_{k=1}^{K-1} Pr(Y = k|\boldsymbol{x})$. The model, including the intercepts, has $(p+1)(K-1)$ parameters overall, and the maximum likelihood solution can be found by the iteratively reweighted least squares algorithm (IRLS), derived from Newton's method. Note that the inputs can be either the original set of predictors or some expansion thereof, so that nonlinear decision boundaries can be achieved. For example, Park & Hastie (2008) include two-level interactions and regularize with an L_2 penalty (no L_1 penalty is considered, though).

The L_1 penalty for logistic regression was first mentioned by Tibshirani (1996). He formulates the binary classification problem with continuous predictors as the minimization of the L_1-penalized negative log-likelihood function. This function is

$$-\sum_{i=1}^{N} \left(y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log\left(1 + e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)\right) + \lambda ||\boldsymbol{\beta}||_1, \tag{9}$$

where the input $\boldsymbol{x}_i$ includes the constant term 1 to integrate the intercept. This problem is solved by applying the original lasso algorithm at each step of the IRLS algorithm. However, the convergence of this method is not guaranteed, and it is not computationally efficient for large-dimensional problems.

For continuous predictors and a binary response, a number of contributions relate the lasso to logistic regression. For example, Roth (2004) adapts the algorithm proposed by Osborne *et al.* (2000), which solves the lasso, to Equation (9), showing the global convergence of the algorithm. Shevade & Keerthi (2003) devise a simple and easy way to implement the algorithm for the same task. Genkin *et al.* (2007) tackle the same problem in a Bayesian context. Also from a Bayesian perspective, Balakrishnan & Madigan (2008) propose online algorithms to fit an L_1-regularized logistic regression model, so that the entire data set does not have to be stored in memory; van Gerven *et al.* (2010) reformulate the Laplace prior on $\boldsymbol{\beta}$ as a scale mixture to force similarity between coefficients of nearby variables; Cawley & Talbot (2006) analytically integrate out the regularization parameter so that computations are accelerated.

An essential milestone for the applicability of logistic regression to high-dimensional data is the recent emergence of efficient methods for computing the entire regularization path of GLMs with convex loss functions. Park & Hastie (2007) develop an efficient regularization path-following algorithm for GLMs based on predictor–corrector methods of convex optimization. Alternatively, Friedman *et al.* (2010) present an extremely efficient coordinate descent method for computing the GLM regression coefficients on a grid of λ parameter values for elastic net penalties. Shi *et al.* (2010) give a comprehensive list of state-of-the-art algorithms for the sparse logistic regression problem. In addition, they propose an algorithm comprising two stages: a fast iterative shrinkage phase and an accurate interior point phase.

Meier *et al.* (2008) adapt the group lasso (Yuan & Lin, 2006), defined in Equation (6), to the binary logistic regression model. The so-called *logistic group lasso* allows for categorical predictors by modeling each categorical predictor as a group of dummy variables. The logistic group lasso then aims to minimize the group L_1-penalized negative log-likelihood function

$$-\sum_{i=1}^{N}\left(y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log\left(1 + e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)\right) + \lambda \sum_{j=1}^{p} w_j ||\boldsymbol{\beta}_j||_2.$$

If the j-th predictor is categorical, $\boldsymbol{\beta}_j$ are the parameters for the set of dummy variables. If the j-th predictor is continuous, $\boldsymbol{\beta}_j$ has only one component. The weights w_j scale the penalty with regard to the dimensionality of $\boldsymbol{\beta}_j$. Meier *et al.* (2008) devise an efficient algorithm based on pathwise coordinate optimization and prove that the resulting estimator is statistically consistent.

Of course, the adaptive lasso idea is useful for logistic regression in particular and GLMs in general, so that the resulting models enjoy the improved theoretical properties of the adaptive penalty under certain mild conditions (Zou, 2006).

There are considerable additional work on L_1-regularized multinomial logistic regression. From a Bayesian perspective, Krishnapuram *et al.* (2005) introduce a new method for sparse multinomial logistic regression, finding the MAP for the formulation in Equation (9) with a Laplacian prior distribution on the parameters. Although a Gaussian prior (which is equivalent to an L_2 penalty) is easily accommodated into the IRLS algorithm, IRLS cannot handle the Laplacian prior. To estimate the $(p + 1)(K - 1)$ regression parameters, Krishnapuram *et al.* (2005) introduce a bound optimization approach with a computational cost equivalent to IRLS.

In situations where the log-likelihood function is not well-behaved and IRLS is not guaranteed to converge, Tian *et al.* (2008) propose a quadratic lower-bound algorithm to solve the binary L_1-regularized logistic regression, also applicable to the multinomial problem. Cawley *et al.* (2007) extend their previous results for a binary response (Cawley & Talbot, 2006) to the multinomial case.

