International Statistical Review

International Statistical Review (2022), **90**, 1, 118–145 doi: 10.1111/insr.12469

A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates

Laura Freijeiro-González, Manuel Febrero-Bande and Wenceslao González-Manteiga

Department of Statistics, Mathematical Analysis and Optimization; Santiago de Compostela University, Santiago de Compostela, Spain
E-mail: laura.freijeiro.gonzalez@usc.es

Summary

The limitations of the well-known LASSO regression as a variable selector are tested when there exists dependence structures among covariates. We analyse both the classic situation with $n \geq p$ and the high dimensional framework with $p > n$. Known restrictive properties of this methodology to guarantee optimality, as well as inconveniences in practice, are analysed and tested by means of an extensive simulation study. Examples of these drawbacks are showed making use of different dependence scenarios. In order to search for improvements, a broad comparison with LASSO derivatives and alternatives is carried out. Eventually, we give some guidance about what procedures work best in terms of the considered data nature.

Key words: Covariates selection; $p > n$; L_1 regularisation techniques; LASSO.

1 Introduction and Motivation

Nowadays, in many important statistical applications, it is of high relevance to apply a first variable selection step to correctly explain the data and avoid unnecessary noise. Furthermore, it is usual to find that the number of variables p is larger than the number of available samples $n(p > n)$. Some examples of fields where this framework arises are processing image, statistical signal processing, genomics or functional magnetic resonance imaging (fMRI), among others. It is in the $p > n$ context where the ordinary models fail, and as a result, estimation and prediction in these settings are generally acknowledged as an important challenge in contemporary statistics.

In this framework, one of the most studied fields is the regression models adjustment. The idea of a regression model is to explain a variable of interest, Y, using p covariates $X_1, \ldots, X_p$. This is carried out by means of a structure $m(X)$, $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$, and an unknown error ε:

$$Y = m(X) + \varepsilon. \tag{1}$$

Here, $m(X)$ denotes the type of relation between the dependent variable Y and the p explanatory covariates, while ε is the term that captures the remaining information and other unobserved fluctuations. This is typically assumed to have null mean and variance σ^2.

Once the $m(X)$ structure of (1) is estimated, it is possible to know the importance of every $X_1, \ldots, X_p$ in terms of explaining Y apart from making predictions. Nevertheless, this estimation when $p > n$ still is a difficult and open problem in many situations. In particular, its easiest expression, the linear regression model, has been widely studied in the last years in order to provide efficient algorithms to fit this (see, e.g. Giraud, 2014 or Hastie *et al*., 2015).

In linear regression, as the name suggests, the relationship between Y and X is assumed to be linear, giving place to the model:

$$Y = X\beta + \varepsilon, \tag{2}$$

where $\beta \in \mathbb{R}^p$ is a coefficients vector to estimate. Note that we assume the covariates $X_1, \ldots, X_p$ and the response Y centred, excluding the intercept from the model without loss of generality.

Having data $(y_i, \mathbf{x_i}) \in \mathbb{R}^{p+1}$ for $i = 1, \ldots, n$ samples, denoting $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p}) \in \mathbb{R}^{n \times p}$ with $\mathbf{x_j} = (x_{ij})_{i=1}^n \in \mathbb{R}^n$ for $j = 1, \ldots, p$, the β vector can be estimated using the classical ordinary least squares (OLS) method solving (3).

$$\hat{\beta}^{OLS} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \tag{3}$$

This estimator, $\hat{\beta}^{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top y$, enjoys some desirable properties such as being an unbiased and consistent β estimator of minimum variance.

Nevertheless, when $p > n$, this estimation method fails, as there are infinite solutions for the problem (3). Then, it is necessary to impose modifications on the procedure or to consider new estimation algorithms able to recover the β values.

In order to overcome this drawback, the LASSO regression (Tibshirani, 1996) is still widely used due to its capability of reducing the dimension of the problem. A survey on this topic can be found in Vidaurre *et al*. (2013). This methodology assumes sparsity in the coefficient vector β, resulting in an easier interpretation of the model and performing variable selection. However, some rigid assumptions on the covariates matrix and sample size are needed to guarantee its good behaviour (see, e.g. Meinshausen & Bühlmann, 2010). Moreover, the LASSO procedure exhibits some drawbacks related to the correct selection of covariates and the exclusion of redundant information (see Su *et al*., 2017), aside from bias. This can be easily showed in controlled simulated scenarios where it is known that only a small part of the covariates are relevant.

It is important to highlight that variable selection can be understood in two different ways: trying to identify the set of true relevant covariates or doing dimension reduction in order to improve predictions without guaranteeing the recovering of the true model. It is well known that they are not compatible (see, e.g. Yang, 2005), and then, one of them must be chosen. For example, in the LASSO case, the optimal value of the penalisation parameter λ may not be the same for both objectives (Leng *et al*., 2006), Bühlmann & Van De Geer (2011). Besides, some drawbacks for the true recovering of the not null elements of the β vector are less harmful for the prediction accuracy target (Dalalyan *et al*., 2017).

In this paper, we focus in the first objective. Henceforth, variable selection is understood from a statistical inference point of view, trying to identify the set of covariates with non-zero β_j in the linear model (2). Then, the discussion does not necessary apply to the second objective.

Thus, is LASSO the best option or at least a good start point to identify the relevant covariates? Although some studies as the one of Su *et al*. (2017) discuss this topic, we have not found a totally convincing answer to this question for dependence scenarios. In order to shed light about this topic, we analyse the theoretical requirements of the LASSO procedure and test its performance under different dependence structures. For this purpose, we start revisiting the existing literature about this topic and its most important adaptations. Furthermore, in view of the LASSO limitations, a global comparison is developed to test which procedures are capable of overcoming these in different dependence contexts, comparing their performance with alternatives that have proved their efficiency. Finally, some conclusions based on the simulation results are drawn.

The article is organised as follows, in Section 2, a complete overview of the LASSO regression is given, including a summary of the requirements and inconveniences this algorithm has to deal with. In Section 3, some special simulation scenarios are introduced and used to illustrate the problems of this methodology in practice, testing the behaviour of the LASSO under different dependence structures. In Section 4, the evolution of the LASSO in the last years is analysed. Besides, other efficient alternatives in covariates selection are briefly described, and their performance is compared with the LASSO one. Eventually, in Section 5, a discussion is provided in order to give some guidance about what types of covariates selection procedures are the best ones in terms of the considered data dependence structure.

2 A Complete Overview of the LASSO Regression

In a linear regression model as the one of (2), there are a lot of situations where not all p explanatory covariates are relevant, but several are unnecessary. In these scenarios, we can assume that the β vector is sparse and then search for the important covariates, avoiding noisy ones. The idea is, somehow, to obtain a methodology able to compare the covariates and select only those most important, discarding irrelevant information and keeping the error of prediction as small as possible. As there are 2^p possible sub-models, it tends to be rather costly to compare all of them using techniques such as forward selection or backward elimination.

One of the most typical solutions is to impose a restriction on the number of included covariates. This is carried out by means of adding some constraints to the OLS problem (3).

This brings up the idea of a model selection criterion, which express a trade-off between the goodness-of-fit and the complexity of the model, such as the AIC (Akaike, 1998) or BIC (Schwarz, 1978). Nevertheless, these approaches are computationally intensive, hard to derive sampling properties and unstable. As a result, they are not suitable for scenarios where the dimension of p is large.

Therefore, we could think in penalising the irrelevant information by means of the number of coefficients included in the final model. This can be carried out by adding a penalty factor $p_\lambda(\beta)$ in (3), resulting in the problem

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + p_\lambda(\beta) \right\}. \tag{4}$$

For this purpose, following the ideas of goodness-of-fit measures, a L_0 regularisation, $\lambda\|\beta\|_0 = \lambda\sum_{j=1}^{p} \mathbf{1}_{\beta_j \neq 0}$, could be applied. This criterion penalises models that include more covariates but do not improve too much the performance. This results in a model with the best trade-off between interpretability and accuracy, as the AIC or BIC criterion philosophy does, obtaining

$$\hat{\beta}^{L_0} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \mathbf{1}_{\beta_j \neq 0} \right\}, \tag{5}$$

where $\lambda > 0$ is a regularisation parameter.

The problem (5) is known as the best subset selection (Beale *et al*., 1967), Hocking & Leslie (1967). This is non-smooth and non-convex, which hinders to achieve an optimal solution. As a result, the estimator $\hat{\beta}^{L_0}$ is infeasible to compute when p is of medium or large size, as (5) becomes a *NP*-hard problem with exponential complexity (Natarajan, 1995). However, when p is small, this estimator can still be used in practice. Moreover, it is known that this estimator is optimal in the minimax sense (Bunea *et al*., 2007), even when the assumptions required for the LASSO are not satisfied.[1] See Hastie *et al*. (2017) for a comparison of this procedure with more current methodologies.

To avoid this drawback, it is possible to replace $\lambda\|\beta\|_0$ by other types of penalisation. Taking into account that this belongs to the family $p_\lambda(\beta_j) = \lambda\|\beta\|_q := \lambda\left(\sum_{j=1}^{p} \sqrt[q]{|\beta_j|}\right)^q$, with $q \geq 0$, we can commute this for a more appropriate one. The problem (4) with this type of penalisation is known as the bridge regression (Fu, 1998). The caveat of this family is that this only selects covariates for the values $1 \geq q > 0$. Moreover, the problem (4) is only convex for the $q = 1$ case (see Figure 1). Then, it seems reasonable to work with the norm $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, which is convex, allows covariates selection and leads to the extensively studied LASSO (least absolute shrinkage and selection operator) regression (see Tibshirani, 1996) and Tibshirani (2011).

The LASSO, also known as basis pursuit in image processing (Chen *et al*., 2001; Donoho *et al*., 2005; Candes *et al*., 2006, was presented by Tibshirani 1996). This proposes the imposition of a L_1 penalisation in (3) with the aim of performing variable selection and overcoming the high dimensional estimation of β drawback when $p > n$. In this way, it would be needed to solve the problem

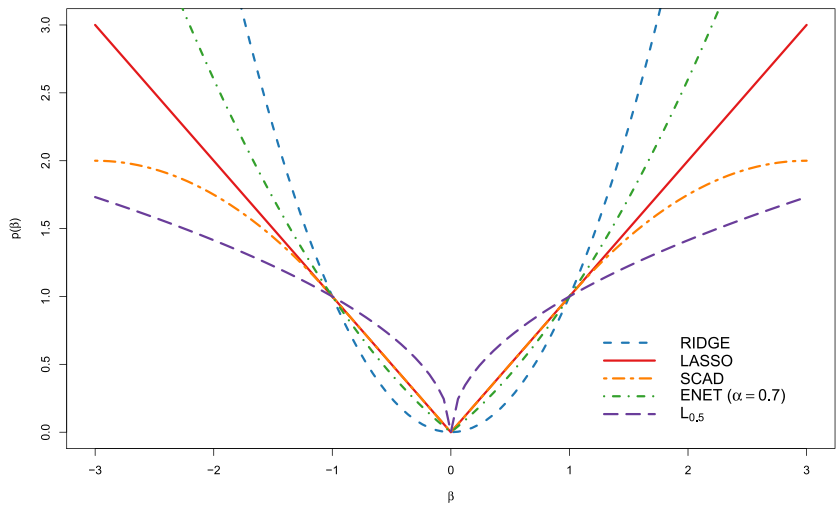

Figure 1. *Comparison of different penalisation methods: L_2 or RIDGE penalisation (RIDGE), L_1 or LASSO penalisation (LASSO), SCAD regularisation (SCAD), elastic net penalisation method for $\alpha = 0.7$ ($ENET\ (\alpha = 0.7)$) and $L_{0.5}$ regularisation ($L_{0.5}$)*

$$\hat{\beta}^{L_1} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{6}$$

The problem (6) is convex, which guarantees that has always one solution at least, although if $p > n$, there may be multiple minimums (see Tibshirani, 2013 or Schneider & Tardivel, 2020, for more details). Besides, assuming the noise term ε to be Gaussian, $\hat{\beta}^{L_1}$ can be interpreted as a penalised maximum likelihood estimate in which the fitted coefficients are penalised in a L_1 sense. As a result, this encourages sparsity.

