Sparse Conformal Predictors

Mohamed Hebiri

Sparse Conformal Predictors

Mohamed Hebiri*

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris 7 - Diderot, UFR de Mathématiques,
175 rue de Chevaleret F-75013 Paris, France.

Abstract

Conformal predictors, introduced by Vovk et al. [16], serve to build prediction intervals by exploiting a notion of conformity of the new data point with previously observed data. In the present paper, we propose a novel method for constructing prediction intervals for the response variable in multivariate linear models. The main emphasis is on sparse linear models, where only few of the covariates have significant influence on the response variable even if their number is very large. Our approach is based on combining the principle of conformal prediction with the ℓ_1 penalized least squares estimator (LASSO). The resulting confidence set depends on a parameter $\varepsilon > 0$ and has a coverage probability larger than or equal to $1 - \varepsilon$. The numerical experiments reported in the paper show that the length of the confidence set is small. Furthermore, as a by-product of the proposed approach, we provide a data-driven procedure for choosing the LASSO penalty. The selection power of the method is illustrated on simulated data.
Keywords: LASSO, LARS, Sparsity, Variable selection, Regularization path, Confidence set.
AMS 2000 subject classifications: Primary 62J05, 62J07; Secondary 62F25, 62L12.

1 Introduction

Consider observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ for $i \geq 1$ from a linear regression model $y_i = x_i'\beta + \xi_i$, where $\beta \in \mathbb{R}^p$ is the unknown parameter and the ξ_i's are the noise variables. Suppose we have already collected the dataset $\mathcal{E}_n = ((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_{new})$ where $x_{new} \in \mathbb{R}^p$ denotes a new observation. Our goal is to predict the label y_{new} corresponding to x_{new} based on $\mathcal{E}_n$ and then exploiting the information in x_{new}. This setup is known as the transduction problem [12]. Our estimation strategy is based on local arguments in order to produce a better estimation for y_{new} [5]. More precisely, we will follow the approach of *conformal prediction* presented by Vovk et al. [16] which relies on two key ideas: one is to provide a confidence prediction (namely, a confidence set containing y_{new} with high probability) and the other is to account for the similarity of the new data x_{new} compared to the previously observed x_i's. The notion of conformal predictor was first described by Vovk et al. [15]. Moreover, in [16], the authors illustrate this approach on the example of ridge regression. Along the paper, this predictor will be referred to as Conformal Ridge Predictor[1] (CoRP). In the present contribution, we

*hebiri@math.jussieu.fr

[1]The Conformal Ridge Predictor was called the Ridge Regression Confidence Machine in Vovk et al. [16].

propose to adapt conformal predictors to the sparse linear regression model, that is a model where the regression vector $\beta \in \mathbb{R}^p$ contains only a few of nonzero components. We introduce a novel conformal predictor called the *Conformal Lasso Predictor* (CoLP) which takes into account the sparsity of the model. Its construction is based on the LASSO estimator [10]. The LASSO estimator for linear regression corresponds to an ℓ_1-penalized least square estimator and it has been extensively studied over the last few years ([7, 8, 1, 19], among others) and several modifications have been proposed ([20, 18, 21, 11, 6] among others). One attractive aspect of the LASSO is that it aims both to provide accurate estimating while enjoying variable selection when the model is sparse. In the approach considered in the present paper, the resulting Conformal Lasso Predictor has a large coverage probability and are small in term of its length in the same time. When we deal with regularized methods like the Ridge or the LASSO estimators, the choice of the penalty is an important task. Contrary to the Conformal Ridge Predictor for which no rule was established to pick the Ridge-penalty [16], the construction of the Conformal Lasso Predictor provides a data-driven way for choosing the LASSO-penalty. Moreover, it turn out that this choice is adapted to variable selection as supported by the numerical experiments.

The paper is organized as follows. We concisely introduce conformal prediction and the LASSO procedure in Section 2 and Section 3 respectively. In Section 4, we give the explicit form of the Conformal Lasso Predictor. An algorithm producing the CoLP is presented in Section 5. Then in Section 6 we discuss a generalization of the Conformal Lasso Predictor to other selection-type procedures; we call these generalized procedures *Sparse Conformal Predictors*. Finally, in Section 7, we illustrate the performance of Sparse Conformal Predictors through some numerical experiments.

2 Conformal prediction

Let us briefly describe the approach based on conformal prediction developed in the book by Vovk et al. [16] where they develop the idea of *conformal* prediction. In order to predict the label y_{new} of a new observation $x_n = x_{new}$, the similarity of pairs of the form (x_{new}, y), where $y \in \mathbb{R}$, to the former observations (x_i, y_i) for $i = 1, \ldots, n-1$ is exploited. This is the purpose of introducing a *nonconformity score* $\alpha(y) = (\alpha_1(y), \ldots, \alpha_n(y))'$ which is based on $\mathcal{E}_n$. Each component α_i describes the efficiency of explaining the observation (x_i, y_i) by a procedure based on the augmented sample $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_{new}, y)\}$. In order to obtain a relative information between different nonconformity scores α_i, we shall use the notion of *p-value*, as introduced in [16], defined as:

$$p(y) = \frac{1}{n} \left| \{ i \in \{1, \ldots, n\} : \alpha_i(y) \geq \alpha_n(y) \} \right| , \tag{1}$$

where for any set $\mathcal{A}$, we denote its cardinality by $|\mathcal{A}|$. The above quantity lies between $1/n$ and 1. Moreover, we note that the smaller this p-value is, the less likely the tested pair (x_{new}, y) is (in other words, y is an outlier when associated to x_{new}). An explicit form of the nonconformity score and the p-value will be given in Section 4 when we will adapt it to the CoLP.

Remark 1. *The notion of p-value introduced in the present paper differs from the classical one. To make the connection with hypothesis testing in mathematical statistics [2], consider*

the following hypotheses:

$$\begin{cases} H_0: & \textit{the pair } (x_{new}, y) \textit{ is conformal,} \\ H_1: & \textit{the pair } (x_{new}, y) \textit{ is not conformal.} \end{cases}$$

Assume the observation $Y = y$ is given. The function $p(y)$ permits to construct a statistical test procedure with critical region $\mathcal{R}_\varepsilon = \{y : p(y) \leq \varepsilon\}$ and H_0 is rejected if $y \in \mathcal{R}_\varepsilon$.