5.2 Poisson Regression

In the Poisson regression setting, the response variable Y can take a positive integer value or zero. Therefore, Poisson regression is useful for modeling count data and contingency tables. The link function $g(\cdot)$ is taken to be the natural logarithm, so that

$$\log(Y|\boldsymbol{x}) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}.$$

Thanks to the log link function, the exponentiated regression coefficients $\exp(\beta_j)$ play the role of multiplicative effects on the response.

If we apply the L_1 penalty to the negative log-likelihood function derived from this configuration, we obtain, for some regularization parameter λ, the following objective function

$$\sum_{i=1}^{N} \left(-y_i \left(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}\right) + \exp\left(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) + \lambda||\boldsymbol{\beta}||_1, \tag{10}$$

which, for the properties of the lasso estimator, leads to a sparse estimate of $\boldsymbol{\beta}$. Because the loss of function is convex, criterion (10) can be efficiently optimized by a pathwise coordinate descent algorithm. The elastic net and adaptive penalties can be used within the Poisson regression setting with no further complication.

5.3 Cox Proportional-hazards Regression

The Cox's proportional hazards model is used for modeling survival data. In this case, we have a data set $\boldsymbol{D} = \{(x_{i1}, \ldots, x_{ip}, y_i, \eta_i), i = 1, \ldots, N\}$, where $\eta_i = 1$ if the subject has passed away and $\eta_i = 0$ if y_i is simply the censoring time. Of course, this formulation can be used for different semantics. Assuming centered data, let us define the hazard function $h(\cdot)$ as

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{11}$$

which gives the hazard at time t for a subject with covariate vector $\boldsymbol{x}$ and where $h_0(t)$ is some baseline hazard function. From Equation (11), it follows the negative log partial likelihood

$$-\sum_{i_1:\eta_{i_1}=1}^{N} \left(\boldsymbol{x}_{i_1}'\boldsymbol{\beta} + \log \sum_{i_2:y_{i_2}\leq y_{i_1}} \exp\left(\boldsymbol{x}_{i_2}'\boldsymbol{\beta}\right)\right).$$

Tibshirani (1997) applies the lasso for the Cox model by imposing a constraint $||\boldsymbol{\beta}|| \leq s$ (or, alternatively, a Lagrange multiplier $\lambda||\boldsymbol{\beta}||$), which, for the properties of the L_1 penalty, drives some coefficients β_j to exactly zero (depending on the value of s). The L_1-penalized Cox model minimization problem can be approached by coordinate descent optimization, which can also deal with the elastic net and adaptive penalties, whose properties will be transferred to the Cox model estimate. Some further discussion about sparse survival data analysis is given, for example, by Porzelius *et al.* (2010).

6 Time Series

This section deals with models for forecasting the evolution and describing the behavior of a variable or set of variables across time. The variables of interest are typically measured at equally-spaced time points. The discretization of an analog signal, sampled over a finite set of intervals, is also a time series. Time series can be modeled considering that the variables of interest depend only on previous data points or include exogenous (external) inputs.

6.1 Wavelet Analysis

Before the lasso for linear regression (Tibshirani, 1996), the L_1 penalty was already successfully used in the context of signal analysis with wavelets; see for example (Donoho & Johnstone, 1994; Antoniadis & Fan, 2001). In contrast to Fourier analysis, which establishes a frequency representation of an analog signal, wavelet theory (Vidakovic, 1999) uses a time–frequency representation. This is very handy for non-stationary signal analysis. Wavelets are used in various statistical contexts.

A wavelet is a function representing an oscillation. Wavelets provide a complete orthonormal basis of an infinite-dimensional space where square-integrable functions can be represented. This orthonormal basis is composed of the infinite set of dilations and integer translations of the so-called mother wavelet and the infinite set of integer translations of some scaling function. The resulting space is structured at various levels, each linked to a given level of detail. This allows for what is called multiresolution analysis, where each basis function is assigned a wavelet coefficient. Because it is computationally impossible to analyze a signal using all wavelet coefficients (and discretized signals convey only finite information), multiresolution analysis can only be performed to a certain level of detail.

In this setting, for equally spaced points, the wavelet coefficients $\boldsymbol{z}$ can be computed by using least squares. Assuming that we have $\boldsymbol{y} = (y_1, \ldots, y_N)$ equidistantly sampled at N lattice-points, the *discrete wavelet transform* is defined as

$$\boldsymbol{z} = \boldsymbol{W}'\boldsymbol{y},$$

where $\boldsymbol{W}'$ is the inverse (transpose) of the $N \times N$ orthonormal basis matrix $\boldsymbol{W}$. Then, $\boldsymbol{W}'$ encloses the (linear) wavelet transform of $\boldsymbol{y}$. Donoho & Johnstone (1994) propose the following penalized estimation of $\boldsymbol{z}$:

$$\hat{\boldsymbol{z}} = \text{argmin}_{\boldsymbol{z}} ||\boldsymbol{y} - \boldsymbol{W}\boldsymbol{z}||_2^2 + 2\lambda||\boldsymbol{z}||_1. \tag{12}$$

Because $\boldsymbol{W}$ is orthonormal, $\boldsymbol{z}$ is efficiently computed by soft thresholding:

$$\hat{z}_t = \text{sign}\left(\hat{z_t}^{ls}\right)(\tilde{z_t} - \lambda)_+, \quad t \in \{1, \ldots, N\},$$

where $\hat{z_t}^{ls}$ is the least squares estimator for z_t. As $||\boldsymbol{z}||_1$ in Equation (12) is an L_1-regularization term, the estimated vector of coefficients turns out to be sparse. This method is known as *Stein Unbiased Estimate of Risk* shrinkage. Antoniadis & Fan (2001) introduce a more general approach for not equally spaced points, called *regularized one-step estimator*.