In this problem, the term $\lambda > 0$ is the shrinkage parameter. For large values of λ the coefficients of β are more penalised, which results in a bigger number of elements shrinkaged to zero. Nevertheless, the estimator $\hat{\beta}^{L_1}$ of (6) has not got an explicit expression.

The LASSO defined in (6) can be viewed as a convex relaxation of the optimisation problem with the L_0 analogue of a norm in (5). Then, the requirement of computational feasibility and statistical accuracy can be met by this estimator.

This method has been widely studied over the last years: it has been showed that this procedure is consistent in terms of prediction (see Van De Geer & Bühlmann, 2009, for an extensive analysis) and this guarantees consistency of the parameter estimates at least in a L_2 sense (Van De Geer & Bühlmann (2009), Meinshausen & Yu (2009), Candes & Tao (2007); besides, this is a consistent variable selector under some assumptions (Meinshausen & Bühlmann (2006), Wainwright (2009), Zhao & Yu (2006)).[2]

2.1 Analysis of the LASSO Regression Requirements and Inconveniences

In spite of all these good qualities, the LASSO regression has some important limitations in practice (see, e.g. Zou & Hastie, 2005 or Su *et al*., 2017). These limitations are analysed in the next subsections, collecting some recent developed theoretical properties and displaying how far it is possible to ensure its good behaviour.

2.1.1 Biased estimator

In the context of having more covariates p, than number of samples n, it depends on the type of optimisation algorithm employed to solve (6) that the LASSO regression can identify more than *n*important covariates. For example, making use of algorithms as the coordinate descent (see, e.g. Section 4.2.4 of Giraud, 2014), this is able to select until p covariates. However, for the LARS algorithm of Efron *et al*. (2004), this can select at most n variables before this saturates (see Zou & Hastie, 2005). This restriction is common for almost all regression adjustment methods which rely on penalisations in this framework. Especially for those based on L_1 ideas.

Related with this, another caveat of penalisation processes is the bias. This produces higher prediction errors. In the LASSO adjustment, the imposition of the L_1 penalisation in the OLS problem (3) as a safe passage to estimate β has a cost, which is translated in bias (see Chapter 3 of Hastie *et al*., 2009), Chapter 4 of Giraud (2014) or Chapter 2 of Hastie *et al*. (2015)). This can be easily explained under orthogonal design, where the L_1 penalisation results in a perturbation of the unbiased OLS estimator $\hat{\beta}^{OLS}$ given by

$$\hat{\beta}_j^{L_1} = \text{sign}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda)_+, \tag{7}$$

where sign(·) denotes the sign of the coefficients and $(\cdot)_+$ equals to zero all quantities which are not positive. This results in a soft threshold of the ordinary mean square estimator ruled by the $\lambda > 0$ parameter, where the coefficients $|\hat{\beta}_j^{OLS}| \leq \lambda$ are adjusted to zero.

In order to correct the bias, it is usual to apply a two-step LASSO-OLS procedure: first, a LASSO regression is employed to select variables, and then, a least squared estimator is obtained over the selected variables. The properties of this procedure have been studied in Belloni & Chernozhukov (2013).

Other options are weighted versions of the LASSO method based on iterative schemes. An example is the popular adaptive LASSO (Zou, 2006; Huang *et al*., 2008; Van de Geer *et al*., 2011). This procedure gives different weights to each covariate in the penalisation part, readjusting these in every step of the iterative process until convergence.

2.1.2 Consistency of the LASSO: neighbourhood stability condition

Despite the LASSO is broadly employed, it is not always possible to guarantee its proper performance as variable selector in practice (Bunea, 2008; Lounici, 2008). As we can see in Bühlmann & Van De Geer (2011), certain conditions are required to guarantee an efficient screening property for variable selection. However, this presents some important limitations as a variable selector when these do not hold.

For example, when the model has several highly correlated covariates with the response, LASSO tends to pick randomly only one or a few of them and shrinks the rest to 0 (see Zou & Hastie, 2005). This fact results in a confusion phenomenon if there are high correlations between relevant and unimportant covariates, and in a loss of information when the subset of important covariates have a strong dependence structure. Some algorithms which result in non-sparse estimators try to relieve this effect, like the Ridge regression (Hoerl & Kennard, 1970) or the elastic net (Zou & Hastie, 2005). An interpretation of their penalties is displayed in Figure 1.

Besides, denoting $S = \{j : \beta_j \neq 0\}$ the set of non-zero real values, for consistent variable selection using $\hat{S}^{L_1} = \{j : \hat{\beta}_j^{L_1} \neq 0\}$, the design matrix of the model, X, needs to satisfy some assumptions. The strongest of which is arguably the so-called 'neighbourhood stability condition' (Meinshausen & Bühlmann, 2006). This condition is equivalent to the irrepresentable condition (Zhao & Yu, 2006; Zou, 2006; Yuan & Lin, 2007):

$$\max_{j \in S^c} |\text{sign}(\beta_S)^\top (X_S^\top X_S)^{-1} X_S^\top X_j| \leq \theta \text{ for some } 0 < \theta < 1, \tag{8}$$

being β_S the subvector of β and X_S the submatrix of X considering the elements of S.

If this condition is violated, all that we can hope for is recovery of the regression vector β in an L_2-sense of convergence by achieving $||\hat{\beta}^{L_1} - \beta||_2 \xrightarrow[n \to \infty]{p} 0$ (see Meinshausen & Bühlmann, 2010, for more details). Moreover, under some assumptions in the design, the irrepresentable condition can be expressed as the called 'necessary condition' (Zou, 2006). It is not an easy task to verify these conditions in practice, especially in contexts where p can be huge.

Quoted Bühlmann & Van De Geer (2011): roughly speaking, the neighbourhood stability or irrepresentable condition (8) fails to hold if the design matrix X is too much 'ill-posed' and exhibits a too strong degree of linear dependence within 'smaller' sub-matrices of X.

In addition, it is needed to assure that there are enough information and suitable characteristics for 'signal recovery' of the sparse β vector. This requires relevant covariates coefficients be

large enough in order to distinguish them from the zero ones. Then, the non-zero regression coefficients need to satisfy

$$\inf_{j \in S} |\beta_j| > \sqrt{s \log(p)/n} \tag{9}$$

in order to guarantee the consistency of the $\hat{\beta}^{L_1}$ estimator of problem (6). This is called a beta-min condition. Nevertheless, this requirement may be unrealistic in practice and small non-zero coefficients may not be detected (in a consistent way). See Bühlmann & Van De Geer (2011) for more information.

Eventually, related with all these requirements, it is important to remind that for all covariates selection procedures, an estimator $\hat{S}$ trying to recover S would be consistent if this verifies

$$\mathbb{P}(\hat{S} = S) \underset{n \to \infty}{\to} 1. \tag{10}$$

The condition (10) places a restriction on the growth of the number p of variables and sparsity $|S|$, typically of the form $|S|\log(p) = o(n)$ (see Meinshausen & Bühlmann (2006)). Then, denoting $s = |S|$, this forces the necessity of $n > s\log(p)$ in order to achieve consistency.

Bunea (2008) explains that, under mild assumptions, the LASSO verifies condition (10), and then this is capable of selecting the relevant variables. However, more assumptions as the irrepresentable condition of (8) are needed to verify (10). This may explain why the LASSO overestimates the support of β.

Owing to these difficulties, different new methodologies based on ideas derived from subsampling and bootstrap have been developed. Examples are the random LASSO (Wang *et al*., 2011), an algorithm based on subsampling, or the stability selection method mixed with randomised LASSO of Meinshausen & Bühlmann (2010). This last searches for consistency although the irrepresentable condition introduced in (8) would be violated.

2.1.3 False discoveries of the LASSO

As it is explained in Su *et al*. (2017): In regression settings where explanatory variables have very low correlations and there are relatively few effects, each of large magnitude, we expect the LASSO to find the important variables with few errors, if any. Nevertheless, in a regime of linear sparsity, there exist a trade-off between false and true positive rates along the LASSO path, even when the design variables are stochastically independent. Besides, this phenomenon occurs no matter how strong the effect sizes are. By linear sparsity, it is understood that the fraction of variables with a non-vanishing effect, that is, $s = |S|$, tends to a constant, however small.

This can be translated as one of the major disadvantages of using LASSO like a variable selector is that exists a trade-off between the false discovery proportion (FDP) and the true positive proportion (TPP), which are defined as

$$FDP(\lambda) = \frac{F(\lambda)}{|\{j : \hat{\beta}_j(\lambda) \neq 0\}| \vee 1} \text{ and } TPP(\lambda) = \frac{T(\lambda)}{s \vee 1}, \tag{11}$$

where $F(\lambda) = |\{j \in S^c : \hat{\beta}_j(\lambda) \neq 0\}|$ denotes the number of false discoveries, $T(\lambda) = |\{j \in S : \hat{\beta}_j(\lambda) \neq 0\}|$ is the number of positive discoveries and $a \vee b = \max\{a, b\}$.

Then, it is unlikely to achieve high power and a low false positive rate simultaneously. Noticing that *FDP* is a natural measure of type I error while $1 - TPP$ is the fraction of missed signals (a natural notion of type II error), the results say that nowhere on the LASSO path can both

types of error rates be simultaneously low. This also happens even when there is no noise in the model and the regressors are stochastically independent. Hence, there exists only a possible reason: it is because of the L_1 shrinkage which results in pseudo-noise. Furthermore, this does not occur with other types of penalisations, like the L_0 penalty. See Su *et al*. (2017) for more details.

In fact, it can be proved in a quite global context, that the LASSO is not capable of selecting the correct subset of important covariates without adding some noise to the model in the best case (see Wasserman & Roeder, 2009 or Su *et al*., 2017).

Then, modifications of the traditional LASSO procedure are needed in order to control the FDP. Some alternatives, such as the boLASSO procedure (see Bach, 2008), which use bootstrap to calibrate the *FDP*, the thresholded LASSO (Lounici, 2008; Zhou, 2010), based on the use of a threshold to avoid noise covariates, or more recent ones, like the stability selection method (see Meinshausen & Bühlmann, 2010) or the use of knockoffs (see Hofner *et al*., 2015, Weinstein *et al*., 2017; Candes *et al*., 2018 and Barber & Candès, 2019), were proposed to solve this drawback. To the best of our knowledge, still there is not a version of this last for the $p > n$ framework.

2.1.4 Correct selection of the penalisation parameter $\boldsymbol{\lambda}$

One of the most important parts of a LASSO adjustment is the proper selection of the penalisation parameter $\lambda \geq 0$. Its size controls both the number of selected variables and the degree to which their estimated coefficients are shrunk to zero, controlling the bias as well. A too large value of λ forces all coefficients of $\hat{\beta}^{L_1}$ to be null, while a value next to zero includes too many noisy covariates. Then, a good choice of λ is needed in order to achieve a balance between simplicity and selection accuracy. See the work of Lahiri (2021) for a current analysis of λ conditions required.

The problem of the proper choice of the λ parameter depends on the unknown error variance σ^2. We can see in Bühlmann & Van De Geer (2011) that the oracle inequality states to select λ of order $\sigma\sqrt{\log(p)/n}$ to keep the mean squared prediction error of LASSO as the same order as if we knew the active set S in advance. In practice, the σ value is unknown and its estimation with $p > n$ is quite complex. To give some guidance in this field, we refer to Fan *et al*. (2012) or Reid *et al*. (2016), although this still is a growing study field.

Thus, other methods to estimate λ are proposed. Following the classification of Homrighausen & McDonald (2018), we can distinguish three categories: minimisation of a generalised information criteria (like AIC or BIC), by means of resampling procedures (such as cross-validation or bootstrap) or reformulating the LASSO optimisation problem. Due to computational cost, the most used criteria to fit a LASSO adjustment are cross-validation techniques. Nevertheless, it can be showed that this criterion achieves an adequate λ value for prediction risk, but this leads to inconsistent model selection for sparse methods (see Meinshausen & Bühlmann, 2006). Then, for recovering the set S, a larger penalty parameter would be needed (Bühlmann & Van De Geer, 2011).

