Stat Comput (2010) 20: 253–266
DOI 10.1007/s11222-009-9167-2

Sparse conformal predictors

SCP

Mohamed Hebiri

Received: 11 February 2009 / Accepted: 8 December 2009 / Published online: 29 December 2009

Abstract Conformal predictors, introduced by Vovk et al. (Algorithmic Learning in a Random World, Springer, New York, 2005), serve to build prediction intervals by exploiting a notion of conformity of the new data point with previously observed data. 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 the total number of covariates 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 and real data.

Keywords LASSO · LARS · Sparsity · Variable selection · Regularization path · Confidence set

M. Hebiri (✉)
Seminar für Statistik, ETH-Zurich, HG G 18 Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: hebiri@stat.math.ethz.ch

M. Hebiri
Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris 7—Diderot, UFR de Mathématiques, 175 rue de Chevaleret, 75013 Paris, France
e-mail: hebiri@math.jussieu.fr

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. Assume that the pairs (x_i, y_i), $i \geq n$ come from an exchangeable distribution $\mathcal{P}$ in the product space $(\mathbb{R}^p \times \mathbb{R})^\infty$. Suppose also that 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 (Vapnik 1998). Our estimation strategy is based on local arguments in order to produce a better estimation for y_{new} (Györfi et al. 2002). More precisely, we will follow the approach of *conformal prediction* presented by Vovk et al. (2005) 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. (1999). Moreover, Vovk et al. (2005) 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 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 (Tibshirani 1996).

[1] The Conformal Ridge Predictor was called the Ridge Regression Confidence Machine by Vovk et al. (2005).

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 (Knight and Fu 2000; Meinshausen and Bühlmann 2006; Bunea et al. 2007; Zhao and Yu 2006) and several modifications have been proposed (Zou 2006; Yuan and Lin 2006; Zou and Hastie 2005; Tibshirani et al. 2005; Hebiri 2008). One attractive aspect of the LASSO is that it aims both to estimate the regression vector 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 terms 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 (Vovk et al. 2005), 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. First, we concisely introduce conformal prediction and the LASSO procedure in Sects. 2 and 3 respectively. In Sect. 4, we give the explicit form of the Conformal Lasso Predictor. An algorithm producing the CoLP is presented in Sect. 5. Then in Sect. 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 Sect. 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. (2005) where they develop the idea of *conformal* prediction. We aim to predict the label y_{new} corresponding to a new observation $x_n = x_{new}$. To this end, we exploit the similarity of pairs of the form (x_{new}, y) to the former observations (x_i, y_i) for $i = 1, \ldots, n-1$, where $y \in \mathbb{R}$. This is the purpose of introducing a *nonconformity score* $\alpha(y) = (\alpha_1(y), \ldots, \alpha_n(y))'$ which is based on $\mathcal{E}_n$. Given a procedure constructed based on the dataset $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_{new}, y)\}$, each value α_i measures the quality of fit on the example (x_i, y_i). For instance, if we use as procedure the least-squares linear regression, then the nonconformity score can be defined as $\alpha_i = \ell(y_i, \hat{\mu}_i)$, where $\hat{\mu}_i$ stands for the linear fit of y_i provided by the least-square procedure and ℓ is any distance. In order to obtain a relative information between different nonconformity scores α_i, we shall use the notion of *p-value*, as introduced by Vovk et al. (2005), and defined as:

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

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 Sect. 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 (Casella and Berger 2001), consider the following hypotheses:

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

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 by Vovk et al. (2005). 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 $\varepsilon_1, \varepsilon_2 > 0$:

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

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 by Vovk et al. (2007). 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:

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

Note that in an on-line perspective, one also focuses on the cumulative error $\text{ERR}_n^\varepsilon = \sum_{i=1}^n \text{err}_i^\varepsilon$. Asymptotic validity

properties of this cumulative error have been studied by Vovk (2002a) and Vovk et al. (2005, Chaps. 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.

Proposition 1 (Vovk 2002a) *Fix a significance level* $\varepsilon \in (0,1)$. *Let* $\alpha \in \mathbb{R}^n$ *be any nonconformity score. If the distribution* $\mathcal{P}$ *is exchangeable, then the conformal predictor*

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

satisfies

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

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

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 Sect. 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 (Tibshirani 1996) has originally been introduced in the linear regression model:

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

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 penalized version of LASSO estimator is defined as follows:

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

where $\lambda \geq 0$ is a tuning parameter. We also refer to the papers by Chen and Donoho (1995) and Santosa and Symes (1986) for anterior utilization of the above estimator in signal processing and in the deconvolution problem. 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 $\mathcal{A}_\lambda = \{j \in \{1, \ldots, p\} : (\hat{\beta}_\lambda)_j \neq 0\}$. Several effective algorithms to compute $\hat{\beta}_\lambda$, the LASSO solution of the minimization problem (3) have been proposed and studied (for instance Interior Points methods (Kim et al. 2007), homotopy algorithms introduced by Osborne et al. (2000b) with a closed form called the LARS (Efron et al. 2004), Pathwise Coordinate Optimization (Friedman et al. 2007), Relaxed Greedy Algorithms (Huang et al.)). In particular a LASSO modification of the LARS algorithm (Efron et al. 2004) can iteratively provide approximations of the LASSO estimator for a few values of the tuning parameters $\lambda = \lambda_0, \ldots, \lambda_K$ such that $\infty = \lambda_0 > \cdots > \lambda_K = 0$ (the indices refer to the algorithm steps and K denotes the last step). These points are the so-called *transition points*.

Let us introduce some notation. First let $\mathbf{x}_\lambda$ denotes the $(n-1) \times |\mathcal{A}_\lambda|$ matrix whose columns are variables $X_j = (x_{1,j}, \ldots, x_{n-1,j})'$, with indices $j \in \mathcal{A}_\lambda$. Then for $\lambda \geq 0$, we denote by $\bar{\beta}_\lambda$ the estimator defined as the minimizer of (3) over the set $\mathcal{A}_\lambda$. That is

$$\bar{\beta}_\lambda = \operatorname*{argmin}_{b \in \mathbb{R}^{|\mathcal{A}_\lambda|}} (\mathbf{y} - \mathbf{x}_\lambda b)'(\mathbf{y} - \mathbf{x}_\lambda b) + \lambda \sum_{j=1}^{|\mathcal{A}_\lambda|} |b_j|, \tag{4}$$

where $\mathbf{y} = (y_1, \ldots, y_{n-1})'$ and $|\mathcal{A}_\lambda|$ is the cardinality of the set $\mathcal{A}_\lambda$. From now on, let us also write $\bar{\beta}_k$, $\mathcal{A}_k$ and $\mathbf{x}_k$ respectively for the estimator $\bar{\beta}_\lambda$ defined in (4), the sparsity set $\mathcal{A}_\lambda$ and the matrix $\mathbf{x}_\lambda$ evaluated at the transition point $\lambda = \lambda_k$, where $k \in \{1, \ldots, K\}$ is one of the LARS algorithm steps. For each λ_k, we assume that the matrix $(\mathbf{x}_k'\mathbf{x}_k)^{-1}$ is invertible. Obviously, the estimator $\bar{\beta}_k$ is an $|\mathcal{A}_k|$-dimensional vector. Furthermore, we denote by s_k the $|\mathcal{A}_k|$-dimensional sign vector whose components are the signs of the components of the estimator $\bar{\beta}_k$ evaluated at the transition point λ_k (i.e., $(s_k)_j = 1$ if $(\bar{\beta}_k)_j > 0$, $(s_k)_j = -1$ if $(\bar{\beta}_k)_j < 0$ where $j \in \mathcal{A}_k$). Here are some characteristics of the LARS algorithm and we refer to the paper by Efron et al. (2004) 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 when $p \leq n$ with the ordinary least square (OLS) estimator. When $p > n$, the LARS cannot select all p variables. It is limited by the sample size n. In such a case, an OLS solution does not exist and the LARS algorithm would end with a solution which consists in the interpolation of the observed variables with the smallest ℓ_1-norm, a solution of little interest.

(ii) For each $\lambda \in (\lambda_{k+1}, \lambda_k]$, the solution of the minimization problem (4) can be expressed in the following form:

$$\bar{\beta}_\lambda(\mathbf{y}, \mathbf{x}_k, s_k) = (\mathbf{x}_k'\mathbf{x}_k)^{-1}\left(\mathbf{x}_k'\mathbf{y} - \frac{\lambda}{2}s_k\right). \tag{5}$$

Let us mention that the set $\mathcal{A}_k$ and the sign vector s_k remain unchanged when λ varies in the interval $(\lambda_{k+1}, \lambda_k]$. We refer to the paper by Osborne et al. (2000a, Sect. 3) for a good way to define the j-th component of the sign vector s_k evaluated at the transition point λ_k, when X_j is an added variable.