On the other hand, the *basis pursuit* approach wherein Chen *et al.* (1998) decomposes a given signal into an optimal combination of dictionary elements (such as, for example, wavelets) from an overcomplete waveform dictionary so that the coefficients of the chosen decomposition have the smallest L_1 norm. Basis pursuit is the counterpart of the lasso in the signal-processing literature.

6.2 Autoregressive Models

The L_1 penalty has also been employed in the context of autoregressive models. Assuming centered data, an m-order autoregressive model is a type of random process that imposes a linear relation

$$y_t = \sum_{i=1}^{m} \beta_i y_{t-i} + \epsilon_t, \quad t \in \{m+1, \ldots, N\},$$

where ϵ_t is Gaussian white noise. This can be easily extended to the multivariate case, yielding the multivariate autoregressive (MAR) model:

$$\boldsymbol{y}_t = \sum_{i=1}^{m} \boldsymbol{B}^{(i)} \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t, \quad t \in \{m+1, \ldots, N\},$$

where $\boldsymbol{y}_t \in \mathbb{R}^p$, $\boldsymbol{B}^{(i)}$ is a $p \times p$ coefficient matrix defined for each time point i, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix.

Valdés-Sosa *et al.* (2005) apply an L_1 penalty to an $m = 1$ order MAR model to achieve sparsity so that the MAR coefficients estimators are computed as

$$\hat{\boldsymbol{B}} = \text{argmin}_{\boldsymbol{B}} \sum_{t=m+1}^{N} ||\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}||_2^2 + \lambda ||\boldsymbol{B}||_1,$$

where the matrix norm $||\cdot||_1$ is the sum of the absolute values of the argument. Hsu *et al.* (2008) apply the L_1 penalty to the general, higher-order MAR model, demonstrating the asymptotics of the estimator such as Knight & Fu (2000) do for the lasso.

Haufe *et al.* (2008) propose a group lasso penalty Yuan & Lin (2006) for the MAR model

$$\begin{aligned}\hat{\boldsymbol{B}} =& \text{argmin}_{\boldsymbol{B}} \sum_{t=m+1}^{N} \sum_{i=1}^{m} ||\boldsymbol{y}_t - \boldsymbol{B}^{(i)} \boldsymbol{y}_{t-i}||_2^2 \\ &+ \lambda \left[|| \text{vec} \left(B_{11}^{(1)}, ..., B_{pp}^{(M)} \right) ||_2 \right. \\ &\left. + \sum_{i_2 < i_1} || \text{vec} \left(B_{i_1 i_2}^{(1)}, ..., B_{i_1 i_2}^{(M)} \right) ||_2 \right],\end{aligned}$$

where $\boldsymbol{B} = \left\{ \boldsymbol{B}^{(1)}, ..., \boldsymbol{B}^{(m)} \right\}$ and vec$(\cdot)$ is the vectorization operation. On the one hand, it penalizes all the squared coefficients to prevent overfitting. On the other hand, it penalizes the subgroups of coefficients that belong to each pair of variables, so that the coefficients will jointly become zero unless they are causally related.

The regression model with autoregressive errors is a linear model where the error follows an autoregressive process:

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \epsilon_t,$$

$$\epsilon_t = \sum_{i=1}^{m} \alpha_i \epsilon_{t-i} + e_t, \quad t \in \{m+1, ..., N\},$$

where $\boldsymbol{x}_t \in \mathbb{R}^p$, $\boldsymbol{\beta}$ is the p-vector of regression coefficients, $\boldsymbol{\alpha} \in \mathbb{R}^m$ is the vector of error autoregression coefficients and e_t is independent and identically distributed $\mathcal{N}\left(0, \sigma^2\right)$ distributed.

Wang *et al.* (2007b) propose to impose an L_1 penalty on both the regression coefficients $\boldsymbol{\beta}$ and the autoregressive coefficients $\boldsymbol{\alpha}$. By using a different and appropriate regularization parameter for each coefficient, we found its statistical properties improve; see also Zou (2006).