A nice feature of this nonconformity score is that it can be related to the confidence of the prediction for y_{new}. We now recall the concept of conformal predictor introduced in [16]. Set $\varepsilon \in (0, 1)$. Given the new observation x_{new}, we search for a subset $\Gamma^\varepsilon = \Gamma^\varepsilon(\mathcal{E}_n)$ of $\mathbb{R}$, in which the expected value of y_{new} lies with a probability of $1 - \varepsilon$. The conformal predictor Γ^ε is defined as the set of labels $y \in \mathbb{R}$ such that $p(y) > \varepsilon$. *In other words, Γ^ε consists of labels y which make the pair (x_{new}, y) more conformal than a proportion ε of the previous pairs (x_i, y_i) for $i = 1, \ldots, n-1$.* Note moreover that the smaller ε, the more confident the predictor. That is to say, for any ε_1, $\varepsilon_2 > 0$:

$$\Gamma^{\varepsilon_1} \subset \Gamma^{\varepsilon_2} \quad \text{whenever } \varepsilon_1 \geq \varepsilon_2 \ .$$

In the present analysis, apart from prediction, we develop an approach for selecting relevant variables. For this reason, we consider three criteria measuring the quality of our procedure: *validity*, *accuracy*, and *selection*. The first two were introduced in [17]. The fact that we consider the issue of sparsity leads us to include the selection power of the predictor.

Validity. This criterion accounts for the power of conformal prediction. The simplest approach is to count the number of times where y_n does not belong to the set Γ^ε. We take the notation:

$$\mathrm{err}_n^\varepsilon = \begin{cases} 1 & \text{if } y_n \notin \Gamma^\varepsilon(\mathcal{E}_n) \\ 0 & \text{otherwise.} \end{cases}$$

Note that in an on-line perspective, one focuses on the cumulative error $\mathrm{ERR}_n^\varepsilon = \sum_{i=1}^n \mathrm{err}_i^\varepsilon$. Asymptotic validity properties of this cumulative error have been studied in [13] and [16, chapters 2 and 8]. In the present work, we will be interested in evaluating the error $\mathrm{err}_n^\varepsilon$ for a fixed n, rather than the cumulative one.

Accuracy. The length of the confidence predictor provides a natural measure of the accuracy. We will see that such a measure is adapted to the variable selection purpose. Note that other choices are possible. We shall discuss this point in Section 5.

Selection. Finally, in the case of sparse linear regression, it is important to include a measure of the capacity of the estimator to select relevant variables, namely those for which the regression parameter β has nonzero components.

3 The LASSO Procedure

The LASSO estimator [10] has originally been introduced in the linear regression model:

$$y_i = x_i' \beta^* + \xi_i, \qquad i = 1, \ldots, n-1 \tag{2}$$

where the design $x_i = (x_{i,1}, \ldots, x_{i,p})' \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown regression vector and the ξ_i's are independent and identically distributed (i.i.d.) centered Gaussian random variables with known variance σ^2. Then the goal is to use the observations to provide an approximation of the label y_{new} of a new observation x_{new} through the estimation of the regression vector β^*. The LASSO estimator is defined as follows:

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n-1} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where $\lambda \geq 0$ is a tuning parameter. Based on $\hat{\beta}_\lambda$, an estimation of the response y_{new} of the new observation $x_n = x_{new}$ is produced by $\hat{\mu}_\lambda = x_{new}'\hat{\beta}_\lambda$. For a large enough λ, the LASSO estimator is sparse. That is many components of $\hat{\beta}_\lambda$ equal zero. Therefore we can naturally define a sparsity (or active) set as $\mathcal{A}_\lambda = \{j \in \{1, \ldots, p\} : \hat{\beta}_\lambda \neq 0\}$. A LASSO modification of the LARS algorithm [3] can iteratively provide approximations of the LASSO estimator for a few values of the tuning parameters $\lambda = \lambda_0, \ldots, \lambda_K$ such that $\infty = \lambda_0 > \ldots > \lambda_K = 0$ (the indices refer to the algorithm steps and K denotes the last step). These points are the so-called *transition points*.
From now on, let us write $\hat{\beta}_k$ and $\mathcal{A}_k$ for the LASSO estimator $\hat{\beta}_\lambda$ and the sparsity set $\mathcal{A}_\lambda$ evaluated at the transition point $\lambda = \lambda_k$. Obviously, the estimator $\hat{\beta}_k$ is an $|\mathcal{A}_k|$-dimensional vector where $|\mathcal{A}_k|$ is the cardinality of the set $\mathcal{A}_k$. Furthermore, we denote by s_k the $|\mathcal{A}_k|$-dimensional sign vector whose components are the signs of the components of the LASSO estimator evaluated at the transition point λ_k (i.e., $(s_k)_j = 1$ if $(\hat{\beta}_k)_j > 0$, $(s_k)_j = -1$ if $(\hat{\beta}_k)_j < 0$ where $j \in \mathcal{A}_k$). Finally, let us denote by $\mathbf{x}_k$, the $(n-1) \times |\mathcal{A}_k|$ matrix whose columns are the variables X_j, with indices $j \in \mathcal{A}_k$. For each λ_k, we assume that the matrix $(\mathbf{x}_k'\mathbf{x}_k)^{-1}$ is invertible. Here are some characteristics of the LARS algorithm and we refer to [2] for more details:

i) At each iteration of the algorithm (i.e., at each transition point), only one variable $X_j = (x_{1,j}, \ldots, x_{n-1,j})'$, $j = 1, \ldots, p$ is added (or deleted) to the construction of the estimator according to its correlation with the current residual. The algorithm begins with only one variable and ends up with the ordinary least square (OLS) estimator[2].

ii) For each $\lambda \in (\lambda_{k+1}, \lambda_k]$, the LASSO estimator can be expressed in the following form:

$$\hat{\beta}_\lambda(\mathbf{y}, \mathbf{x}_k, s_k) = (\mathbf{x}_k'\mathbf{x}_k)^{-1}(\mathbf{x}_k'\mathbf{y} - \frac{\lambda}{2} s_k), \tag{4}$$

where $\mathbf{y} = (y_1, \ldots, y_{n-1})'$. Note that (4) is obtained by minimizing (3) over the set $\mathcal{A}_k$. Let us also mention that the set $\mathcal{A}_k$ and the sign vector s_k remain unchanged when λ varies in the interval $(\lambda_{k+1}, \lambda_k]$.

iii) As highlighted by (4), the LASSO estimator is piecewise linear in λ and linear in $\mathbf{y}$ for every fixed λ [9]. Using the LASSO modification of the LARS algorithm, this property

[2]When $p > n$, the LARS cannot select all p variables. It is limited by the sample size n. In such a case, the last iteration does not correspond to the OLS.

helps us to provide the regularization path of the LASSO estimator, which is defined as $\{\hat{\beta}_\lambda : \lambda \in [0, \infty)\}$ (each point of the regularization path corresponds to the evaluation of the regression vector estimator for a given value of λ). Indeed, the slope of the LASSO regularization path changes at a finite number of points which coincide with the transition points $\lambda_1, \ldots, \lambda_K$.

iv) Piecewise linearity is an important property of the LASSO modification of the LARS algorithm. Indeed, let $\lambda \in (\lambda_{k+1}, \lambda_k]$ where λ_{k+1} and λ_k are two transition points. In this interval, the LASSO estimator $\hat{\beta}_\lambda$ uses the same variables (variables with indices in $\mathcal{A}_k$). By using (4), it is easy to see [22] that the linearity of the LASSO estimator implies that, for any $\lambda \in (\lambda_{k+1}, \lambda_k]$:

$$\sum_{i=1}^{n-1} \left(y_i - x_i' \hat{\beta}_\lambda\right)^2 > \sum_{i=1}^{n-1} \left(y_i - x_i' \hat{\beta}_{\lambda_{k+1}}\right)^2 .$$

This last observation indicates that the transition points are the most interesting points in the regularization path.