Su *et al*. (2017) argue that, when the regularisation parameter λ is needed to be large for a proper variable selection, the LASSO estimator is seriously biased downwards. The residuals still contain much of the effects associated with the selected variables, which is called shrinkage noise. As many strong variables get picked up, this gets inflated and its projection along the directions of some of the null variables may actually dwarf the signals coming from the strong regression coefficients, selecting null variables.

Nevertheless, to the best of our knowledge, there is not a common agreement about the way of choosing this λ value. Hence, cross-validation techniques are widely used to adjust the LASSO regression. See Homrighausen & McDonald (2018) for more details.

Modifications of the LASSO algorithm as the square-root LASSO (Belloni *et al*., 2011), which does not need to known σ to obtain an optimal λ, the work of Städler *et al*. (2010) or the scaled LASSO (Sun & Zhang, 2012), which simultaneously estimate σ and β, have been proposed to relieve these inconveniences. A complete survey on this topic is carried out in Giraud *et al*. (2012).

3 A Comparative Study with Simulation Scenarios

Next, performance of LASSO as variable selector under different dependence structures is tested in practice. Scenarios verifying and do not the aforementioned conditions are simulated, and its results are compared with those of other procedures. For this purpose, a Monte Carlo study taking $M = 500$ simulations is carried out. Three dependence scenarios are introduced, simulating them under the linear regression model structure given by (2). We consider β as a sparse vector of length p with only $s < p$ values not equal zero, $X \in \mathbb{R}_{n \times p}$ where $n$is the sample size and $\varepsilon \in N_n(0, \sigma^2 I_n)$. We fix $p = 100$ and choose σ^2 verifying that the percentage of explained deviance is explicitly the 90%. Calculation of this parameter is collected in Section 1 of the supporting information. To guarantee the optimal LASSO performance, it is needed that $n > 4.61s$ as we saw in (10), $\inf|\beta_j| > 2.15\sqrt{s/n}$ for $j \in S$ as in (9) and taking λ of order $2.15\sigma\sqrt{1/n}$. To test its behaviour, we consider different combinations of parameters values taking $n = 25, 50, 100, 200, 400$ and $s = 10, 15, 20$. A study of when these conditions hold is showed in Section 2 of the supporting information. In every simulation, we count the number of covariates correctly selected ($|\hat{S} \cap S|$) and the noisy ones ($|\hat{S} \setminus S|$). Besides, the prediction power of the algorithm is measured by means of the mean squared error (MSE) and the percentage of explained deviance %Dev $= (RSS - RSS_0)/(RSS_0)$, being $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}X_i)^2$ the residual sum of squares of the model and $RSS_0 = \sum_{i=1}^{n} y_i^2$. The *MSE* gives us an idea about the bias produced by the LASSO (see Section 2.1.1).

- **Scenario 1** (*Orthogonal design*). Only the first s values are not equal zero for β_j with $j = 1, \ldots, s$ and $p > s > 0$, $\beta_1 = \ldots = \beta_s = 1.25$, while $\beta_j = 0$ for all $j = s + 1, \ldots, p$. X is simulated as a $N_n(0, I_p)$.
- **Scenario 2** (*Dependence by blocks*). The vector β has the first $s < p$ components not null, of the form $\beta_1 = \ldots = \beta_s = 1$ and $\beta_j = 0$ for the rest. X is simulated as a $N_n(0, \Sigma)$, where $\sigma_{jj} = 1$ and $\sigma_{jk} = cov(X_j, X_k) = 0$ for all pairs (j, k) except if $mod_{10}(j) = mod_{10}(k)$, in that case $\sigma_{jk} = \rho$, taking $\rho = 0.5, 0.9$.
- **Scenario 3** (*Toeplitz covariance*). Again, only s ($p > s > 0$) covariates are important, simulating X as a $N_n(0, \Sigma)$ and assuming $\beta_j = 0.5$ in the places where $\beta \neq 0$. In this case, $\sigma_{jk} = \rho^{|j-k|}$ for $j, k = 1, \ldots, p$ and $\rho = 0.5, 0.9$. Now, we analyse two different dependence structures varying the location of the s relevant covariates:
  - **Scenario 3.a**: we assume that the relevant covariates are the first $s = 15$.
  - **Scenario 3.b**: consider $s = 10$ relevant variables placed every 10 sites, which means that only the $\beta_1, \beta_{11}, \beta_{21}, \ldots, \beta_{91}$ terms of β are not null.

The first choice, the orthogonal design of Scenario 1, is selected as the best possible framework. This verifies the consistency conditions for values of n large enough and avoids the confusion phenomenon given that there are not correlated covariates.

In contrast, to assess how the LASSO behaves in case of different dependence structures, Scenarios 2 and 3 are proposed. In the dependence by blocks context (Scenario 2), we force the design to have a dependence structure where the covariates are correlated 10 by 10. As a result, we induce a more challenging scenario for the LASSO, in which the algorithm has to

overcome a fuzzy signal produced by irrelevant covariates. Different magnitudes of dependence are considered in Scenario 2 with $\rho = 0.5$ and $\rho = 0.9$ to test the effect of the confusion phenomenon. As a result, different sizes of n are needed in terms of s to guarantee the proper behaviour of the LASSO. This scenario has been studied in other works, like in Meinshausen & Bühlmann (2010).

Eventually, the LASSO performance is tested in a scenario were all the covariates are correlated: the Toeplitz covariance structure (Scenario 3). This mimics a time series dependence pattern. This is an example where the irrepresentable condition holds (see Section 2.6.1 of Bühlmann & Van De Geer 2011), but the algorithm suffers from highly correlated relations between the true set of covariates and unimportant ones. This framework has been studied for the LASSO case (see, e.g. Meinshausen & Bühlmann 2010 or Bühlmann & Van De Geer 2011). Because the distance between covariates is relevant to establish their dependence, we study two different frameworks. In the first scenario (Scenario 3.a), the important covariates are highly correlated among them and little with the rest. Particularly, there are only notable confusing correlations in the case of the last variables of $S = \{1, \ldots, 15\}$ with their noisy neighbours. Here, the LASSO is only able to recover S in the $n = 400$ case. In contrast, in the Scenario 3.b, the important covariates are markedly correlated with unimportant ones, contributing to magnify the spurious correlations phenomenon. For this scenario, we need a size of $n = 200, 400$.

We start testing the performance of the standard LASSO using the library `glmnet` (Friedman *et al*., 2010) implemented in R (R Core Team, 2019). This uses K-fold cross-validation (CV) to select the λ parameter which minimises the MSE, $\lambda^{\min}$. We denote this procedure by LASSO.min. See Friedman *et al*. (2010) for more details. As it was explained in Section 2.1.4, this is one of the most popular ways of estimating λ. In order to be capable of comparing different models and following recommendations of the existing literature, we have fixed $K = 10$ for all simulations. Besides, we work with the response y centred and with the matrix $\mathbf{X}$ standardised by columns. This last would not be really necessary in these frameworks because the covariates are all in the same scale. However, this is carried out to keep the usual implementation of LASSO type algorithms in practice. We apply the two-step LASSO-OLS version introduced above to adjust the model (see Belloni & Chernozhukov, 2013). This scheme is also following for the rest of procedures. The grid of tuning parameter values is taken of length 100 and is calculated based on the sample data and methodology employed, following author's recommendation. More details are given in Section 3.1 of the supporting information.

There are other faster algorithms available in R, such as the famous LARS procedure (Efron *et al*., 2004; Hastie & Efron, 2013). However, we decided to make use of the `glmnet` library due to its easy implementation and interpretation, as well as its simple adaptation to other derivatives of the LASSO we test in this document.

3.1 Performance of the LASSO in Practice

In order to test the inconveniences of the LASSO when there exists dependence among covariates, a complete simulation study is carried out. For this purpose, we make use of the simulation scenarios introduced above. Complete results are displayed in the Section 5 of the supporting information.

In the orthogonal design of Scenario 1, we would expect the LASSO to recover the whole set of important covariates and not to add too much noise into the model for a large enough value of n. However, we have observed different results.

Firstly, we can appreciate that it does not really matter the number of relevant covariates considered ($s = 10, 15, 20$) in relation with the capability of recovering this set. It is because the algorithm only includes the complete set under the $n \geq p$ framework except for the $s = 10$

scenario taking $n = 50$. See this fact in Figure 2. It can be easily explained in terms of the consistence requirements given in (10). Besides, although we are under orthogonal design assumption, this includes a lot of noisy variables in the model. What is shocking is the fact that the number of irrelevant covariates selected is always larger than the important ones. This exemplifies the existing trade-off between FDP and TPP introduced in (11) as well as that both quantities cannot be simultaneously low.

In second place, we notice that this procedure clearly overestimates its results. This obtains values for the MSE and percentage of explained deviance less and greater, respectively, of the oracle ones (see values in brackets in Table 1). In conclusion, with this toy example, we can illustrate how the LASSO.min procedure performs very poorly and present important limitations even in an independence framework.

The overestimation of the set S is likely to be because of a larger value of λ is needed for proper covariates selection. In Section 3.2 of the supporting information, some values greater than $\lambda^{\min}$ are chosen and tested. These outperform the LASSO.min performance in terms of recovering S and avoid irrelevant information. Nevertheless, some guidance criterion is needed to select a penalisation value in practice. Friedman *et al*. (2010) proposed the alternative of estimating the mean cross-validated error for every value of the λ grid and taking λ^{1se}. This value is the largest value of λ such that error is within 1 standard error of the minimum ($\lambda^{\min}$). Complete results of LASSO using λ^{1se} (LASSO.1se) are collected in Section 5 of the supporting information. In view of the results, we can appreciate that this selection of the penalisation makes sense and improves the LASSO.min performance.

As it was pointed out for one referee, the inclusion of too noisy covariates could also be due to the selection criterion. Cross-validation search for the λ value minimising the mean squared error, which is helpful for estimation of $X\beta$ but can fail for covariates selection. Thus, other techniques as information theory based criterion may achieve a better performance recovering S. To fill this gap, a comparison with the Bayesian information criterion (BIC) is carried out. We denote this methodology by LASSO.BIC. A summary of its results is displayed in Figure 3 and

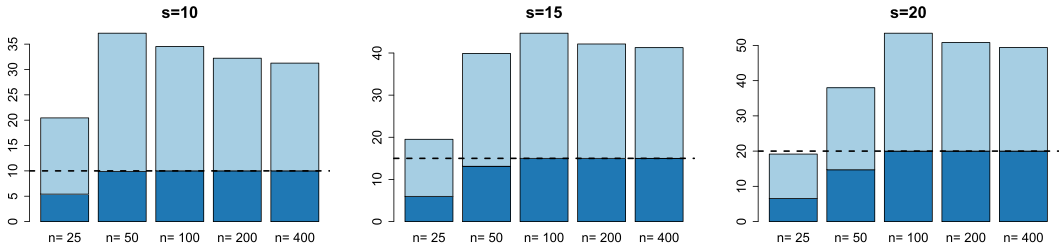

Figure 2. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.min in Scenario 1. The dashed line marks the s value*

Table 1. *Summary of the LASSO.min results for Scenario 1.*

	$s = 10$		$s = 15$		$s = 20$	
	MSE (1.736)	*% Dev*	*MSE (2.604)*	*% Dev*	*MSE (3.472)*	*% Dev*
$n = 50$	0.169	0.990	0.579	0.973	1.747	0.944
$n = 100$	0.702	0.959	0.841	0.967	0.914	0.973
$n = 200$	1.164	0.932	1.628	0.936	2.060	0.940

The oracle value for the deviance is 0.9, and those for the MSE are in brackets.