6.3 Other Regression Models

Furthermore, there are some L_1-regularized approaches based on the usual linear regression model for accommodating time-course data. They consider that data instances close in time should share common properties. Let us consider the following model:

$$\boldsymbol{y}(t) = \boldsymbol{X}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \quad t \in \{1, ..., N\},$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ is the design matrix, $\boldsymbol{y}(t) \in \mathbb{R}^N$ is the response vector, $\boldsymbol{\beta}(t) \in \mathbb{R}^p$ and $\boldsymbol{\epsilon}(t)$ is the independent and identically distributed error vector whose components are $\mathcal{N}\left(0, \sigma^2\right)$ distributed. Hence, we have a linear regression at each time point. A generalization of this formulation could make $\boldsymbol{X}$ to be also dependent on t.

The fused lasso Tibshirani *et al.* (2005) described in Section 4 could be used here by penalizing the difference of the coefficients of contiguous time points $|\beta_j(t) - \beta_j(t-1)|$. However, the resulting model has a total of Np parameters, because all the time points have to be simultaneously modeled. To solve this problem at each time point t, the *smoothed lasso* Meier & Bühlmann (2007) estimates $\boldsymbol{\beta}(t)$ as the minimizer of

$$\sum_{s=1}^{N} w_\tau(s,t)\, ||\boldsymbol{y}(s) - \boldsymbol{X}\boldsymbol{\beta}(t)||_2^2 + \lambda ||\boldsymbol{\beta}(t)||_1,$$

where the weights $w_\tau(s,t)$ depend on $|t-s|$ and a bandwidth parameter τ. The weights are chosen so that $\sum_{s=1}^{N} w_\tau(s,t) = 1$. This formulation intends to achieve a parsimonious model that is smooth on the time scale, combining information across different time points. Besides the pure L_1 penalty, an adaptive penalty Zou (2006) is also tried.

The smoothed lasso can solve this problem for high-dimensional settings and allows for bandwidth selection (by, for example, cross-validation).

In the binary classification arena, Balakrishnan & Madigan (2007) use a combination of the fused lasso Tibshirani *et al.* (2005) and the group lasso Yuan & Lin (2006) to find contiguous (given some ordering) groups of variables with high predictive power. This approach, which can be applied in the temporal domain, is called *Lasso with Attribute Partition Search*.

6.4 Change Point Analysis

Change point analysis accounts for the detection of significant changes in a signal whose underlying generating process is assumed to be piecewise constant plus some white noise. This problem arises, for example, in neuroscience electroencephalography segmentation. The model can be formulated as

$$\begin{aligned} y_t = \mu_k + \epsilon_t, \quad & t \in \{1, ..., N\} \\ \text{s.t.} \quad & \delta_{k-1} \leq t \leq \delta_k, \end{aligned}$$

where ϵ_t is independent and identically distributed $\mathcal{N}\left(0, \sigma^2\right)$ and (μ_k, δ_k), $k \in \{1, ..., K\}$ are parameters to be estimated. We assume $\delta_0 = 0$. Also, K needs to be estimated if no prior information is available. This problem can also be formulated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{X}$ is an $N \times N$ lower triangular matrix with nonzero elements equal to one, $\boldsymbol{\beta} \in \mathbb{R}^N$ is a vector with all elements equal to zero except for those corresponding to the

change point instants and ϵ_t is independent and identically distributed $\mathcal{N}\left(0, \sigma^2\right)$. Harchaoui & Lévy-Leduc (2008) estimate $\boldsymbol{\beta}$ by the minimization of

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

which is equivalent to estimating the extended vector $\boldsymbol{\mu}^* \in \mathbb{R}^N$, whose elements correspond to elements of $\boldsymbol{\mu}$, by

$$||\boldsymbol{y} - \boldsymbol{\mu}^*||_2^2 + \lambda \sum_{t=1}^{N} |\mu_{t+1}^* - \mu_t^*|.$$

7 Discussion

The lasso, that is, the minimization of the L_1-penalized negative log-likelihood function, is supported by well-grounded theoretical analysis. A number of penalties based on the L_1 penalty have been proposed for adaptation to specific types of problems or improvement of the statistical properties. This methodology and its interesting properties is of paramount importance in a wide variety of problems. In particular, L_1 regularization usually enriches the models with variable selection and a reasonable bias-variance trade off.

In this paper, we have described the basics about L_1-penalized linear regression, providing some insight on model selection, theoretical properties, regularization paths, computational algorithms, and the connections to the popular boosting framework. We have also presented the lasso from a Bayesian perspective, giving a brief discussion about advantages and drawbacks.

Extending the Gaussian-distributed response case, we have discussed some representative generalized linear models, such as logistic regression, Poisson regression and the Cox's proportional hazards model. Most of the theory and considerations of the lasso for linear regression apply to these models, as well as the methodology for nonlinear modeling. Generalized linear models usually have computational particularities that need to be addressed, although the (re)discovery of pathwise coordinate optimization has considerably eased this point.

Finally, a lot has also been written about time series. We have briefly introduced some selected research on L_1-regularized wavelet analysis. Also, we have discussed L_1-regularized autoregressive models (and related models) and change point analysis. We have omitted regularized Fourier analysis, for example, which has also interesting developments; see for example Yang *et al.* (2010).