All these nice properties encourage the use of the LASSO as a selection procedure. In the sequel, we will consider the LASSO modification of the LARS algorithm which provides an approximate solution to the LASSO.

Remark 2. *Through the paper, one should keep in mind the analogy between each iteration k of the modification of the LARS algorithm and its corresponding tuning parameter value λ_k. Decrease of tuning parameter λ is reflected through the increase of the number of iterations of the modification of the LARS algorithm.*

4 Sparse predictor with conformal Lasso

For the reasons exposed above, we focus on the transition points $\lambda_1, \ldots, \lambda_K$ and construct conformal predictors for each of these λ_k. We then propose to select the best conformal predictor among them according to its performance in terms of accuracy (cf. Section 2).

Now let us detail the construction of the CoLP for each λ_k. To this end, denote by $X_j = (x_{1,j}, \ldots, x_{n-1,j}, x_{new,j})'$, $j = 1, \ldots, p$ the augmented variable j. Define the augmented matrix $\widetilde{\mathbf{x}} = (x_1, \ldots, x_{n-1}, x_{new})' = (X_1, \ldots, X_p)$ and the augmented response vector $\widetilde{\mathbf{y}} = (y_1, \ldots, y_{n-1}, y)'$ where y is a candidate value for y_{new}. Using the notation introduced in Section 3, for the fixed λ_k, we also define the LASSO estimator $\hat{\beta}_k(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_k, s_k)$ from expression (4) with the augmented data. From now on, we denote this estimator by $\hat{\beta}_k$. Define $\hat{\mu}_k := \widetilde{\mathbf{x}}_k \hat{\beta}_k$. Moreover, the matrix $\mathbf{H}_k$ will be the $n \times n$ projection matrix onto the subspace generated by $\widetilde{\mathbf{x}}_k$ and $\mathbf{I}$ identity matrix of the same size. For each λ_k, we define a corresponding nonconformity score $\alpha^k = \left(\alpha_1^k, \ldots, \alpha_n^k\right)'$ by:

$$\begin{aligned} \alpha^k(y) &:= |\widetilde{\mathbf{y}} - \hat{\mu}_k| = |\left(\mathbf{I} - \mathbf{H}_k\right) \widetilde{\mathbf{y}} + \frac{\lambda_k}{2} \widetilde{\mathbf{x}}_k \left(\widetilde{\mathbf{x}}_k' \widetilde{\mathbf{x}}_k\right)^{-1} s_k| \\ &= |A_k + B_k\, y|, \end{aligned}$$

where $|\cdot|$ is meant here componentwise and

$$\begin{cases} A_k = (a_1^k, \ldots, a_n^k)' := \left(\mathbf{I} - \mathbf{H}_k\right) (y_1, \ldots, y_{n-1}, 0)' + \frac{\lambda_k}{2} \widetilde{\mathbf{x}}_k \left(\widetilde{\mathbf{x}}_k' \widetilde{\mathbf{x}}_k\right)^{-1} s_k, \\ B_k = (b_1^k, \ldots, b_n^k)' := \left(\mathbf{I} - \mathbf{H}_k\right) (0, \ldots, 0, 1)', \end{cases} \tag{5}$$

Note that each component $\alpha_i^k(y)$ is piecewise linear with respect to y. Then the corresponding p-value $p_k(y)$ as defined by (1) clearly can change only at points y where the sign of $\alpha_i^k(y) - \alpha_n^k(y)$ changes. Hence, we do not have to evaluate all the possible values of y. We only focus on points y for which the i-th nonconformity measure $\alpha_i^k(y)$ equals $\alpha_n^k(y)$. For this purpose, we define, for each observation $i \in \{1, \ldots, n\}$

$$S_i^k = \left\{ y : \alpha_i^k(y) \geq \alpha_n^k(y) \right\}, \tag{6}$$

which corresponds to the range of values y such that the new pair (x_{new}, y) has a better conformity score than the i-th pair (x_i, y_i). Moreover, let l_i^k and u_i^k denote two real defined respectively as

$$l_i^k = \min\{-\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k}\}, \qquad \text{and} \qquad u_i^k = \max\{-\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k}\}, \tag{7}$$

where a_i^k and b_i^k are given by (5).

Proposition 1. *Let us fix a $k \in \{1, \ldots, K\}$ and an $i \in \{1, \ldots, n-1\}$. Assume that both b_i^k and b_n^k are non-negative. Then*

i) if $b_i^k \neq b_n^k$, we have either $S_i^k = [l_i^k; u_i^k]$ or $S_i^k = (-\infty; l_i^k] \cup [u_i^k; -\infty)$, with l_i^k and u_i^k given by (7).

ii) if $b_i^k = b_n^k \neq 0$, then $l_i^k = u_i^k = -\frac{a_i^k + a_n^k}{2b_n^k}$ and we have either $S_i^k = (-\infty; l_i^k]$ or $S_i^k = [l_i^k; -\infty)$. Moreover if $a_i^k = a_n^k$, we have $S_i^k = \mathbb{R}$.

iii) if $b_i^k = b_n^k = 0$, we have either $S_i^k = \mathbb{R}$ or $S_i^k = \emptyset$.

The assumption that all the b_i^k are non-negative does not make loose any generality as one can multiply a_i^k, b_i^k and c_i^k by -1 if $b_i^k < 0$. With this definition of S_i^k, we may rewrite the definition of the conformal predictor as follows

$$\Gamma_k^\varepsilon = \{y : \sum_{i=1}^n \mathbb{I}(\alpha_i^k(y) \geq \alpha_n^k(y)) \geq n\varepsilon\} = \{y : \sum_{i=1}^n \mathbb{I}(S_i^k)(y) \geq n\varepsilon\}, \tag{8}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function. This approach leads to a whole collection of confidence intervals $\Gamma_1^\varepsilon, \ldots, \Gamma_K^\varepsilon$. We propose below a strategy for choosing one one particular Γ_k^ε, the performance of which will be studied through numerical simulations.

It is worth mentioning that in view of [14, Theorem 1] (see also [16, Proposition 2.3 page 26]), each of predictor Γ_k^ε would have a coverage probability at least equal to $1 - \varepsilon$, if the corresponding value λ_k of the tuning parameter were deterministic. In fact, the following result holds.