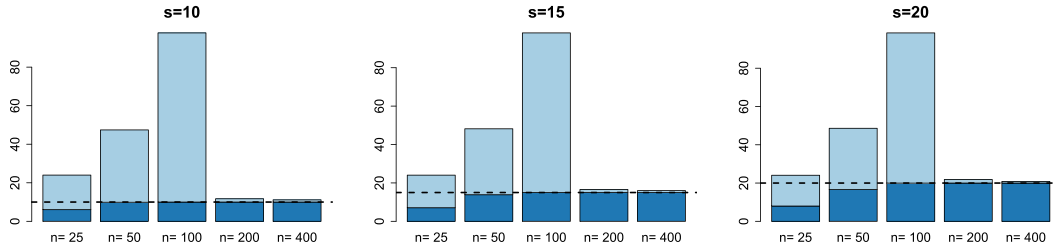

Figure 3. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.BIC in Scenario 1. The dashed line marks the s value*

Table 2. *Summary of the LASSO.BIC results for Scenario 1.*

	$s = 10$		$s = 15$		$s = 20$	
	MSE (1.736)	*% Dev*	*MSE (2.604)*	*% Dev*	*MSE (3.472)*	*% Dev*
$n = 50$	0.001	1	0.001	1	0.001	1
$n = 100$	0.014	0.999	0.011	1	0.005	1
$n = 200$	1.544	0.910	2.272	0.911	2.950	0.913

The oracle value for the deviance is 0.9, and those for the MSE are in brackets.

Table 2. See the complete results and details about its implementation in Section 3.2 of the supporting information.

We can appreciate as the BIC criterion helps to reduce the inclusion of noisy covariates for $n > p$. Besides, this corrects a bit the overestimation in this framework, although this is not removed. In contrast, its performance is pretty bad for $p \geq n$. The BIC criterion strongly overfits the results in this last cases: this adds more noise to the model and produces more overestimation than the LASSO.min. See an analysis of this topic in Giraud *et al*. (2012).

Next, we analyse the results of the dependence by blocks context. In case of dependence, it is expected for a 'smart' algorithm to be capable of selecting a portion of relevant covariates and explaining the remaining ones making use of the existing correlation structure. The subset of S which is really necessary to explain this type of models is denote as 'effective covariates'. These can be calculated measuring how many terms are necessary to explain a certain percentage of Σ_S variability, being Σ_S the submatrix of Σ considering the elements of S. This number is inversely proportional to the dependence strength. For example, to explain the 90–95%, we found that for the Scenario 2 with $\rho = 0.5$, there are needed about 12–14 covariates taking $s = 15$ and about 16–18 for the case of $s = 20$. In contrast, only 10 are necessary in Scenario 2 with $\rho = 0.9$. Complete calculation for the different simulation scenarios is displayed in Section 4 of the supporting information. Again, the LASSO.min presents some difficulties for an efficient recovery.

A summary of the results for the Scenario 2 with $\rho = 0.5$ is displayed in Table 3 and Figure 4, while for the Scenario 2 with $\rho = 0.9$ is showed in Table 4 and Figure 5. If we consider only $s = 10$ important explanatory variables, its behaviour is quite similar to the Scenario 1. Besides, in both scenarios with $s = 10$, LASSO.min almost recovers the complete set S, even for $n = 25$ although its proper recovery is guaranteed from $n = 50$. However, more noise is included in this last.

In contrast, the situation is different if we simulate with $s = 15$ or $s = 20$ relevant covariates. Then, the LASSO.min does not tend to recover the covariates of S, not even for values of

Table 3. *Summary of the LASSO.min results for Scenario 2 with $\rho = 0.5$.*

	$s = 10$		$s = 15$		$s = 20$	
	MSE (0.556)	*% Dev*	*MSE (1.389)*	*% Dev*	*MSE (2.222)*	*% Dev*
$n = 50$	0.438	0.956	1.095	0.956	1.752	0.956
$n = 100$	0.495	0.951	1.238	0.951	1.981	0.951
$n = 200$	0.523	0.951	1.307	0.950	2.091	0.951

The oracle value for the deviance is 0.9, and those for the MSE are in brackets.

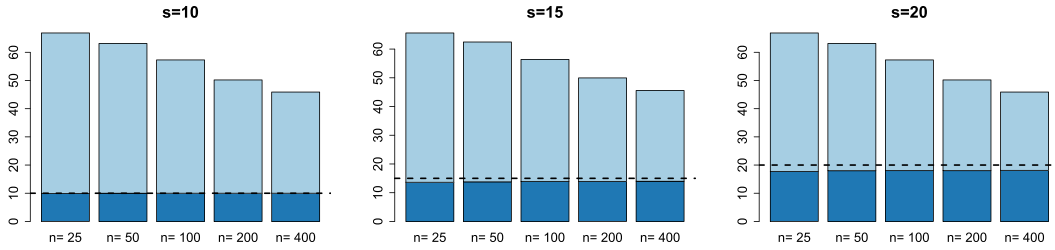

Figure 4. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.min in Scenario 2 with $\rho = 0.5$. The dashed line marks the s value*

Table 4. *Summary of the LASSO.min results for Scenario 2 with $\rho = 0.9$.*

	$s = 10$		$s = 15$		$s = 20$	
	MSE (1)	*% Dev*	*MSE (2.5)*	*% Dev*	*MSE (4)*	*% Dev*
$n = 50$	0.784	0.926	1.96	0.925	3.137	0.926
$n = 100$	0.888	0.918	2.22	0.918	3.551	0.918
$n = 200$	0.939	0.913	2.347	0.913	3.756	0.913

The oracle value for the deviance is 0.9 and those for the MSE are in brackets.

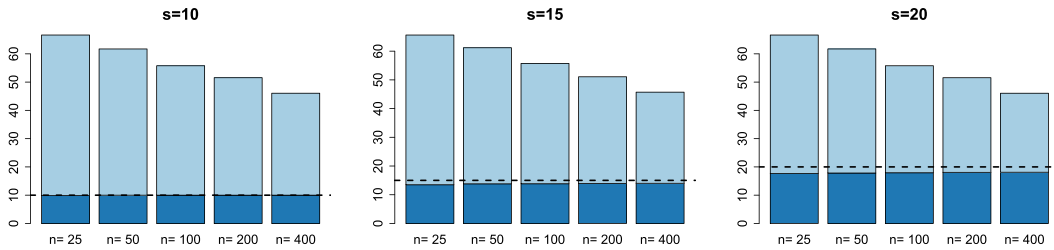

Figure 5. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.min in Scenario 2 with $\rho = 0.9$. The dashed line marks the s value*

*n*verifying $n \geq p$ as well as conditions (10) and (9). See Section 2 of the supporting information for more information. However, this selects more than the effective number of covariates. It seems the LASSO.min tries to recover the set S but, due to the presence of spurious correlations, this chooses randomly between two highly correlated important covariates. It can be appreciated by means of Tables 30 and 31 in Section 6.1 of the supporting information that the 10 first

covariates are selected with high probability, near 1. However, due to the confusion phenomenon, some of them are interchanged by a representative one. The following $s - 10$ relevant variables have a lower selection rate, and there are some irrelevant ones selected a larger number of times, adding quite noise to the model. This inconvenient seems not to be overcome increasing the number of samples n. Again, the LASSO.min keeps overestimating its results as we can see by the percentage of explained deviance and the MSE.

A similar behaviour is observed selecting the λ value greater than $\lambda^{\min}$ (see LASSO.1se) and by means of the BIC criterion, although this last tends to add more noise. Furthermore, for $p \geq n$, the LASSO.BIC is not capable to deal with the dependence structure, selecting less than s covariates in some cases. This translates in overestimation. In contrast, the LASSO.1se does not improve too much the LASSO.min performance in this scenario. The same phenomenon has been observed for greater values than λ^{1se}, see Section 3.2 of the supporting information. Results for the LASSO.BIC and LASSO.1se algorithms are collected in Sections 3.1 and 5 of the supporting information, respectively.

Finally, we study the results of the Toeplitz covariance structure by means of the Scenario 3.a, where the relevant covariates are the first $s = 15$ (Table 5 and Figure 6), and the Scenario 3.b, where there are only $s = 10$ important variables placed every 10 sites (Table 5 and Figure 7).

Interpreting their results, we see that the LASSO.min procedure recovers the important set of covariates for $\rho = 0.5$, taking a value of $n = 100, 200, 400$ verifying the consistent condition, in both cases. Nevertheless, this exceeds the number of efficient covariates in Scenario 3.a for $\rho = 0.9$, because with 10 covariates it is explained the 98% of variability. Moreover, this algorithm returns to include many pointless covariates in the model and overestimates the prediction accuracy.

Table 5. *Summary of the LASSO.min results for Scenarios 3.a and 3.b.*

	Scenario 3.a				*Scenario 3.b*			
	$\rho = 0.5$		$\rho = 0.9$		$\rho = 0.5$		$\rho = 0.9$	
	MSE (1.139)	*% Dev*	*MSE (3.807)*	*% Dev*	*MSE (0.278)*	*% Dev*	*MSE (0.53)*	*% Dev*
$n = 50$	0.19	0.983	1.894	0.950	0.034	0.987	0.147	0.971
$n = 100$	0.546	0.951	2.815	0.928	0.123	0.955	0.309	0.94
$n = 200$	0.825	0.927	3.302	0.916	0.195	0.929	0.417	0.920

The oracle value for the deviance is 0.9, and those for the MSE are in brackets.

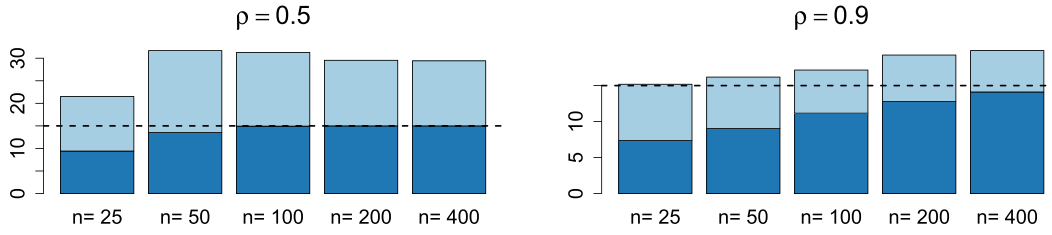

Figure 6. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.min in Scenario 3.a. The dashed line marks the $s = 15$ value*

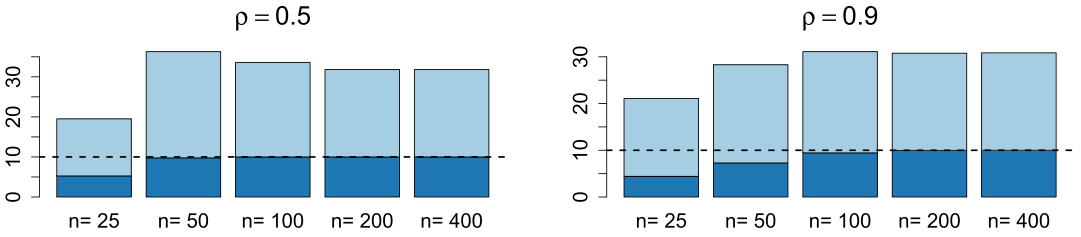

Figure 7. *Number of important covariates (dark area) and noisy ones (soft area) selected by the LASSO.min in Scenario 3.b. The dashed line marks the* $s = 10$ *value*

The LASSO.1se behaviour is quite similar to the LASSO.min, although the LASSO.1se reduces the noise and corrects a bit the overestimation. Taking larger values than λ^{1se} the performance of the LASSO is even better. However, it seems to be quite difficult to establish a common rule for the optimal λ selection in the different frameworks of Scenarios 3.a and 3.b. Specifically, the Scenario 3.b with $\rho = 0.9$ seems to need a larger value than the remaining ones. See Section 3.2 of the supporting information for a graphical comparison.