Although not discussed in this paper, the idea of linear L_1-regularized regression can be extended to cope with nonlinearity. Two approaches are possible: we can either use complex models that fit the entire data domain better or we can fit simpler (linear) models for different areas of the data domain. The former approach works with a dictionary of functions, which contains a potentially very large set of elements. These functions are transformations of the original input terms and can involve one Ravikumar *et al.* (2009) or more input terms Lin & Zhang (2006) & Radchenko & James (2010). These ideas can be implemented within the generalized linear model framework Zhang *et al.* (2004) & Ravikumar *et al.* (2009). The second approach uses only the closest data to build a (typically) linear model for each point of interest. Here, the cornerstone is the selection of the bandwidth that parametrizes the function of distance between points. A few methods for local estimation are related to the lasso, for example, Lafferty & Wasserman (2008). These methods, however, rely on heuristics to a greater or lesser extent.

Important advances are being made towards the extension of the theoretical foundations to many of these variants and extensions. However, as noted by Hesterberg *et al.* (2008), much

work still remains to be carried out. Although considerable progress has been made since Hesterberg *et al.*'s review, further investigation of the properties of the proposed models would still be of great interest.

Another promising line of research is on techniques for making rigorous inference with lasso-related methods. Whereas this is straightforward for the Bayesian approximation, the frequentist lasso and extensions need a careful treatment in this sense. Bootstrapping and other subsampling techniques, for example, are a possibility but are not free of problems. Finally, the Bayesian formulation of some of the extensions of the lasso can be challenging. There are already some relevant work in this sense; see for example Li & Lin (2010) and Kyung *et al.* (2010).

Some applications of L_1 regularization that, albeit interesting, are omitted from this review include quantile regression Li & Zhu (2008), matrix completion Candès & Tao (2010), compressed sensing Donoho (2006), sparse canonical correlation analysis Hardoon & Shawe-Taylor (2011), or sparse coding Sprechmann *et al.* (2010) and Lee *et al.* (2007), among others.

In a future review paper, we plan to cover the use of L_1 regularization in relevant machine learning applications, such as supervised classification, cluster analysis, graphical models, and feature extraction techniques.

Acknowledgements

This research was partially supported by the Spanish Ministry of Science and Innovation, projects TIN2010-20900-C04-04 and Cajal Blue Brain.

References

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Stat. Soc. Ser. B*, **36**(1), 99–102.

Antoniadis, A. & Fan, J. (2001). Regularization of Wavelets Approximations. *J. Amer. Statist. Assoc.*, **96**(455), 939–955.

Bach, F. R. (2008a). Bolasso: Model Consistent Lasso Estimation Through the Bootstrap. 15th International Conference on Machine Learning; 33–40.

Bach, F. R. (2008b). Consistency of the Group Lasso and Multiple Kernel Learning. *J. Machine Learn. Res.*, **9**(1), 1179–1225.

Balakrishnan, S. & Madigan, D. (2007). Finding Predictive Runs with LAPS. IEEE International Conference on Data Mining; 415–420.

Balakrishnan, S. & Madigan, D. (2008). Algorithms for Sparse Linear Classifiers in the Massive Data Setting. *J. Machine Learn. Res.*, **9**(1), 313–337.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

Bickel, P. J. & Li, B. (2006). Regularization in Statistics. *TEST*, **15**(2), 271–344.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *Ann. Statist.*, **37**(4), 1705–1732.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-dimensional Data.* Berlin-Heidelberg: Springer.

Bunea, F. (2008). Honest Variable Selection in Linear and Logistic Regression models via l_1 and $l_1 + l_2$ penalization. *Electron. J. Statist.*, **2**(1), 1935–7524.

Bunea, F., Tsybakov, A. & Wegkamp, M. (2007). Sparsity Oracle Inequalities for the Lasso. *Electron. J. Statist.*, **1**, 169–194.

Candès, E. J. & Tao, T. (2007). The Dantzig Selector: Statistical Estimation when p is Much Larger than n. *J. Roy. Stat. Soc. Ser. B*, **35**(6), 2313–2351.

Candès, E. J. & Tao, T. (2010). The Power of Convex Relaxation: Near-optimal Matrix Completion. *IEEE Trans. Inform.*, **56**(5), 2053–2080.

Cawley, G. C. & Talbot, N. L. C. (2006). Gene Selection in Cancer Classification Using Sparse Logistic Regression with Bayesian Regularization. *Bioinformatics*, **22**(19), 2348–2355.

Cawley, G. C., Talbot, N. L. C. & Girolami, M. (2007). Sparse Multinomial Logistic Regression via Bayesian l_1 Regularisation, Neural Information Processing Systems; 209–216.

Chatterjee, A. & Lahiri, S. N. (2011). Bootstrapping Lasso Estimators. *J. Amer. Statist. Assoc.*, **106**(494), 608–625.

Chatterjee, S. (2013). Assumptionless Consistency of the Lasso. *arXiv preprint arXiv:1303.5817*.