Proposition 2. *Fix the significance level $\varepsilon \in (0, 1)$ and the tuning parameter $\lambda > 0$. Let $\hat{\beta}_{\lambda,n}(y)$ be the Lasso estimate for the augmented dataset $(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})$ and let us define $\alpha^\lambda(y) = |\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\hat{\beta}_{\lambda,n}(y)|$. Then, the conformal predictor*

$$\Gamma_\lambda^\varepsilon = \left\{ y : \sum_{i=1}^n \mathbb{I}(\alpha_i^\lambda(y) \geq \alpha_n^\lambda(y)) \geq n\varepsilon \right\},$$

satisfies

$$\mathbb{P}(y_{new} \in \Gamma_k^\varepsilon) \geq 1 - \varepsilon,$$

for any $n \in \mathbb{N}$.

Actually, in the proof of Proposition 2 detailed in [14], one needs the exchangeability of the pairs $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y)$ in the definition of the predictor. This property is not fulfilled when the tuning parameter λ is chosen in the set $\{\lambda_1, \ldots, \lambda_K\}$ of Lasso's transition points, since the elements of this set depend only on the first $n-1$ observations and not on (x_n, y). We believe that under some additional assumptions a result similar to Proposition 2 can be obtained for the predictor Γ_k^ε as well, for each $k = 1, \ldots, K$. This is the topic of an ongoing work. In the present paper, we content ourselves by proposing a data-driven choice of the conformal predictor from the collection of predictors $\{\Gamma_k^\varepsilon; 1 \leq k \leq K\}$ and by exploring its empirical properties.

Remark 3. *Of course, one can also apply the well-known sample splitting technique for choosing the values* $\lambda_1, \ldots, \lambda_K$ *based on a first sample, and then use the methodology described below for selecting the data-driven predictor based on a second sample which is assumed to be independent of the first sample. However, this technique is not attractive from the practical standpoint, that is why we do not develop this approach.*

As discussed above, we believe that all the predictors Γ_k^ϵ share nearly the $1-\varepsilon$ validity property, which is supported by our empirical study. We suggest to select among them the one which has the smallest Lebesgue measure. We denote this confidence set by Γ_{opt}^ϵ, that is

$$\Gamma_{opt}^\varepsilon = \Gamma_\nu^\varepsilon, \qquad \nu = \operatorname*{argmin}_k |\Gamma_k^\varepsilon|. \tag{9}$$

In general, since ν is a random variable, the $1-\varepsilon$ validity of all Γ_k^ε would not imply the $1-\varepsilon$ validity of Γ_{opt}^k, but only $1 - K\varepsilon$ validity. However, $1 - K\varepsilon$ is a worst case majorant obtained by a simple application of the union bound, whereas numerical examples we considered (some of them are reported below) suggest that the validity is much better than $1 - K\varepsilon$ and could even be equal to $1-\varepsilon$ when $p \leq n$.

5 Implementation

We provide here a three-step algorithm which enables us to easily construct the CoLP. We start in **Step 1** by applying the LASSO modification of the LARS algorithm to the dataset $((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}))$. This step provides all transition points $\lambda_1, \ldots, \lambda_K$, the corresponding design matrices $\mathbf{x}_k$ and sign vectors s_k for $k = 1, \ldots, K$. Then, in **Step 2**, we construct the conformal predictor Γ_k^ε associated to each λ_k. Thanks to Proposition 1, for each λ_k, we can construct the sets S_i^k for $i = 1 \ldots, n$ defined by (6). We use these sets in order to construct the conformal predictor Γ_k^ε. To do this, we take advantage from the fact that the function $y \mapsto \sum_{i=1}^n \mathbb{I}(S_i^k(y))$ is piecewise constant. Furthermore, the endpoints of the intervals where this function is constant belong to the set of the all endpoints of intervals forming the sets S_i^k. Thus, to determine Γ_k^ε, we sort the set U consisting of the all endpoints of the intervals described in Proposition 1 and include an interval having as endpoints two

successive elements of U in Γ_k^ε if the center of this interval belongs to at least $[n\varepsilon]$ sets S_i^k.

Algorithme 1 : Lasso Conformal Predictor

Step 1: Run the LASSO modification of the LARS algorithm on the data set $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}))$

Step 2: Construct the Conformal Lasso Predictors for each $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$ **begin**

  Step 2a: Initialization : Define A_k and B_k as in (5). Set $U^k \longleftarrow \emptyset$

  Step 2b: Harmonization

  for $i = 1$ **to** n **do**

    if $b_i^k < 0$ **then**

      $a_i^k = -a_i^k$ and $b_i^k = -b_i^k$

    end

  end

  Step 2c: Actualize the set U^k

  for $i = 1$ ***to*** n **do**

    if $b_i^k \neq b_n^k$ **then**

      Add l_i^k and u_i^k (7) to U^k

    end

    if $b_i^k = b_n^k \neq 0$ *and* $a_i^k \neq a_n^k$ **then**

      Add $l_i^k = u_i^k$ (7) to U^k

    end

  end

  Step 2d: Sort U^k. Let $m \longleftarrow |U^k|$. Then $y_{(0)} \longleftarrow -\infty$ and $y_{(m+1)} \longleftarrow +\infty$

  Step 2e: Evaluate N_j^k for $j = 1, \ldots, m$. Initialize $N_j^k \longleftarrow 0$. Then actualize

  for $i = 1$ **to** n **do**

    for $j = 1$ **to** m **do**

      if $|a_i^k + b_i^k\, y| \geq |a_n^k + b_n^k\, y|$ *for* $y \in (y_{(j)}, y_{(j+1)})$ **then**

        Increment $N_j^k = N_j^k + 1$

      end

    end

  end

  Step 2f: For a fixed threshold $\varepsilon > 0$, output the conformal predictor

$$\Gamma_k^\varepsilon = \cup_{j:\frac{N_j}{n} > \varepsilon} [y_{(j)}, y_{(j+1)}]$$

end

Step 3: Output the Conformal Lasso Predictor Γ_{opt}^ε as the smallest (w.r.t. their Lebesgue measure) confidence set among the constructed conformal predictors

Finally, in a **Step 3**, we provide the CoLP, says Γ_{opt}^ε, which is defined as the smallest confidence set, according to its Lebesgue measure, among the constructed conformal predictors Γ_k^ε, $k = 1, \ldots, K$. According to Proposition 2, each Γ_k^ε is valid. Moreover the criterion for choosing the CoLP is adapted to variable selection as conformal predictors constructed here for different values of λ_k, $k = 1, \ldots, K$ bring into play different variables. This is illustrated in Figure 5 (left) where we constructed the conformal predictors when $n = 300$. One can observe that all the conformal predictors are valid since they contain the true value of the label y_{new}. Hence our construction is suitable when the sample size is larger than the number