In contrast, the BIC adjustment selects less covariates, tending $|\hat{S}| = s$ as n increases, except in Scenario 3.b with $\rho = 0.9$. For Scenario 3.a this selection procedure tends to recover S for a large enough value of n, avoiding irrelevant information. Nevertheless, something different happens for Scenario 3.b. In this last, the algorithm interchanges relevant covariates with irrelevant ones quite correlated with the ones of S. In spite of this, the algorithm includes with high probability representatives of the $s = 10$ relevant covariates, especially for $\rho = 0.9$, capturing the essential information. This can be appreciated in Section 6.2.2 of the supporting information. Moreover, this procedure corrects a bit the overestimation of the CV technique. Results are collected in the supporting information: by means of figures in Sections 3.1 and 6.2, and one table in Section 3.1.

4 Evolution of the Lasso in the Last Years and Alternatives

Once the LASSO and its inconveniences have been displayed, we want to compare its performance with other approaches. Then, we test different methodologies designed to select the relevant information and adjust a regression model in the $p > n$ framework. Nevertheless, it is impossible to include all the existing algorithms here. Instead, we attempted to collect the most relevant ones, providing a summary of the most used methodologies nowadays.

Methods proposed to alleviate the limitations of the LASSO algorithm are based in a wide range of different philosophies. Some of them opt to add a second selection step after solving the LASSO problem, such as the relaxed LASSO (Meinshausen, 2007) or thresholded LASSO (Lounici (2008), Zhou (2010), Van de Geer *et al*., 2011), while other alternatives are focused on giving different weights to the covariates proportional to their importance, as the adaptive LASSO (Zou, 2006; Huang *et al*., 2008; Van de Geer *et al*., 2011). Others pay attention to the group structure of the sparse vector β when this exists, like the grouped LASSO procedure (Yuan & Lin, 2006) or the fused LASSO (Tibshirani *et al*., 2005).

The resampling or iterative procedures are other approaches which make use of subsampling or computational power, algorithms like boLASSO (Bach, 2008), stability selection with randomised LASSO (Meinshausen & Bühlmann, 2010), the random LASSO (Wang *et al*., 2011), the scaled LASSO (Sun & Zhang, 2012) or the combination of classic estimators with variable selection diagnostics measures (Nan & Yang, 2014), among others, are based on this idea. Furthermore, more recent techniques like the Knockoff filter (Barber & Candès, 2015),

Candes *et al*. (2018) or SLOPE (Bogdan *et al*., 2015) have been introduced to control some measures of the type I error. One drawback of the famous Knockoff filter is that, to the best of our knowledge, this is not yet available for the $p > n$ case.

There are other alternatives, which modify the constraints of the LASSO problem (6) in order to achieve better estimators of β, like the elastic net (Zou & Hastie, 2005) or the Dantzig selector (Candes & Tao, 2007). Moreover, other different alternatives have been developed recently, such as the Elem-OLS Estimator (Yang *et al*., 2014), the LASSO-Zero (Descloux & Sardy, 2021), the spike-and-slab LASSO (Ročková & George, 2018) or Bayesian approaches (see, e.g. Castillo *et al*., 2015 or Bhadra *et al*., 2019).

Quoted Descloux & Sardy (2021): although differing in their purposes and performance, the general idea underlying these procedures remains the same, namely to avoid overfitting by finding a trade-off between the fit $y - X\beta$ and some measure of the model complexity.

Along the many papers, we have found that a modest classification of the different proposals can be performed, although in this classification some of the procedures does not only fit in a single class. These categories are

- **Weighted LASSO:** weighted versions of the LASSO which attach the particular importance of each covariate for a suitable selection of the weights. Joint with iteration, this modification allows a reduction of the bias.
- **Resampling LASSO procedures:** mix of the LASSO adjustment with resampling procedures for randomising the covariates selection process to reduce unavoidable random noise.
- **Thresholded versions of the LASSO:** a second thresholding step in the covariates selection is implemented in order to reduce the irrelevant ones.
- **Alternatives to the LASSO:** procedures with different nature and aims designed to solve the LASSO drawbacks.

This extensive list of procedures makes noticeable the impact the LASSO has nowadays. A brief summary is displayed in Table 6.

Owing to the computational cost required for the resampling LASSO procedures, such as the boLASSO of Bach (2008) or the random LASSO algorithm (Wang *et al*., 2011), these algorithms are too slow. Even for small values of p, we found that the computational costs were high. For this reason, they are excluded for the comparative analysis studio. The LASSO-Zero technique of Descloux & Sardy (2021) suffers from the same issue, so it is excluded too.

Other problem springs up for the thresholded versions of the LASSO. In this case, we noticed that the complexity of finding a correct threshold is similar to the one of obtaining the optimal value of λ for the LASSO adjustment. In both cases, we would need to know in advance the dispersion of the error σ^2, which is usually impossible in practice. Then, procedures as the thresholded LASSO algorithm of Lounici (2008) are shut out to avoid more complications in the adjustment.

One method in the middle of both groups is the stability selection procedure proposed by Meinshausen & Bühlmann (2010). This methodology pays attention to the probability of each covariate to be selected. Only the covariates with probability greater than a fixed threshold $q \in (0, 1)$ are added to the final model. Although the authors recommend to take $q \in (0.6, 0.9)$, we have observed in practice that a proper choice of the threshold value seems to depend on the sample size considered, n, as well as the sparsity of the vector β. Besides, one more tuning parameter is needed: the bound for expected number of false positives. See Dezeure *et al*. (2015) for more practical details. For all these reasons, this approach is not included in the comparison neither.

Table 6. *Formulated problems to estimate the β vector in a linear regression high dimensional framework ($p > n$), showing if the optimisation problems are convex (✓) or not (×). Their main advantages in comparison with the LASSO are displayed in column (**PROS**) and it is showed if they are a weighted version of the LASSO (●) or alternatives (▲).*

PROBLEM FORMULATION		PROS
▲ **Best subset selection** – Beale et al. (1967), Hocking and Leslie (1967) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\mathbf{1}_{\beta_j\neq 0}\right\}$	×	*Better selection*
LASSO – Tibshirani (1996) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\lvert\beta_j\rvert\right\}$	✓	–
▲ **SCAD** – Fan (1997) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+p_\lambda(\beta)\right\}$ with $p_\lambda(\beta)=\begin{cases}\lambda\lvert\beta\rvert, & \text{if } \lvert\beta\rvert\leq\lambda,\\ \dfrac{2a\lambda\lvert\beta\rvert-\beta^2-\lambda^2}{2(a-1)}, & \text{if } \lambda<\lvert\beta\rvert\leq a\lambda \quad (a>2)\\ \dfrac{\lambda^2(a+1)}{2}, & \text{otherwise.}\end{cases}$	×	*Better selection* *Bias reduction*
Basis Pursuit Denoising – Chen et al. (2001) $\min_{\beta}\lVert\beta\rVert_1 \quad \text{subject to } \lVert y-X\beta\rVert_2\leq\theta$	×	–
▲ **Elastic Net** – Zou and Hastie (2005) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left(\alpha\lvert\beta_j\rvert+(1-\alpha)\beta_j^2\right)\right\}$ with $\alpha\in(0,1)$	✓	*Better prediction* *Possible selection of more than n covariates ($p > n$)*
▲ **Fused LASSO** – Tibshirani et al. (2005) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda_1\sum_{j=1}^{p}\lvert\beta_j\rvert+\lambda_2\sum_{j=2}^{p}\lvert\beta_j-\beta_{j-1}\rvert\right\}$	✓	*Ordered structure*
● **Adaptive LASSO** – Zou (2006) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}w_j\lvert\beta_j\rvert\right\}$ (taking $w_j=1/\lvert\hat{\beta}_j^{RR}\rvert^q$ where $\hat{\beta}^{RR}$ is the ridge estimator (Hoerl and Kennard (1970)) and $q\geq 1$)	✓	*Better selection* *Bias reduction*
▲ **Group LASSO** – Yuan and Lin (2006) $\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{k=1}^{K}X_k\beta_k\right)^2+\lambda\sum_{k=1}^{K}\lVert\beta_k\rVert_{Z_k}\right\}$ with $\lVert w\rVert_{Z_k}=(w^\top Z_k w)^{1/2}$ (Z_k are kernel matrices of the functional space induced by the kth factor)	✓	*Group structure*

▲ **Dantzig selector** – Candes and Tao (2007) $\min_{\beta}\|\beta\|_1 \quad \text{subject to } \|X^\top r\|_\infty \le \lambda_p \cdot \sigma$ $(\text{with } \|X^\top r\|_\infty := \sup_{1\le j\le p} \|(X^\top r)_j\| \text{ and } r = y - X\beta)$	×	*Consistent to orthogonal transformations*
▲ **Relaxed LASSO** – Meinshausen (2007) $\min_{\beta}\left\{n^{-1}\sum_{i=1}^{n}\left(y_i - x_i^\top\{\beta\cdot \mathbf{1}_{\mathcal{M}_\lambda}\}\right)^2 + \phi\lambda\|\beta\|_1\right\} \text{ with } \phi\in(0,1]$	✓	*Faster convergence rates* *More accurate predictions*
▲ **Square-root LASSO** – Belloni et al. (2011) $\min_{\beta}\left\{\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right]^{1/2} + \lambda\sum_{j=1}^{p}\|\beta_j\|\right\}$	✓	*It is not needed to known σ to obtain an optimal λ*
▲ **Scaled LASSO** – Sun and Zhang (2012) $\hat\sigma \leftarrow \|y - X\hat\beta^{old}\|_2/n^{1/2}, \quad \lambda \leftarrow \hat\sigma\lambda_0$ $\hat\beta^{new} = \min_{\beta}\begin{cases} x_j^\top(y - X\hat\beta)/n = \lambda\text{sign}(\hat\beta_j), & \hat\beta_j \neq 0, \\ x_j^\top(y - X\hat\beta)/n \in \lambda[-1,1], & \hat\beta_j = 0. \end{cases}$ $\hat\beta \leftarrow \hat\beta^{new},\ L_\lambda(\hat\beta^{new}) \le L_\lambda(\hat\beta^{old})$ $\left(\text{where } L_\lambda(\beta) = \frac{\|y - X\beta\|_2^2}{2n} + \lambda\sum_{j=1}^{p}\|\beta_j\|\right)$	✓	*Simultaneous estimation of σ and β*
● **SLOPE** – Bogdan et al. (2015) $\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \sum_{j=1}^{p}\lambda_j|\beta|_{(j)}\right\}$	✓	*Control of the False Discovery Rate FDR*
▲ **Debiased LASSO** – Javanmard and Montanari (2018) $\hat\beta^{debiased} = \hat\beta^{L_1} + \frac{1}{n}MX^\top(y - X\hat\beta^{L_1}) \sim N(\beta, \sigma^2/n)$ with $M = (m_1,\ldots,m_p)^\top \in \mathbb{R}^{p\times p}$, where each $m_i \in \mathbb{R}^p$ is the solution of $\min_{m} m^\top\hat\Sigma m \quad \text{subject to } \|\hat\Sigma m - e_i\|_\infty \le \mu$ ($e_i \in \mathbb{R}^p$ is a standard unit vector, $\hat\Sigma = (X^\top X)/n$ and μ a constraint)	✓	*Characterization of the probability distribution for the estimator of β ($\hat\beta^{debiased}$)*
▲ **LASSO-Zero** – Descloux and Sardy (2021) $\min_{\beta}\|\beta\|_1 + \|\gamma\|_1$ subject to $y = \tilde X\beta + G\gamma$ ($G \in \mathbb{R}^{n\times q}$ a noise dictionary and $\tilde X = (X\|G)$)	×	*Excellent trade-off between high true positive rate (TPR) and low false discovery rate (FDR)*

Next, we display the results of the simulation study comparing the performance of different procedures which have showed suitable properties.