Chen, S., Donoho, D. & Saunders, M. (1998). Atomic Decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33–61.

Donoho, D. & Johnstone, I. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, **81**(3), 425–455.

Donoho, D. L. (2006). Compressed Sensing. *IEEE Trans. Inform.*, **52**(4), 1289–1306.

Efron, B. (2004). The Estimation of Prediction Error: Covariance Penalties and Cross-validation. *J. Amer. Statist. Assoc.*, **99**(467), 619–632.

Efron, B., Johnstone, I., Hastie, T. & Tibshirani, R. (2004). Least Angle Regression. *Ann. Statist.*, **32**(2), 407–499.

Fan, J. & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Amer. Statist. Assoc.*, **96**(456), 1348–1360.

Fan, J. & Lv, J. (2010). A selective Overview of Variable Selection in High Dimensional feature space. *Statist. Sinica*, **20**(1), 101–148.

Figueiredo, M. A. T. (2003). Adaptive Sparseness for Supervised Learning. *IEEE Trans. Pattern Anal. Machine Intelligence*, **25**(9), 1150–1159.

Fraley, C. & Hesterberg, T. (2009). Least Angle Regression and LASSO for Large Datasets. *Statist. Anal. Data Mining*, **1**(4), 251–259.

Frank, I. E. & Friedman, J. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**(2), 109–148.

Friedman, J., Hastie, T., Höfling, H. & Tibshirani, R. (2007). Pathwise Coordinate Optimization. *Ann. Appl. Statist.*, **1**(2), 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Statist. Software*, **33**(1), 1–22.

Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Machine Intelligence*, **6**(1), 721–741.

Genkin, A., Lewis, D. D. & Madigan, D. (2007). Large-scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, **49**(3), 291–304.

Greenshtein, E. & Ritov, Y. (2004). Persistence in High-dimensional Linear Predictor-selection and the Virtue of Over-parametrization. *Bernoulli*, **10**(6), 971–988.

Harchaoui, Z. & Lévy-Leduc, C. (2008). Catching Change-points with Lasso, Neural information processing systems; 161–168.

Hardoon, D. R. & Shawe-Taylor, J. (2011). Sparse cCnonical Correlation Analysis. *Machine Learning*, **83**(3), 331–353.

Hastie, T., Taylor, J., Tibshirani, R. & Walther, G. (2007). Forward Stagewise Regression and the Monotone Lasso. *Electron. J. Statist.*, **1**(1), 1935–7524.

Hastie, T., Tibshirani, R. & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Predictions*, 2nd ed. New York: Springer.

Haufe, S., Müller, K. R., Nolte, G. & Krämer, N. (2008). Sparse Causal Discovery in Multivariate Time Series, Neural information processing systems; 97–106.

Hesterberg, T., Choi, N. M., Meier, L. & Fraley, C. (2008). Least Angle and l_1 Penalized Regression: A Review. *Statistics Surveys*, **2**(1), 61–93.

Hoerl, A. & Kennard, R. (1970). Ridge Regression: Biased Estimates for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

Hsu, N. J., Hung, H. L. & Chang, Y. M. (2008). Subset Selection for Vector Autoregressive Processes Using Lasso. *Comput. Statist. Data Anal.*, **52**(7), 3645–3657.

Huang, J., Ma, S., Xie, H. & Zhang, C. H. (2009). A Group Bridge Approach for Variable Selection. *Biometrika*, **96**(2), 339–355.

Huang, J., Ma, S. & Zhang, C. H. (2008). Adaptive Lasso for Sparse High-dimensional Regression Models. *Statist. Sinica*, **18**(374), 1603–1618.

Jacob, L., Obozinski, G. & Vert, J. P. (2009). Group Lasso with Overlap, 26th International Conference on Machine learning; 433–440.

James, G. M., Radchenko, P. & Lv, J. (2009). DASSO: Connections Between the Dantzig Selector and Lasso. *J. Roy. Stat. Soc. Ser. B*, **27**(1), 127–142.

James, W. & Stein, C. (1961). Estimation with Quadratic Loss. In *4th Berkeley Symposium*, Vol. 1, pp. 361–379.

Khan, J. A., van Aelst, S. & Zamar, R. H. (2007). Robust Linear Model Selection Based on Least Angle Regression. *J. Amer. Statist. Assoc.*, **102**(480), 1289–1299.