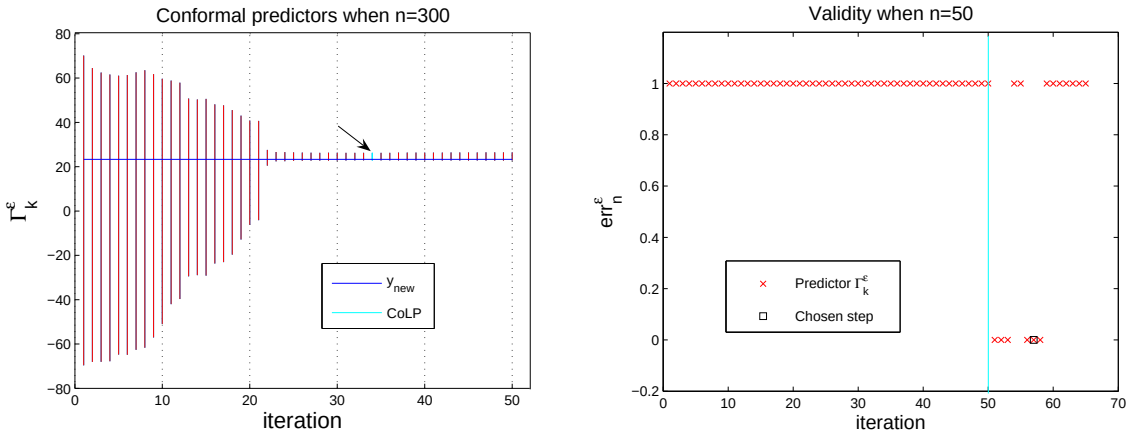

Figure 1: *Left*: Conformal predictors Γ_k^ε evolution through the iterations of the LASSO modification of the LARS algorithm when $n = 300$ (the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The CoLP is drawn in cyan and corresponds to the 34-th iteration. The horizontal blue line corresponds to the value of y_{new}. *Right*: Validity analysis (err_n^ε) of the conformal predictors Γ_k^ε through the iterations of the LASSO modification of the LARS algorithm when $n = 50$ (the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The CoLP is marked by a black square and corresponds to the 57-th iteration. The vertical line represents a separation between a stable and an unstable zone.

of variables (i.e., $n > p$) but may be not appropriated when $p \geq n$. Figure 5 (right) shows an example where almost all the constructed conformal predictors Γ_k^ε, $k = 1, \ldots, K$, using the above algorithm are valid. Only six are not. One of them is the selected CoLP (iteration 57 in Figure 5 (right)) which corresponds to the smallest predictor. In such cases ($p \geq n$), a correction can be made and other choices for the accuracy measure are possible. We discuss this criterion in Section 7. Let us add that we only illustrated the validity of the conformal predictors in Figure 5 (right) as the unstable zone (on the right side of the vertical line) makes the representation hard to be analyzed. More details are given in Section 7.

Remark 4. *In **Step 1** of Algorithm 1, we use the LARS algorithm for its ability to generate a small number of tuning parameter values of interest. It is an important aspect as it considerably reduces the computational cost. On-line versions could be implemented by plugging in an on-line version of the LASSO solution as in [4]. The analysis of such on-line versions is the object of work under progress.*

6 Extension to others procedures

In this section we generalize the construction of the confidence predictor to a family of estimators which includes selection-type procedures as the Elastic-Net [21] and the Smooth-Lasso [6]. As for CoLP (Section 4), we are interested in two properties of estimators: the *piecewise linearity w.r.t. the response* y (to easily compute the nonconformity scores α_i, $i = 1, \ldots, n$), and the *piecewise linearity w.r.t the tuning parameter* λ *[9]* (to reduce computational effort by using a modification of the LARS algorithm).
We use the same notation as in Section 3 for the LASSO estimator. Set $\hat{\beta}$ to be an estimator of the regression vector β based on $\mathbf{x}$ and $\mathbf{y}$. Let also s be the sign vector of the estimator $\hat{\beta}$. On the other hand, using the notation in Section 4, we set $\hat{\mu} = \widetilde{\mathbf{x}}\hat{\beta}$ where this time $\hat{\beta}$ is based on the augmented dataset $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$.

Assumption 1. *The estimator $\hat{\mu}$ can be written as:*

$$\hat{\mu} = u(\widetilde{\mathbf{x}}, s)\widetilde{\mathbf{y}} + v(\widetilde{\mathbf{x}}, s), \tag{10}$$

where $u(\cdot)$ and $v(\cdot)$ are piecewise constant functions w.r.t. $\widetilde{\mathbf{y}}$.

As soon as Assumption 1 holds, we can construct a conformal predictor corresponding to the estimator $\hat{\mu}$. Then many estimators can be considered. The CoLP and CoRP obviously belong to this class of predictors and we introduce here the Conformal Elastic Net Predictor (CENeP) which is a conformal predictor constructed based on the Elastic-Net modification of the LARS instead of the LASSO one (**Step1** in Algorithm 1). This predictor is defined by $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \mu_k\mathbf{I}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$ where λ_k and μ_k correspond respectively to the LASSO and Ridge tuning parameters in the definition of the Elastic-Net estimator and $\mathbf{I}_k$ is the $|\mathcal{A}_k| \times |\mathcal{A}_k|$ identity matrix [21]. In the same way, we can define the Conformal Smooth Lasso Predictor (CoSmoLaP) based on a Smooth-Lasso modification of the LARS algorithm [6]. Here $u(\widetilde{\mathbf{x}}, s) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \mu_k\mathbf{J}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}, s) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$. The difference between the CoSmoLaP definition the CENeP one is the identity matrix $\mathbf{I}_k$ which is replaced by the $|\mathcal{A}_k| \times |\mathcal{A}_k|$ matrix $\mathbf{J}_k$ whose components are such that $(\mathbf{J}_k)_{i,i} = 1$ if $i = 1$ or $i = |\mathcal{A}_k|$ and $(\mathbf{J}_k)_{i,i} = 2$ otherwise. Moreover for $(i, j) \in \{1, \ldots, \mathcal{A}_k\}^2$ with $i \neq j$, we have $(\mathbf{J}_k)_{i,j} = -1$ if $|i - j| = 1$ and zero otherwise. Note that the definition of $\mathbf{J}_k$ makes the CoSmoLaP more appropriated to model with successive correlation between successive variables.

As for CoLP, we can define the nonconformity score of an expected label y associated to the estimator $\hat{\mu}$ as follows:

$$\begin{aligned}(\alpha_1(y), \ldots, \alpha_n(y))' &:= |\widetilde{\mathbf{y}} - \hat{\mu}| \\ &= |\,(\mathbf{I} - u(\widetilde{\mathbf{x}}, s))\,\widetilde{\mathbf{y}} - v(\widetilde{\mathbf{x}}, s)| \\ &= |A + B\,y|,\end{aligned}$$

with

$$\begin{cases} A = (a_1, \ldots, a_n)' := (\mathbf{I} - u(\widetilde{\mathbf{x}}, s))\,(y_1, \ldots, y_{n-1}, 0)' - v(\widetilde{\mathbf{x}}, s), \\ B = (b_1, \ldots, b_n)' := (\mathbf{I} - u(\widetilde{\mathbf{x}}, s))\,(0, \ldots, 0, 1)', \end{cases}$$

and $\mathbf{I}$ is the $n \times n$ identity matrix. The quantities A and B are the analogues of A_k and B_k respectively, when we considered the CoLP at the transition point λ_k, $k = 1, \ldots, K$. Then replacing A_k and B_k by respectively A and B in **Step 2.a** of Algorithm 1, we obtain the conformal predictors associated to the estimator $\hat{\mu}$.