4.1 Comparison with Other Approaches by Means of a Simulation Study

As we advanced above, we have made a selection of the most relevant methodologies in terms of good qualities as well as reasonable computational time. Because of the nature of the

simulation scenarios of Section 3, we have discarded some procedures due to their unsuitable characteristics. Moreover, we took into account to choose methodologies with available code in R (R Core Team, 2019), so everyone can make use of them. We have chosen libraries which provide us with enough resources to fit the models, selecting those created for the author's methodology or the most recently updated option in case of doubt. This selection resulted in:

- **LASSO**: `glmnet` of Friedman *et al*. (2010), last update February 21, 2021.
- **SCAD**: `ncvreg` of Breheny & Huang (2011), last update July 9, 2020.
- **Adaptive LASSO (AdapL)**: `glmnet` of Friedman *et al*. (2010), last update February 21, 2021.
- **Dantzig selector (Dant)**: `flare` of Li *et al*. (2019), last update December 1, 2020.
- **Relaxed LASSO (RelaxL)**: `relaxo` of Meinshausen (2012), last update February 20, 2015.
- **Square-root LASSO (SqrtL)**: `flare` of Li *et al*. (2019), last update December 1, 2020.
- **Scaled LASSO (ScalL)**: `scalreg` of Sun (2019), last update January 25, 2019.

Now, we analyse the performance of these algorithms in comparison with the LASSO ones (Section 3.1), following the scheme introduced in Section 3. Details of its tuning parameters selection is given in Section 3.1 of the supporting information. Furthermore, we compare their results with an innovative procedure which has proved its efficiency, the distance correlation algorithm for variable selection (DC.VS) of Febrero-Bande *et al*. (2019). This makes use of the correlation distance (Székely *et al*., 2007; Szekely & Rizzo, 2017) to implement an iterative procedure (forward) deciding in each step which covariate enters the regression model. For this purpose, this borrows the idea of the LARS algorithm of Efron *et al*. (2004). A complete explanation of the algorithm is given in Febrero-Bande *et al*. (2019). Because of its selection improvements, this methodology is tested instead of the usual forward selection. A comparison between the forward selection and the LASSO can be found in Hastie *et al*. (2017). For this purpose, the library `fda.usc` (Febrero-Bande & Oviedo de la Fuente, 2012) is employed. The complete simulation results are provided in the supporting information.

We start studying the easiest framework: the orthogonal design (Scenario 1). In case of simulating under independence between covariates we can see that any of the studied algorithms perform better than the LASSO.min. These obtain good results searching for the s relevant covariates when $p > n$, and they seem to be able to recover the set S for a large enough value of n (see Figure 8). Besides, all of them add less noise to the model and do not overestimate too much the predictions, as the LASSO.min does. See Table 7 for a brief comparison. Nevertheless, the only method which selects the complete set S without including any noise to the model, for a large enough value of n, is the AdapL.1se algorithm. This last performs incredibly well. The performance of the LASSO.BIC and RelaxL is also remarkable, although the first one only outperforms the LASSO.min for values of $n > p$. The Dant achieves good results in terms of avoiding noise too; however, its convergence to the set S seems more slow.

Once we have seen that the proposed alternatives to the LASSO.min improve the results when there is not a correlation structure between covariates, we want to test their performance under dependence. The first considered model with dependence is the dependence by blocks correlation structure (Scenario 2), simulating a correlation structure of value ρ every ten places. In Section 3.1 we saw that the LASSO.min does not select a representative subset of S formed by a bunch of efficient covariates as expected, instead this always tries to recover the complete set adding a lot of noisy ones, which translates in overestimation. A comparative example of all algorithms performance in this scenario, for $s = 15$ and $n = 400$, is displayed

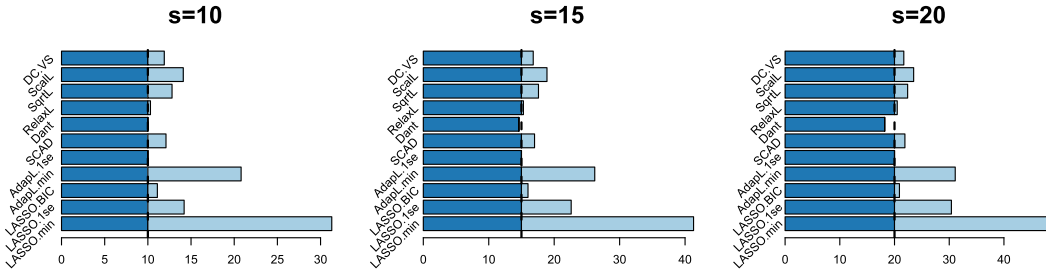

Figure 8. *Comparison of the important covariates number (dark area) and noisy ones (soft area) for* $n = 400$ *in Scenario 1. The dashed line marks the* s *value*

Table 7. *Comparison of all proposed algorithms for Scenario 1 taking* $n = 400$ *and* $s = 15$.

Scenario	$\lvert\hat{S} \cap S\rvert$ *(15)*	$\lvert\hat{S} \setminus S\rvert$	$\lvert\hat{S}\rvert$	*MSE (2.604)*	*% Dev (0.9)*
LASSO.min	15	26.3	41.3	2.091	0.919
LASSO.1se	15	7.6	22.6	2.300	0.911
LASSO.BIC	15	1	16	2.443	0.905
AdapL.min	15	11.2	26.2	2.235	0.913
AdapL.1se	15	0	15	2.491	0.903
SCAD	15	2	17	2.425	0.906
Dant	14.6	0	14.6	2.976	0.885
RelaxL	15	0.3	15.3	2.478	0.904
SqrtL	15	2.6	17.6	2.403	0.907
ScalL	15	3.9	18.9	2.371	0.908
DC.VS	15	1.8	16.8	2.421	0.906

The oracle values are in brackets.

in Table 8 taking $\rho = 0.5$ (Scenario 2 with $\rho = 0.5$) and in Table 8 simulating with $\rho = 0.9$ (Scenario 2 with $\rho = 0.9$). Visual examples are showed in Figure 9 and Figure 10 respectively.

The LASSO.1se, LASSO.BIC, Dant algorithm and the SqrtL suffer from the same issue. We can see as these algorithms are not able of interpreting the structure of the data and select almost the p covariates in some cases. Here, the Dant mimics the performance of the LASSO.min when there exists the same correlation between important covariates and noisy ones as in the $s = 15$ and $s = 20$ framework. In these situations, this algorithm recovers 10 out of the s relevant variables but then, this is unable to distinguish between the rest of important covariates and noise. This seems due to the dependence by blocks structure: important covariates already selected by the model have the same correlation with the rest of relevant ones as with noisy covariates placed every ten locations. Then, the LASSO.1se, LASSO.BIC, Dant and SqrtL do not overcome the spurious correlations phenomenon and tend to select too many covariates. Besides, these procedures perform even worse than the LASSO.min in both frameworks of the Scenario 2 adding more noise and overestimating the prediction accuracy. As a result, we can conclude that the four methods are not suitable when we have the same strong correlation between remaining important covariates and noisy ones.

In contrast, the rest of alternatives seem to perform better, trying to select a representative subset of length 10 approximately. However, not all the remaining procedures select a representative subset between the s important variables. Instead, the majority change relevant covariates for noisy ones quite correlated with the previous ones, covering the complete set S. Into words,

Table 8. *Comparison of all proposed algorithms for Scenario 2 with $\rho = 0.5$ and with $\rho = 0.9$ taking $n = 400$ and $s = 15$.*

	$\rho = 0.5$					$\rho = 0.9$				
	$\lvert\hat{S} \cap S\rvert$ *(15)*	$\lvert\hat{S} \setminus S\rvert$	$\lvert\hat{S}\rvert$	*MSE (1.389)*	*% Dev (0.9)*	$\lvert\hat{S} \cap S\rvert$ *(15)*	$\lvert\hat{S} \setminus S\rvert$	$\lvert\hat{S}\rvert$	*MSE (2.5)*	*% Dev (0.9)*
LASSO.min	14	31.6	45.6	1.346	0.949	14	31.7	45.7	2.42	0.912
LASSO.1se	13.8	32.1	45.9	1.346	0.949	13.7	31.6	45.2	2.420	0.912
LASSO.BIC	14.4	40.3	54.7	1.346	0.949	14.4	41.7	56.1	2.420	0.912
AdapL.min	10	0	10	1.346	0.949	10	0	10	2.423	0.912
AdapL.1se	10	0	10	1.346	0.949	9.9	0.1	10	2.423	0.912
SCAD	10.1	0	10.1	1.346	0.949	10.1	0	10.1	2.423	0.912
Dant	11.2	50.4	61.6	4.968	0.811	11.1	50.2	61.3	6.013	0.781
RelaxL	3.7	8	11.7	1.377	0.947	4.5	7.2	11.7	2.438	0.911
SqrtL	15	85	100	1.346	0.949	15	84.9	100	2.42	0.912
ScalL	3.4	7.1	10.4	1.374	0.948	4	6.5	10.4	2.567	0.906
DC.VS	3.4	6.6	10	1.346	0.949	3.8	6.2	10	2.423	0.912

The oracle values are in brackets.

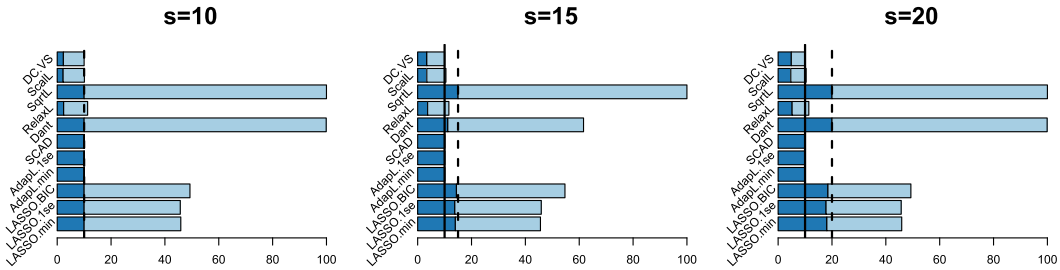

Figure 9. *Comparison of the important covariates number (dark area) and noisy ones (soft area) for $n = 400$ in Scenario 2 with $\rho = 0.5$. The dashed line marks the considered s value while the continuous line where $s = 10$*

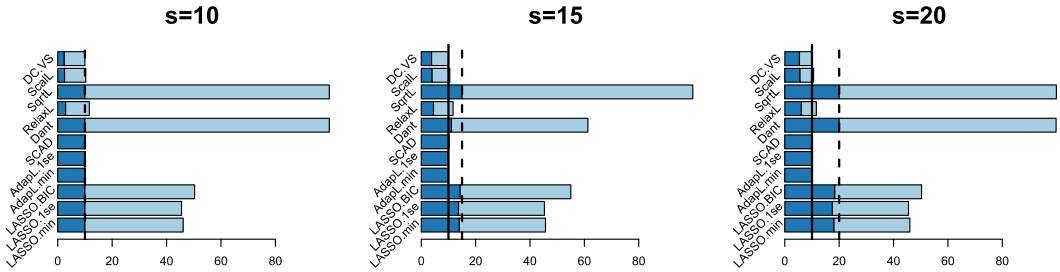

Figure 10. *Comparison of the important covariates number (dark area) and noisy ones (soft area) for $n = 400$ in Scenario 2 with $\rho = 0.9$. The dashed line marks the considered s value while the continuous line where $s = 10$*

if a procedure chooses a noise covariate it is expected that this last is a representative of some not included relevant covariate to achieve a good explanation of the data. We can see a proof of this phenomenon for the RelaxL, ScalL and DC.VS in Section 6.1 of the supporting information. Only the AdapL.min, AdapL.1se and the SCAD algorithms seem to behave properly in this sense, recovering 10 elements of the set S. All these methodologies correct a bit the overestimation produced by the LASSO.min.

Finally, we simulate under the Toeplitz covariance structure of Scenario 3. We consider a first scenario, where the relevant covariates are located in the first $s = 15$ placements (Scenario 3.a), and a second one, where we simulate only $s = 10$ important variables and they are placed every ten sites (Scenario 3.b). Hence, we expect for the Scenario 3.a to obtain a representative subset of the set S, with cardinal less than s as we explained in Section 3.1. In particular, when the correlation between covariates is strong, as for $\rho = 0.9$. It is owing to the fact that we have, in this scenario, several relevant covariates with a representative correlation between them. Roughly speaking, because of the Toeplitz covariance structure, one variable could be 'easily' explained by others in its neighbourhood. This translates in the possibility of interchanging last variables of S with nearby ones. Then, for $\rho = 0.5$, because $0.5^5 \leq 0.05$, we consider as good representatives those covariates which distance is less than 4 to some position of S. When $\rho = 0.9$ this distance is enlarged and there are many more possibilities. In contrast, simulating the Scenario 3.b, we would expect the algorithm to select all the 10 relevant covariates in the best case or a representative subset following this criteria.