Knight, K. & Fu, W. (2000). Asymptotics for Lasso-type Estimators. *Ann. Stat.*, **28**(5), 1356–1378.
Krishnapuram, B., Carin, L. & Figueiredo, M. (2005). Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**(6), 957–968.
Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, **5**(2), 369–412.
Lafferty, J. & Wasserman, L. (2008). Rodeo: Sparse, Greedy Nonparametric Regression. *Ann. Stat.*, **36**(1), 28–63.
Lee, H., Battle, A., Raina, R. & Ng, A. Y. (2007). Efficient Sparse Coding Algorithms. In *Neural Information Processing Systems*, pp. 801–808.
Li, C. & Li, H. (2008). Network-constrained Regularization and Variable Selection for Analysis of Genomic Data. *Bioinformatics*, **24**(9), 1175–1182.
Li, Q. & Lin, N. (2010). The Bayesian Elastic Net. *Bayesian Analysis*, **5**(1), 151–170.
Li, Y. & Zhu, J. (2008). l_1-norm Quantile Regression. *J. Comput. Graph. Stat.*, **17**(1), 163–185.
Lin, Y. & Zhang, H. H. (2006). Component Selection and Smoothing in Multivariate Nonparametric regression. *Ann. Stat.*, **34**(5), 2272–2297.
Lorbert, A., Eis, D., Kostina, V., Blei, D. M. & Ramadge, P. J. (2010). Exploiting Covariate Similarity in Sparse Regression Via the Pairwise Elastic Net. In *14th International Conference on Artificial Intelligence and Statistics*, pp. 477–484.
McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
Meier, L. & Bühlmann, P. (2007). Smoothing l_1-penalized Estimators for High-dimensional Time-course Data. *Electon. J. Statist.*, **1**(1), 597–615.
Meier, L., van de Geer, S. & Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *J. Roy. Stat. Soc. Ser. B*, **70**(1), 53–71.
Meinshausen, N. (2007). Lasso with Relaxation. *Comput. Stat. Data Anal.*, **52**(1), 374–393.
Meinshausen, N. & Bühlmann, P. (2006). High Dimensional Graphs and Variable Selection with the Lasso. *Ann. Stat.*, **34**(3), 1436–1462.
Meinshausen, N. & Bühlmann, P. (2010). Stability Selection. *J. Roy. Stat. Soc. B*, **72**(4), 417–473.
Meinshausen, N. & Yu, B. (2009). Lasso-type Recovery of Sparse Representations for High-dimensional Data. *Ann. Stat.*, **37**(1), 246–270.
Osborne, M., Presnell, B. & Turlach, B. (2000). On the LASSO and Its Dual. *J. Comput. Graph. Stat.*, **9**(2), 319–337.
Park, M. Y. & Hastie, T. (2007). l_1-regularization Path Algorithm for Generalized Linear Models. *J. Roy. Stat. Soc. B*, **69**(4), 659–677.
Park, M. Y. & Hastie, T. (2008). Penalized Logistic Regression for Detecting Gene Interactions. *Biostatistics*, **9**(1), 30–50.
Park, T. & Casella, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.*, **103**(482), 681–686.
Porzelius, C., Schumacher, M. & Binder, H. (2010). Sparse Regression Techniques in Low-dimensional Survival Data Settings. *Stat. Comput.*, **20**(2), 151–163.
Pötscher, B. M. & Schneider, U. (2009). On the Distribution of the Adaptive LASSO Estimator. *J. Statist. Plann. Inference*, **139**(8), 2775–2790.
Radchenko, P. & James, G. M. (2008). Variable Inclusion and Shrinkage Algorithms. *J. Amer. Statist. Assoc.*, **103**(483), 1304–1315.
Radchenko, P. & James, G. M. (2010). Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions. *J. Amer. Statist. Assoc.*, **105**(492), 1541–1553.
Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009). Sparse additive models. *J. Roy. Stat. Soc. B*, **71**(5), 1009–1030.
Rosset, S. & Zhu, J. (2007). Piecewise Linear Regularized Solution Paths. *Ann. Stat.*, **35**(3), 1012–1030.
Rosset, S., Zhu, J. & Hastie, T. (2004). Boosting as a Regularized Path to a Maximum Margin Classifier. *J. Mach. Learn. Res.*, **5**(1), 941–973.
Roth, V. (2004). The Generalized LASSO. *IEEE Transactions in Neural Networks*, **15**(1), 16–28.
Shevade, S. & Keerthi, S. (2003). A Simple and Efficient Algorithm for Gene Selection Using Sparse Logistic Regression. *Bioinformatics*, **19**(17), 2246–2253.
Shi, J., Yin, W., Osher, S. & Sajda, P. (2010). A Fast Hybrid Algorithm for Large-scale l_1-regularized Logistic Regression. *J. Mach. Learn. Res.*, **11**(1), 713–741.
Similä, T. & Tikka, J. (2006). Common Subset Selection of Inputs in Multiresponse Regression. In *International Joint Conference on Neural Networks*, pp. 1908–1915.
Similä, T. & Tikka, J. (2007). Input Selection and Shrinkage in Multiresponse Linear Regression. *Comput. Stat. Data Anal.*, **52**(1), 406–422.
Sprechmann, P., Ramírez, I., Sapiro, G. & Eldar, Y. (2010). Collaborative Hierarchical Sparse Modeling. In *44th Annual Conference on Information Sciences and Systems*, pp. 1–6.