Note that the dependency in the tuning parameter, noted λ, can be included in $u(\widetilde{\mathbf{x}}, s)$ (as for CoRP) or $v(\widetilde{\mathbf{x}}, s)$ or in both of them (as for the CoLP). For instance, in the construction of the CoLP, this dependency is underlined in the matrix $\widetilde{\mathbf{x}}_k$ and the sign vector s_k as they were computed by the LARS algorithm for a specified value λ_k of the tuning parameter λ.

Computational cost of the construction of conformal predictors has also to be considered. Three main points interfere. First, one run of the LARS algorithm requires the same cost as the computation of the least square estimation. Then we have to consider the number of conformal predictors we have to construct: each value of the tuning parameter λ provides a conformal predictor Γ_λ using the algorithm described in Section 5. The final conformal predictor Γ_{opt} is then the one with the minimal length. As for the CoRP, the main problem is: how many λ's do we have to test? One way is to use a grid of value for λ which lets open the problem of the choice of the grid and the window of this grid.

On the other hand, we saw how the LARS algorithm permits to reduce considerably the number of tuning parameters to be considered. Indeed the grid of tuning parameters values is directly described by the transition points $\lambda_1, \ldots, \lambda_K$ obtained from the run of the LARS algorithm. Finally, let us consider *the construction of the conformal predictor itself:* this point has been treated in Vovk et al. [16, Chapter 2.3 and 4.1]. It turns out that sparse conformal predictors and the CoLP requires computation time $\mathcal{O}(n^2)$ and can be reduced to $\mathcal{O}(n \log(n))$.

7 Experimental Results

In the section we present the experimental performances of the Sparse Conformal Predictors (SCP) w.r.t. their validity, their accuracy and also their selection power. As benchmark, we use the CoRP[3] for its validity and accuracy and the original LASSO and Elastic-Net estimators for their selection[4] power.
We consider three SCPs: the Conformal Lasso Predictor (CoLP was introduced in Sections 4 and 5) and the Conformal Elastic Net Predictor (CENeP was described in Section 6). The last SCP called Conformal Ridge Lasso Predictor (CoRLaP) is a mix of the CoRP and the CoLP. To construct the CoRLaP, we use the variables selected by the LASSO modification of the LARS algorithm (**Step 1** in Algorithm 1 described in Section 5). Then we use these variables to construct a CoRP. This conformal predictor can be seen as a restricted CoRP. All conformal predictors are constructed with confidence level $1 - \varepsilon = 90\%$.

7.1 Simulated Experiments

We consider four simulations from the linear regression model

$$y = \mathbf{X}'\beta + \sigma\xi, \quad \xi \sim \mathcal{N}(0,1),\ \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{50})' \in \mathbb{R}^{50},$$

with $\beta \in \mathbb{R}^{50}$. Hence $p = 50$ through the simulations. Noise level σ and the sample size n are let free. They will be specified during experiments.

Example **(a)** $[n/\sigma]$: *Very Sparse and Correlated.* Here only β_1 is nonzero and equals 5. Moreover, the design correlations matrix Σ is described by $\Sigma_{j,k} = \exp(-|j-k|)$ for $(j,k) \in \{15, \ldots, 35\}^2$ and $\Sigma_{j,k} = \mathbb{I}(j = k)$ otherwise where $\mathbb{I}(\cdot)$ is the indicator function.

Example **(b)** $[n/\sigma]$: *Sparse and Correlated.* The correlations are defined as in Example (a) and the regression vector is given by $\beta_j = -5 + 0.2j$ for $j = 1, \ldots, 5$; $\beta_j = 4 + 0.2j$ for $j = 10, \ldots, 25$ and zero otherwise.

Example **(c)** $[n/\sigma]$: *Sparse and Highly correlated.* We have $\beta_j = 5$ for $j \in \{1, \ldots, 15\}$ and zero otherwise. We construct three groups of correlated variables: $\Sigma_{j,k} = 1$ when (j,k)

[3] We construct the CoRP associated to same tuning parameters as the CoLP (i.e., the transition points λ_k observed in Section 5). Note that the performance would not be inflected as conformal predictors according to this method are almost embedded and changes sensitively while the tuning parameter varies. See [16, page 39] for more details.

[4] We use a BIC-type criterion to select the optimal tuning parameter. Such a criterion is adapted to variable selection.

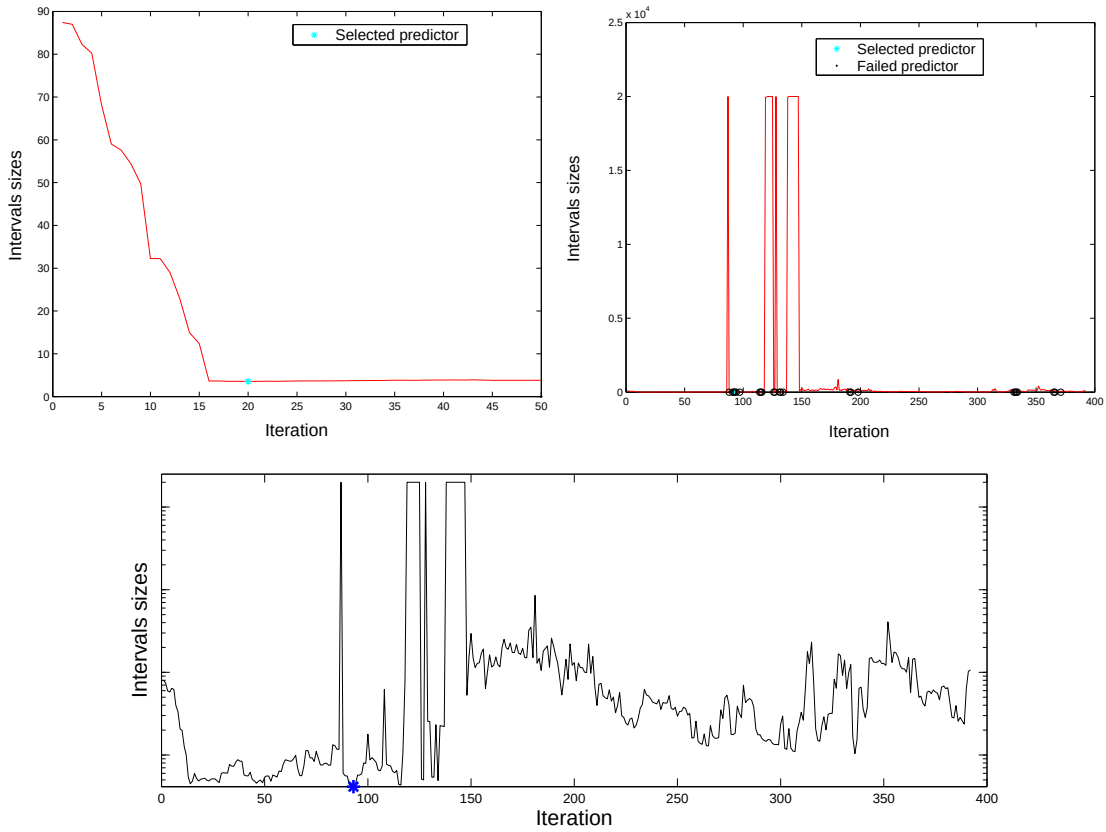

Figure 2: Analysis of conformal predictors length (y-axis) through the LASSO modification of the LARS algorithm iterations (x-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}) in Example (c)[300/1] (top left) and in Example (c)[50/1] (top right). The iteration associated to the CoLP is marked by a blue star. Predictors which are non valid are marked by a black circle. The panel of bottom shows the lengths of intervals in a logarithmic scale.