For the Scenario 3.a we can appreciate in Figure 11 a similar phenomenon as the one observed in Scenario 2. This is translated in the existence of algorithms which try to recover the complete set S, like the LASSO.min, the AdapL.min, the SCAD, the RelaxL, the SqrtL or the ScalL. We could also include the LASSO.1se, the LASSO.BIC and the DC.VS to the list

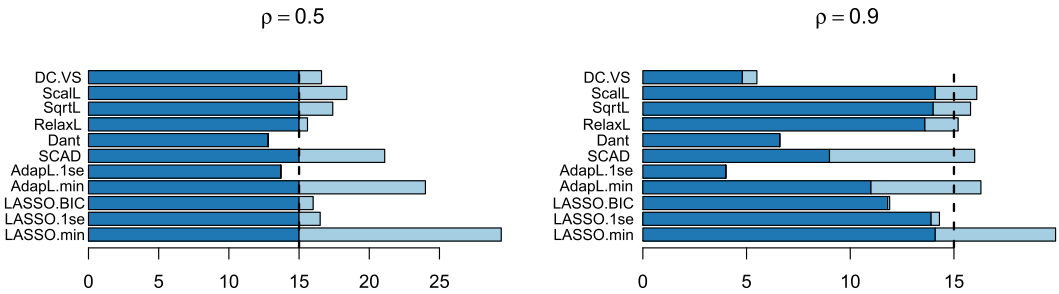

Figure 11. *Comparison of the important covariates number (dark area) and noisy ones (soft area) for* $n = 400$ *in Scenario 3. a. The dashed line marks the s value*

Table 9. *Comparison of all proposed algorithms for Scenario 3.a taking* $n = 400$ *with* $\rho = 0.5$ *and* $\rho = 0.9$.

	$\rho = 0.5$					$\rho = 0.9$				
	$\lvert\hat{S}\cap S\rvert$ *(15)*	$\lvert\hat{S}\setminus S\rvert$	$\lvert\hat{S}\rvert$	*MSE (1.139)*	*% Dev (0.9)*	$\lvert\hat{S}\cap S\rvert$ *(15)*	$\lvert\hat{S}\setminus S\rvert$	$\lvert\boxtimes\rvert$	*MSE (3.807)*	*% Dev (0.9)*
LASSO.min	15	14.4	29.4	0.972	0.914	14.1	5.8	19.9	3.620	0.908
LASSO.1se	15	1.5	16.5	1.063	0.906	13.9	0.4	14.3	3.762	0.905
LASSO.BIC	15	1	16	1.066	0.906	11.8	0.1	11.9	3.824	0.903
AdapL.min	15	9	24	0.995	0.912	11	5.3	16.3	3.593	0.909
AdapL.1se	13.7	0	13.7	1.195	0.894	4	0	4	4.543	0.885
SCAD	15	6.1	21.1	1.016	0.910	9.1	7	16	3.658	0.907
Dant	12.8	0	12.8	1.443	0.873	6.6	0	6.6	4.92	0.875
RelaxL	15	0.6	15.6	1.078	0.905	13.6	1.6	15.2	3.728	0.906
SqrtL	15	2.4	17.4	1.053	0.907	14.1	1.8	15.9	3.717	0.906
ScalL	15	3.4	18.4	1.039	0.908	14.1	2	16.1	3.701	0.906
DC.VS	15	1.6	16.6	1.061	0.906	4.8	0.8	5.6	4.285	0.891

The oracle values are in brackets.

for the $\rho = 0.5$ case. The rest of algorithms, the AdapL.1se as well as the Dant algorithm, search always for a representative subset without including noise. A summary of their performance is displayed In Table 9. Taking $\rho = 0.5$ we appreciate that the AdapL.1se and the Dant are the only procedures which select the number of efficient covariates needed to explain, at least, the 90% of the covariance. A similar behaviour could be considered for the DC.VS but this adds more noise and selects more than $s = 15$ covariates for $\rho = 0.5$. In Section 6.2.1 of the supporting information the percentage of times the relevant covariates are selected for these algorithms is displayed.

Studying the provided results for $\rho = 0.9$ (Table 9) we can claim that DC.VS achieves the best results in terms of prediction when the correlation is large, but this pays the price of including more irrelevant information than the AdapL.1se or the Dant. In contrast, when $\rho = 0.5$, the selection of covariates made by the DC.VS results in an overestimation of the model. Finally, if we compare the AdapL.1se with the Dant algorithm results, we see that it seems like the first one obtains a better trade-off between selection of covariates and estimation. We notice that, for $\rho = 0.9$, the AdapL.1se selects less covariates of S but achieves a better performance in terms of explanation of the data. All these three approaches select less than $s = 15$ covariates for $\rho = 0.9$ but a number large enough to guarantee a good explanation of the covariance. See Section 4 of the supporting information for more details.

At this point, it is interesting to notice that the Dant performs correctly in this dependence context in comparison with the Scenario 2 framework. Now, this is able to recover a representative subset of Swithout adding noise to the model. This phenomenon could be explained taking into account that in Scenario 3.a we have not too many noisy covariates highly correlated with the ones of S, especially for $\rho = 0.5$. Only those in the neighbourhood of the 15^{th} could be a threat. However, in the Scenario 2 we had covariates correlated ten by ten and, as a result, every relevant covariate is correlated with 8 irrelevant ones at least. In contrast, the SqrtL keeps its bad behaviour and the SCAD algorithm starts to perform poorly. This last brings out the fact that the SCAD procedure suffers when all the covariates are correlated among them. This happens when the important covariates are close in location as in the case of Scenario 3.a, however, when these are more scattered, like in Scenario 3.b, the algorithm performs better.

Next, we compare the results obtained for the Scenario 3.b. An example is displayed in Figure 12 and the rest of results are provided in the supporting information. Simulating for $\rho = 0.5$ we observe that all the proposed algorithms outperform the LASSO.min results. At first sight, it may seem that the LASSO.BIC does not perform properly, however, this selects a representative subset of Schanging important covariates for correlated ones. The remained procedures try to recover the complete set S as it is expected taking into account that the

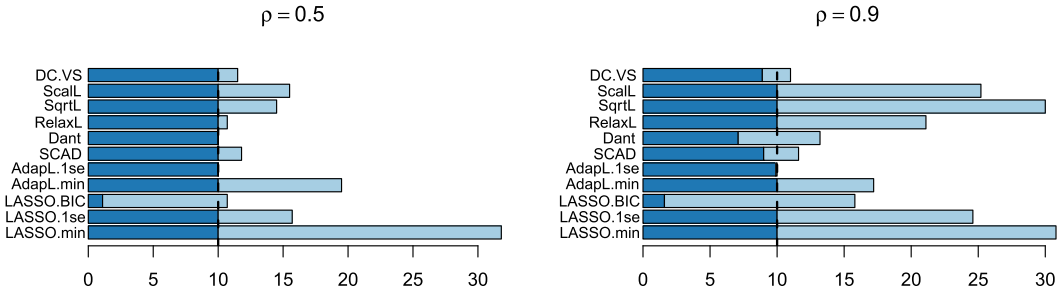

Figure 12. *Comparison of the important covariates number (dark area) and noisy ones (soft area) for* $n = 400$ *in Scenario 3. b. The dashed line marks the s value*

number of efficient covariates is 10 now. Nevertheless, when $\rho = 0.9$, some drawbacks come up. Some of them interchange relevant covariates with irrelevant ones quite correlated with these. This is product of the strong correlation structure of the Toeplitz covariance. These are the LASSO.BIC, the SCAD and the DC.VS algorithms. Maybe, we can include in this last group the Dant, although it is doubtful. Section 6.2.2 of the supporting information collects the percentage of times a representative of the 10 relevant covariates enters the model. Other procedures, like the RelaxL, the SqrtL or the ScalL add unnecessary noise, overestimating the model. Even the LASSO.BIC could be included in this list. Only one algorithm is almost capable of recovering the s variables without adding more noise to the model, this is the AdapL.1se algorithm. All the alternatives correct the overestimation in the prediction made by the LASSO.min though.

Eventually, it is important to highlight that computational time varies from one methodology to another. This is because the way these are implemented and their nature. Some of them have the cross-validation scheme integrated in the employed R library considered. For example, the famous library `glmnet`(Friedman, *et al*. 2010) has implemented an optimal cross-validation algorithm in Fortran code, which improves the computational cost of the R language. This results in a quite competitive computational time for the LASSO and AdapL adjustments. The library `ncvreg` (Breheny & Huang, 2011) for the SCAD has utilities for carrying out cross-validation also. In contrast, other methods have not implemented this scheme as for the `flare` library used for the Dant and the SqrtL. In this case, it is needed to program the cross-validation scheme, resulting in higher computational cost. Besides, the SqrtL pays the price of higher computational time required to be able to select an optimal λ without knowing σ. As a result, SqrtL is the slowest algorithm of the study. Other different procedure, the ScalL, is fitted by means of an iterative algorithm rather than a cross-validation process. Then, its time complexity depends on convergence criteria. Last, the DC.VS of Febrero-Bande *et al*. (2019) has a different nature to the previous ones. This apply a special forward selection scheme recalculating the distance matrices between samples on every step to obtain the correlation distance coefficients (Székely *et al*., 2007; Szekely & Rizzo, 2017), increasing its time complexity in terms on n.

5 Discussion: Some Guidance about LASSO

Currently, the LASSO regression keeps being a broadly employed covariates selection technique. Despite its several advantages, some strict necessary requirements could make difficult a correct performance of this methodology, as it was explained in Section 2. As we argued at the beginning of the document, there are no global recommendations about the use of the LASSO in terms of the nature of the data under dependence or when some of these conditions do not hold. With the aim of shed light on this topic, we have analysed the LASSO drawbacks, studying modifications and alternatives to overcome these. Besides, an extensive simulation study has been carried out to illustrate the behaviour of LASSO in the best possible scenario and in trickier ones carefully chosen (Section 3), comparing this with the one of recent modifications and other alternatives (Section 4). In view of the results, we give next some guidance on how to choose a proper covariates selector according to the nature of data. Results are summarised in Table 10.

We have seen that even in scenarios where there no exists dependence, the LASSO procedure performs poorly in terms of recovering the relevant covariates and avoiding noisy ones. This adds more noise than relevant covariates to the model when λ is selected by cross-validation techniques minimising a prediction criterion, like the LASSO.min or LASSO.1se (see Section 3.1). Nevertheless, this recovers the complete set Spaying the price of noise addition

Table 10. *Most competent procedures in terms of the considered simulation scenarios.*

Orthogonal design	*Dependence by blocks*	*Toeplitz covariance*	
		3.a	*3.b*
AdapL.1se Dant RelaxL DC.VS	AdapL.min AdapL.1se SCAD	AdapL.1se Dant DC.VS	AdapL.1se
LASSO.1se, LASSO.BIC, AdapL.min, SCAD, SqrtL, ScalL	*RelaxL, ScalL, DC.VS*	*LASSO.1se, LASSO. BIC*	*SCAD, Dant, DC.VS*

Under the dashed line, we show other studied techniques that improve the LASSO.min performance.

As a result, this selection of covariates overestimates the prediction errors. This is also appreciated for the BIC version (LASSO.BIC) when $p > n$, although this improves the results for $n \geq p$ performing a good covariates selection for a large enough value of n. These drawbacks can be easily overcome making use of other penalisation techniques, keeping the ideas of the L_1 regularisation, as the ones proposed in Section 4.1. All these procedures improve the LASSO results in this independence context, decreasing the number of selected noisy covariates and correcting the overestimation. We highlight the adaptive LASSO of Zou (2006) (AdapL.1se), the relaxed LASSO of Meinshausen (2007) (RelaxL), the Dantzig selector of Candes & Tao (2007) (Dant) and the distance correlation algorithm of Febrero-Bande *et al*. (2019) (DC.VS) as the best of the proposed algorithms for this framework. They are able to recover the complete set S adding little noise for a great enough value of n. Besides, they correct the prediction errors.