Tian, G. L., Tang, M. L., Fang, H. B. & Tan, M. (2008). Efficient Methods for Estimating Constrained Parameters with Applications to Lasso Logistic Regression. *Comput. Stat. Data Anal.*, **52**(7), 3528–3542.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. Ser. B*, **58**(1), 267–288.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Stat. Med.*, **16**(4), 385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and Smoothness via the Fused Lasso. *J. Roy. Stat. Soc. Ser. B*, **67**(1), 91–108.

Tikhonov, A. N. (1943). On the Stability of Inverse Problems (in Russian). *Doklady Akademii Nauk SSSR*, **39**(5), 176–179.

Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nonsmooth separable minimization. *J. Optimiz. Theory App.*, **109**(3), 475–494.

Tseng, P. & Yun, S. (2009). A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization. *Math. Program B*, **117**(1-2), 387–423.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-Garía, L. & Canales-Rodríguez, E. (2005). Estimating Brain Functional Connectivity with Sparse Multivariate Autoregression. *Philos. T. R. Soc. B*, **360**(1457), 969–981.

van de Geer, S. & Bühlmann, P. (2009). On the Conditions used to Prove Oracle Results for the Lasso. *Electron. J. Stat.*, **3**, 1360–1392.

van Gerven, M. A. J., Cseke, B., de Lange, F. P. & Heskes, T. (2010). Efficient Bayesian Multivariate fMRI Analysis Using a Sparsifying Spatio-temporal Prior. *Neuroimage*, **50**(1), 150–161.

Vidakovic, B. (1999). *Statistical Modeling with Wavelets.* New York: Wiley.

Vidaurre, D., Bielza, C. & Larrañaga, P. (2013). Classification of Neural Signals from Sparse Autoregressive Features. *Neurocomputing*, **111**(2), 21–26.

Wang, H. & Leng, C. (2007). Unified LASSO Estimation via Least Squares Approximation. *J. Amer. Statist. Assoc.*, **102**(479), 1039–1048.

Wang, H. & Leng, C. (2008). A note on adaptive group lasso. *Comput. Stat. Data Anal.*, **52**(12), 5277–5286.

Wang, H., Li, G. & Jiang, G. (2007a). Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *J. Bus. Econ. Stat.*, **25**(3), 347–355.

Wang, H., Li, G. & Tsai, C. L. (2007b). Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. Ser. B*, **69**(1), 63–78.

Wang, S., Nan, B., Rosset, S. & Zhu, J. (2011). Random Lasso. *Ann. Appl. Stat.*, **5**(1), 468–485.

Williams, P. M. (1995). Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, **7**(1), 117–143.

Wu, T. T. & Lange, K. (2008). Descent Algorithms for Lasso Penalized Regression. *Ann. Appl. Stat.*, **2**(1), 224–244.

Xu, H., Caramanis, C. & Mannor, S. (2010). Robust Regression and Lasso. *IEEE Trans. Inf. Theory*, **56**(7), 3561–3574.

Yang, J., Zhang, Y. & Yin, W. (2010). A Fast TVL1-L2 Minimization Algorithm for Signal Reconstruction from Partial Fourier Data. *IEEE J. Sel. Top. Signa.*, **4**(2), 288–297.

Yuan, M. & Lin, Y. (2005). Efficient Empirical Bayes Variable Selection and Estimation in Linear Models. *J. Amer. Statist. Assoc.*, **103**(472), 1215–1225.

Yuan, M. & Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *J. Roy. Stat. Soc. Ser. B*, **70**(1), 53–71.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2004). Variable Selection and Model building via Likelihood Basis Pursuit. *J. Amer. Statist. Assoc.*, **99**(467), 659–672.

Zhao, P., Rocha, G. & Yu, B. (2009). The Composite Sbsolute Penalties Family for Grouped and Hierarchical Variable Selection. *Ann. Stat.*, **37**(6A), 3468–3497.

Zhao, P. & Yu, B. (2006). On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.*, **7**(1), 2541–2567.

Zhao, P. & Yu, B. (2007). Stagewise Lasso. *J. Mach. Learn. Res.*, **8**(1), 2701–2726.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, **101**(12), 1418–1429.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. Roy. Stat. Soc. Ser. B*, **67**(2), 301–320.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "Degrees" of Freedom of the Lasso. *Ann. Stat.*, **35**(5), 2173–2192.

Résumé

La régularisation L_1, ou régularisation par pénalisation L_1, est une notion populaire en statistique et en "machine learning". Cet article examine le concept et les applications en régression de ces méthodes de régularisation. Notre but n'est pas de présenter une liste exhaustive des usages de la pénalisation L_1 dans les problèmes de régression; au contraire, nous nous concentrons sur ce que nous croyons être l'ensemble des usages les plus représentatifs de cette technique, et les décrivons en détail. Ainsi, nous traitons d'un certain nombre de méthodes faisant intervenir la régularisation L_1 en régression linéaire, dans les modéles linéaires généralisés, et en analyse des séries temporelles. Bien que cette revue cible la pratique plutôt que la théorie, nous donnons quelques précisions théoriques sur la méthode couramment désignée sous le nom de "lasso".

[*Received September 2012, accepted May 2013*]