belongs to $\{1,\dots,5\}^2$, $\{6,\dots,10\}^2$ and $\{11,\dots,15\}^2$; $\Sigma_{j,k}=1$ for $(j,k)\in\{16,\dots,p\}^2$ if $j=k$ and zero otherwise.

Example **(d)** $[n/\sigma]$: *Non Sparse and correlated.* Here $\beta_j = 3+0.2j$ for $j\in\{1,\dots,p\}$ and the correlations are described by $\Sigma_{j,k}=\exp(-|j-k|)$ for $(j,k)\in\{1,\dots,p\}^2$.

We consider separately the three points of interest: accuracy, validity and selection.

Accuracy. First of all, let us consider the length of the predictors Γ^{ε}_k, $k=1,\dots,K$ obtained at the end of **Step 2** in Algorithm 1 described in Section 5. We remind that each of these predictors is associated to an iteration of a modification of the LARS algorithm, that is the transition points λ_k, $k=1,\dots,K$. Figure 7.1 illustrates the predictors lengths for the construction of the CoLP, when applied to Example (c)$[n/1]$ with $n=300$ and $n=50$. When $n=300$, we note that the length of the Γ^{ε}_ks sensitively changes from one iteration to the following and that the larger predictor has a reasonable length compared

Table 1: Validity frequencies [with precision ±95%] of the CoRP, CoLP, CoRLaP, CENeP, the Early-Stopped CoLP and the 2-PN CoLP based on 1000 replications.

Example	σ	CoRP	CoLP	CoRLaP	CENeP
(a)[300/σ]	1	0.897± 0.019	0.876± 0.020	0.854± 0.022	0.878± 0.020
	7	0.894± 0.019	0.908± 0.018	0.894± 0.019	0.899± 0.019
	15	0.893± 0.019	0.893 ± 0.019	0.879± 0.020	0.887± 0.020
(b)[300/σ]	1	0.901± 0.018	0.875± 0.020	0.869± 0.021	0.874± 0.021
(c)[300/σ]	1	0.900± 0.019	0.900± 0.019	0.891± 0.019	0.901± 0.018
(d)[300/σ]	1	0.892± 0.019	0.895± 0.019	0.895± 0.019	0.895± 0.019
(a)[50/σ]	3	0.887± 0.020	0.668± 0.029	0.414± 0.030	0.789± 0.025
(a)[20/σ]	3	0.865± 0.021	0.596± 0.030	0.304± 0.028	0.685± 0.029
Example	σ	CoRP	CoLP	Stopped-CoLP	2-PN-CoLP
(a)[50/σ]	7	0.853± 0.022	0.620± 0.030	0.815± 0.024	0.881± 0.020
(b)[50/σ]	1	0.875± 0.020	0.558± 0.031	0.814± 0.024	0.907 ± 0.018
(c)[20/σ]	15	0.875± 0.020	0.608± 0.030	0.769± 0.026	0.893± 0.019
(d)[20/σ]	1	0.900± 0.019	0.602± 0.030	0.793± 0.025	0.892± 0.019

to the smallest one (about 10 times larger). Then the construction is stable. We also observe that in the neighborhood of the optimal iteration (that is iteration 20), the conformal predictors have approximately the same size. Such an observation can also be made when we take a look at Figure 5 (left) when applied to Example (b)[300/1]. On the other hand, when $n = 50$, it appears that the predictors length grows drastically at some iteration (around iteration 85). We even can not compare the lengths of the bigger and smaller predictors (more than 10^4 times larger). In the same time, it seems that the construction becomes unstable as violent variations often happen after this iteration 85. We will consider in the next point the validity of these predictors. However let us mention that in Example (c)[50/1], the CoLP which is the smallest Γ_k^ε and then the selected predictor is not valid (in Figure 7.1 (right), the selected predictor at iteration 93 is not valid). This aspect can also be observed in Figure 5 (right) (the graph corresponds to Example (b)[50/1]) where the selected CoLP at iteration 57 is not valid. Similar violent variations of the corresponding predictors lengths would have been observed after iteration 49 if we have provided a graph as Figure 7.1 (right).

Validity. Now, we consider the validity of the selected predictors (cf. **Step 3** in Algorithm 1). As shown in Table 1, we observe that variations on the noise level, the variables correlations and the sparsity of the model do to not perturb the validity whereas the sample size relatively to the dimension p does. When $n = 300 > p$, all the procedures seem to be quite similar and produce good predictors. In the other cases, i.e., when $n = p = 50$ and $n = 20 < p$, the selected confidence predictors have worst performance than expected (validity with smaller proportion than $1 - \varepsilon = 90\%$). Moreover, Sparse Confidence Predictors perform worst than the CoRP as observed in Table 1. As pointed in the accuracy part, one explication can be observed in Figure 7.1 as the selected predictor which also is not valid (iteration 93) corresponds to an iteration in the unstable

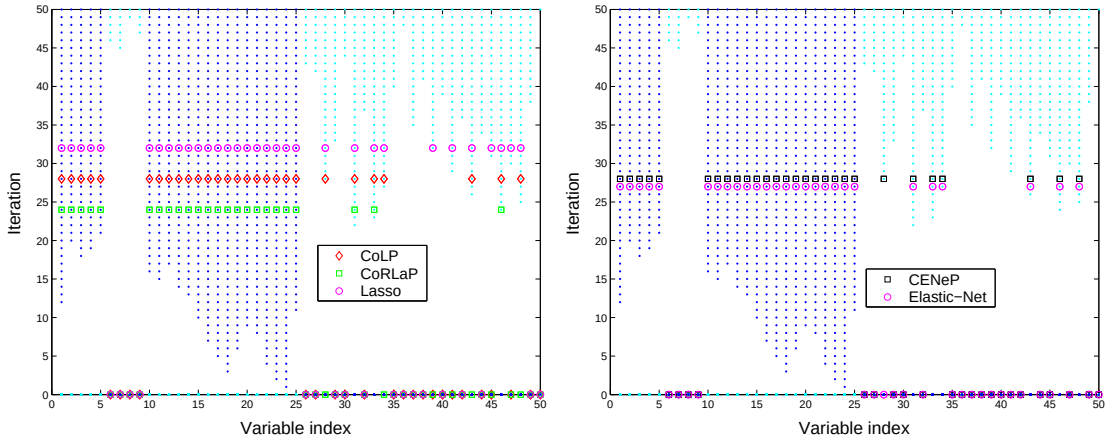

Figure 3: Variable selection analysis for the CoLP, the CoRLaP and the CENeP in Example (b)[300/1] (variables 1 to 5 and 10 to 25 are relevant; see variables in dark blue on the plot). On the left, we consider the CoLP and the CoRLaP selected variables (x-axis) with respect to the LASSO modification of the LARS algorithm iterations (y-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). On the right, we consider the CENeP selected variables (x-axis) with respect to the Elastic-Net modification of the LARS algorithm iterations (y-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The selected iteration is marked by red diamonds for the CoLP, green squares for CoRLaP and black squares for the CENeP.