These disadvantages of the LASSO are also transferred to dependence structures. The confusion phenomenon appears in these situations involving an increment of false discoveries and overestimation. Here, not all the proposed methods of Section 4.1 performs properly. It depends on the nature of the correlation which methodologies will be efficient. In order to test their adequacy, we have consider different scenarios under a dependence by blocks structure and under a time series one. We found that one version of the adaptive LASSO of Zou (2006) (AdapL.1se) and the distance correlation algorithm of Febrero-Bande *et al*. (2019) (DC.VS) are the only ones quite competent in all these scenarios, regarding to different types of dependence.

The quality of some procedures performance vary according to the type of correlation structure of the data. Examples of this are the SCAD penalisation of Fan (1997) (SCAD) and the Dantzig selector of Candes & Tao (2007) (Dant). The first one achieves a good performance except for the case when there exists strong correlations between all the relevant covariates. In contrast, the Dantzig selector performs properly in these scenarios, but this is not capable of recovering the important covariates, avoiding noise, under a dependence structure by blocks.

The rest of analysed methods: relaxed LASSO (RelaxL), square-root LASSO (SqrtL) and scaled LASSO (ScalL), present a deficient behaviour when there exists some type of dependence structure between the covariates. In case of the dependence by blocks, as in Scenario 2, the relaxed LASSO and the scaled LASSO mix relevant covariates with unimportant ones even for $\rho = 0.5$, whereas the square-root LASSO does not take advantage of the correlation structure. For the Toeplitz covariance scenario, all of them mimic the LASSO behaviour trying to recover the complete set S instead of making use of the structure of the data to correctly adjust the regression model.

As mentioned in Section 3, in the different considered dependence structures, all covariates are in the same scale. Analysis about the effect of different scales on the covariates, combined with dependence structures, are interesting for future work.

ACKNOWLEDGEMENTS

We are grateful to the Joint Editor, Associate Editor and two referees for the constructive comments that significantly improved the paper. This work has been partially supported by the Spanish Ministerio de Economía, Industria y Competitividad grant MTM2016-76969-P, Xunta de Galicia Competitive Reference Groups 2017-2020 (ED431C 2017/38) and the Xunta de Galicia grant ED481A-2018/264. Besides, we want to acknowledge to the Supercomputing Center of Galicia (CESGA) for the computational facilities.

Notes

[1]These assumptions will be treated in Section 2.1.
[2]Some of these procedures will be introduced later in Section 4.

References

Akaike, H. (1998). Information theory & an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, pages 33–40. ACM.

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annal. Stat.*, **43**(5):2055–2085.

Barber, R. F. & Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *Annal. Stat.*, **47**(5):2504–2537.

Beale, E. M. L., Kendall, M. G. & Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, **54**(3–4):357–366.

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Ther. Ber.*, **19**(2):521–547.

Belloni, A., Chernozhukov, V. & Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**(4):791–806.

Bhadra, A., Datta, J., Polson, N. G. & Willard, B. (2019). Lasso meets horseshoe: a survey. *Stat. Sci.*, **34**(3):405–427.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W. & Candès, E. J. (2015). SLOPE-adaptive variable selection via convex optimization. *Annal. Appl. Stat.*, **9**(3):1103.

Breheny, P. & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annal. Appl. Stat.*, **5**(1):232–253.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods*, *Theory & Applications*. Springer Science & Business Media.

Bunea, F. (2008). Honest variable selection in linear & logistic regression models via l1 & l1 + l2 penalization. *Electron. J. Stat.*, **2**. 1154-1194

Bunea, F., Tsybakov, A. B. & Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Annal. Stat.*, **35**(4):1674–1697.

Candes, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: "model-X" knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Series B Stat. Methodology*, **80**(3):551–577.

Candes, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annal. Stat.*, **35**(6):2313–2351.

Candes, E. J., Romberg, J. K. & Tao, T. (2006). Stable signal recovery from incomplete & inaccurate measurements. *Comm. Pure Appl. Math.*, **59**(8):1207–1223.

Castillo, I., Schmidt-Hieber, J. & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Annal. Stat.*, **43**(5):1986–2018.

Chen, S. S., Donoho, D. L. & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Rev.*, **43**(1):129–159.

Dalalyan, A. S., Hebiri, M. & Lederer, J. (2017). On the prediction performance of the LASSO. *Ther. Ber.*, **23**(1):552–581.

Descloux, P. & Sardy, S. (2021). Model selection with lasso-zero: adding straw to the haystack to better find needles. *J. Comput. Graph. Stat.*, 1–14.

Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values & R-software hdi. *Stat. Sci.*, **30**(4):533–558.
Donoho, D. L., Elad, M. & Temlyakov, V. N. (2005). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, **52**(1):6–18.
Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Annal. Stat.*, **32**(2):407–499.
Fan, J. (1997). Comments on <<wavelets in statistics: A review>> by A. Antoniadis. *J. Italian Stat. Soc.*, **6**(2):131-138.
Fan, J., Guo, S. & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Series B Stat. Methodology*, **74**(1):37–65.
Febrero-Bande, M., González-Manteiga, W. & Oviedo de la Fuente, M. (2019). Variable selection in functional additive regression models. *Comput. Stat.*, **34**(2):469–487.
Febrero-Bande, M. & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: the R package fda.usc. *J. Stat. Softw.*, **51**(4):1–28.
Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**(1):1–22.
Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Stat.*, **7**(3):397–416.
Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC.
Giraud, C., Huet, S. & Verzelen, N. (2012). High-dimensional regression with unknown variance. *Stat. Sci.*, **27**(4):500–518.
Hastie, T. & Efron, B. (2013). lars: least angle regression, LASSO & forward stagewise. R package version 1.2.
Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference & Prediction*. Springer Science & Business Media.
Hastie, T., Tibshirani, R. & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection & the LASSO. *arXiv preprint arXiv:1707.08692*.
Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning With Sparsity: The Lasso & Generalizations*. CRC press.
Hocking, R. R. & Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Dent. Tech.*, **9**(4):531–540.
Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Dent. Tech.*, **12**(1):55–67.
Hofner, B., Boccuto, L. & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinform.*, **16**:144.
Homrighausen, D. & McDonald, D. J. (2018). A study on tuning parameter selection for the high-dimensional LASSO. *J. Stat. Comput. Simul.*, **88**(15):2865–2892.
Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statist. Sinica*, 1603–1618.
Javanmard, A. & Montanari, A. (2018). Debiasing the LASSO: Optimal sample size for Gaussian designs. *Annal. Stat.*, **46**(6A):2593–2622.
Lahiri, S. N. (2021). Necessary & sufficient conditions for variable selection consistency of the LASSO in high dimensions. *Annal. Stat.*, **49**(2):820–844.
Leng, C., Lin, Y. & Wahba, G. (2006). A note on the LASSO & related procedures in model selection. *Statist. Sinica*, **16**(4):1273–1284.
Li, X., Zhao, T., Wang, L., Yuan, X. & Liu, H. (2019). *flare: family of LASSO regression*. R package version 1.6.0.2.
Lounici, K. (2008). Sup-norm convergence rate & sign concentration property of LASSO & Dantzig estimators. *Electron. J. Stat.*, **2**.
Meinshausen, N. (2007). Relaxed LASSO. *Comput. Stat. Data Anal.*, **52**(1):374–393.
Meinshausen, N. (2012). relaxo: Relaxed LASSO. R package version 0.1-2.
Meinshausen, N. & Bühlmann, P. (2006). High dimensional graphs & variable selection with the LASSO. *Annal Stat*, **34**(3):1436–1462.
Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Series B Stat. Methodology*, **72**(4):417–473.
Meinshausen, N. & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annal. Stat.*, **37**(1):246–270.
Nan, Y. & Yang, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *J. Comput. Graph. Stat.*, **23**(3):636–656.
Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, **24**(2):227–234.
R Core Team (2019). *R: A Language & Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, S., Tibshirani, R. & Friedman, J. (2016). A study of error variance estimation in LASSO regression. *Statist. Sinica*, **26**:35–67.
Ročková, V. & George, E. I. (2018). The spike-and-slab LASSO. *J. Am. Stat. Assoc.*, **113**(521):431–444.
Schneider, U. & Tardivel, P. (2020). The geometry of uniqueness, sparsity & clustering in penalized estimation. *arXiv: Statistics Theory.*
Schwarz, G. (1978). Estimating the dimension of a model. *Annal. Stat.*, **6**(2):461–464.
Städler, N., Bühlmann, P. & van de Geer, S. (2010). L1-penalization for mixture regression models. *TEST*, **19**:209–256.
Su, W., Bogdan, M. & Candès, E. (2017). False discoveries occur early on the LASSO path. *Annal. Stat.*, **45**(5):2133–2150.
Sun, T. (2019). scalreg: scaled sparse linear regression. R package version 1.0.1.
Sun, T. & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, **99**(4):879–898.
Szekely, G. J. & Rizzo, M. L. (2017). The energy of data. *Annu. Rev. Stat. Appl.*, **4**:447–479.
Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring & testing dependence by correlation of distances. *Annal. Stat.*, **35**(6):2769–2794.
Tibshirani, R. (1996). Regression shrinkage & selection via the LASSO. *J. R. Stat. Soc. B. Methodol.*, **58**(1):267–288.
Tibshirani, R. (2011). Regression shrinkage & selection via the LASSO: a retrospective. *J. R. Stat. Soc. Series B Stat. Methodology*, **73**(3):273–282.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity & smoothness via the fused LASSO. *J. R. Stat. Soc. Series B Stat. Methodology*, **67**(1):91–108.
Tibshirani, R. J. (2013). The LASSO problem & uniqueness. *Electron. J. Stat.*, **7**:1456–1490.
Van de Geer, S., Bühlmann, P. & Zhou, S. (2011). The adaptive & the thresholded LASSO for potentially misspecified models (and a lower bound for the LASSO). *Electron. J. Stat.*, **5**:688–749.
Van De Geer, S. A. & Bühlmann, P. (2009). On the conditions used to prove oracle results for the LASSO. *Electron. J. Stat.*, **3**:1360–1392.
Vidaurre, D., Bielza, C. & Larranaga, P. (2013). A survey of l_1 regression. *Int. Stat. Rev.*, **81** 361 387.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional & noisy sparsity recovery using l_1-constrained quadratic programming (LASSO). *IEEE Trans. Inf. Theory*, **55**(5):2183–2202.
Wang, S., Nan, B., Rosset, S. & Zhu, J. (2011). Random LASSO. *Annal. Appl. Stat.*, **5**(1):468.
Wasserman, L. & Roeder, K. (2009). High dimensional variable selection. *Annal. Stat.*, **37**(5A):2178.
Weinstein, A., Barber, R. & Candès, E. (2017). A power & prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.
Yang, E., Lozano, A. & Ravikumar, P. (2014). Elementary estimators for high-dimensional linear regression. In Xing, E. P. & Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research*, pages 388–396, Bejing, China. PMLR.
Yang, Y. (2005). Can the strengths of AIC & BIC be shared? A conflict between model identification & regression estimation. *Biometrika*, **92**:937–950.
Yuan, M. & Lin, Y. (2006). Model selection & estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodology*, **68**(1):49–67.
Yuan, M. & Lin, Y. (2007). Model selection & estimation in the Gaussian graphical model. *Biometrika*, **94**(1):19–35.
Zhao, P. & Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.*, **7**:2541–2563.
Zhou, S. (2010). Thresholded LASSO for high dimensional variable selection & statistical estimation. *arXiv preprint arXiv:1002.1583*.
Zou, H. (2006). The adaptive LASSO & its oracle properties. *J. Am. Stat. Assoc.*, **101**(476):1418–1429.
Zou, H. & Hastie, T. (2005). Regularization & variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodology*, **67**(2):301–320.

[*Received September 2020; accepted July 2021*]