zone (that is, after iteration 85). Then in order to reduce the gap between SCP and CoRP in the cases $p \geq n$, we suggest to modify the selection criterion in **Step 3** in two ways. i) *Early Stopping CoLP:* do not consider (and do not construct) all the conformal predictors Γ_k^ε. Stop the construction of the predictors Γ_k^ε as soon as the length of Γ_k^ε (predictor at iteration k) has a length at least 10 times larger than Γ_{k-1}^ε; ii) *N Previous Neighbors CoLP:* we can enforce the Early Stopping rule by considering as final predictor: $\Gamma_{opt}^\varepsilon = \bigcup_{j:\, 0 \leq k-j < N} \Gamma_j^\varepsilon$, where k is the index of the (selected) smallest predictor and N is the number of neighbors we consider. Note that this method does not alter selection properties as Γ_k^ε is usually constructed with more variables than Γ_j^ε, $j < k$. It further does not alter a lot the accuracy as the Early Stopping rule ensures that we are in stable zone (cf. Figure 7.1 (right) and Figure 5 (right)). Table 1 sums up the performances of the early-stopped CoLP and the 2-PN CoLP in term of validity. We observe the good adaptation of both methods to the case $p = n$ and we remark that 2-PN CoLP nicely produce valid predictor even in the case $p > n$. This improvement in the term of validity can also be illustrated by Figure 5 (right) where we observe that in Example (b)[50/1], the early-stopped CoLP is valid whereas the original CoLP is not.

Selection. The selection ability of Sparse Conformal Predictors is here in concern. First, note that the selected variables in SCPs are directly linked to the selection ordering through the iterations of the LASSO or Elastic-Net modification of the LARS algorithm. Then, if the used modification of the LARS algorithm fails to recover the true model, we can not hope to get a predictor which contains only the true variables. Figure 7.1 illustrates the evolution of the variable selection of CoLP, CoRLaP and the LASSO on one hand and the CENeP and the Elastic-Net on the other hand, in Example (b)[300/1]. It turns out

that CoLP and CENeP select larger model that expected (that is, some noise variables are selected), as the LASSO and the Elastic-Net do. Moreover CoRLaP uses to select a smaller subset of variables than the CoLP. Then it often produces a better variable selection performance than the other methods. It often provides closer model to the true one. Compared to the LASSO, it seems that the CoLP and the CoRLaP perform better in this example. However, we can not conclude the superiority of the CoLP on the LASSO in term of variable selection. A similar conclusion can be given when we compare the CENeP and the Elastic-Net. Nevertheless, the CENeP seems to select little larger models than the Elastic-Net. Finally, analogously to the superiority of the Elastic-Net compared to the LASSO, we can remark that the CENeP manages to have better selection performances compared to the CoLP and the CoRLaP when a group structure may exist between different variables (for instance in Example (d)$[n/\sigma]$). This is due to the LASSO modification of the LARS algorithm which uses to select some noise variables before relevant ones in such cases.

7.2 Real data

We applied SCPs on 150 randomly permutations of the House Boston dataset[5], in which we randomly choose one row to be the new pair (x_{new}, y_{new}). The original dataset consists of 506 observations with 13 variables. When we consider variable selection, we note that almost all SCPs are constructed without the variable $X_7 = (x_{1,7}, \ldots, x_{505,7})$. This variable is selected with frequencies lower than 3%. The CoRLaP also does not consider the variable X_3 as relevant with a frequence equal to 17%. Conforming to Section 7.1, we would better consider X_3 irrelevant as the CoRLaP uses to produce better performance when variable selection is in concern. Then we conclude that the proportion of non-retail business acres per town and the proportion of owner-occupied units built prior to 1940 do not interfere in the value of owner-occupied homes. We also can notice that variable selection sligtly improved accuracy of conformal predictors in all presented experiments. Here, we can for instance remark that the median lengths of the CoLP, the CoRLaP and the CENeP are respectively 13.61, 13.50 and 13.58, whereas CoRP length is 14.45.

8 Conclusion

We presented Sparse Conformal Predictors, a family of l_1 regularized conformal predictors. We focused on LASSO and Elastic-Net versions of these Sparse Conformal Predictors. We illustrated their performance in term of accuracy, validity and variable selection. We concluded that such Sparse Conformal Predictors are valid and nicely exploit the sparsity of the model when the sample size is larger than the the number of variables (i.e, when $n > p$). We also provided a way to adopt these sparse predictors to the case $p \geq n$ through a pair of rules we called Early Stopping and N Previous Neighbors rules.
Several extensions of this work can be explored such as the construction of SCP with Adaptive LASSO [20] and they will be investigated in future work.

□

Acknowledgement. We would like to thank Professor Arnak Dalalyan and Professor Nicolas Vayatis for insightful comments.

[5]The data and their description are available at http://archive.ics.uci.edu/ml/datasets/Housing.

References

[1] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[2] George Casella and Roger L. Berger. *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression - with discussion. *Ann. Statist.*, 32(2):407–499, 2004.

[4] P. Garrigues and L. El Ghaoui. An homotopy algorithm for the lasso with online observations. *To appear in Neural Information Processing Systems (NIPS) 21*, 2008.

[5] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[6] M. Hebiri. Regularization with the smooth-lasso procedure. *Technical Report*, 2008.

[7] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 2000.

[8] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[9] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3):1012–1030, 2007.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[11] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[12] V. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[13] V. Vovk. Asymptotic optimality of transductive confidence machine. In *Algorithmic learning theory*, volume 2533 of *Lecture Notes in Comput. Sci.*, pages 336–350. Springer, Berlin, 2002.

[14] V. Vovk. On-line confidence machines are well-calibrated. *In: Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*, pages 187–196, 2002.

[15] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. *In: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, 1999.

[16] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005.

[17] V. Vovk, G. Nouretdinov Ilia, and A. Gammerman. On-line predictive linear regression. *Technical Report*, 2007.

[18] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[19] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[20] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[21] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

[22] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007.