(iii) Clearly, one can compute the LASSO estimator $\hat{\beta}_\lambda$ defined in (3) using the estimator $\bar{\beta}_\lambda(\mathbf{y}, \mathbf{x}_k, s_k)$ given by (5). This is done by setting (if necessary) to zero the components j, with $j \notin \mathcal{A}_\lambda$ in the vector $\hat{\beta}_\lambda$. The remaining components of $\hat{\beta}_\lambda$ coincide with the corresponding components of $\bar{\beta}_\lambda$. As highlighted by (5), the LASSO estimator is piecewise linear in λ (Rosset and Zhu 2007). Using the LASSO modification of the LARS algorithm, this property 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 matches an 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) An important property of the LASSO modification of the LARS algorithm is piecewise linearity. Indeed, let $\lambda \in (\lambda_{k+1}, \lambda_k]$ where λ_{k+1} and λ_k are two successive transition points. In this interval, the LASSO estimator $\hat{\beta}_\lambda$ uses the same variables (variables with indices in $\mathcal{A}_k$). By using (5), it is easy to see Zou et al. (2007), 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 LARS algorithm (more precisely its modified version) 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 LARS algorithm.

4 Sparse predictor with conformal Lasso

For the reasons exposed above section, 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. Sect. 2).

Now let us detail the construction of the CoLP for each λ_k. To this end, denote by $\widetilde{X}_j = (x_{1,j}, \ldots, x_{n-1,j}, x_{new,j})'$, $j = 1, \ldots, p$ the augmented variable j. Define the augmented design matrix $\widetilde{\mathbf{x}} = (x_1, \ldots, x_{n-1}, x_{new})' = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$ 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 Sect. 3, for the fixed λ_k, we also define the estimator $\bar{\beta}_k(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_k, s_k)$ from expression (5) based on these augmented data. That is

$$\bar{\beta}_k(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_k, s_k) = (\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}\left(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{y}} - \frac{\lambda_k}{2}s_k\right). \tag{6}$$

Let us mention that in the above expression, the transition point λ_k and the corresponding sign vector s_k are obtained as described in Sect. 3. In particular, they do not depend on x_{new} nor on y. They are only dependent on the $n-1$ pairs $\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$. From now on, we denote the estimator (6) by $\bar{\beta}_k$. We define $\bar{\mu}_k := \widetilde{\mathbf{x}}_k\bar{\beta}_k$. 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:

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

where $|\cdot|$ is meant here componentwise, $A_k = (a_1^k, \ldots, a_n^k)'$ and $B_k = (b_1^k, \ldots, b_n^k)'$ with

$$\begin{cases} A_k := (\mathbf{I} - \mathbf{H}_k)\,(y_1, \ldots, y_{n-1}, 0)' + \frac{\lambda_k}{2}\widetilde{\mathbf{x}}_k\left(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k\right)^{-1}s_k, \\ B_k := (\mathbf{I} - \mathbf{H}_k)\,(0, \ldots, 0, 1)'. \end{cases} \tag{7}$$

We defined the above nonconformity score based on the absolute difference between the observed and fitted value. In light of (1), let us mention that one can use other measure, as the squared difference for instance, without inducing any modification in the resulting conformal predictor. Note also 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 $i \in \{1, \ldots, n\}$

$$S_i^k = \left\{ y : \alpha_i^k(y) \geq \alpha_n^k(y) \right\}, \tag{8}$$

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 reals defined respectively as

$$l_i^k = \min \left\{ -\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k} \right\}, \tag{9}$$

and

$$u_i^k = \max \left\{ -\frac{a_i^k - a_n^k}{b_i^k - b_n^k}; -\frac{a_i^k + a_n^k}{b_i^k + b_n^k} \right\}, \tag{10}$$

where a_i^k and b_i^k are given by (7).

Proposition 2 *Let us fix $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, n-1\}$. 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 (9) and (10) respectively.*

(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 lose any generality as one can multiply a_i^k and b_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

$$\begin{aligned} \Gamma_k^\varepsilon &= \left\{ y : \sum_{i=1}^n \mathbb{I}(\alpha_i^k(y) \geq \alpha_n^k(y)) \geq n\varepsilon \right\} \\ &= \left\{ y : \sum_{i=1}^n \mathbb{I}(S_i^k)(y) \geq n\varepsilon \right\}, \end{aligned} \tag{11}$$

where $\mathbb{I}(\cdot)$ stands for the indicator function. The above approach leads to a whole collection of confidence intervals $\Gamma_1^\varepsilon, \ldots, \Gamma_K^\varepsilon$. We propose below a strategy for choosing one particular Γ_k^ε, the performance of which will be studied in Sect. 7 through numerical experiments.

It is worth mentioning that in view of the paper by Vovk (2002b, Theorem 1) (see also the book by Vovk et al. 2005, Proposition 2.3, p. 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 3 *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 $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ and let us define $\alpha^\lambda(y) = |\widetilde{\mathbf{y}} - \widetilde{\mathbf{x}}\hat{\beta}_{\lambda,n}(y)|$. If the distribution $\mathcal{P}$ is exchangeable, 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 $n \in \mathbb{N}$.

In the proof of Proposition 3 detailed by Vovk (2002b), one needs the exchangeability of $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ and the last pair (x_n, y) in the definition of the predictor. Actually, 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 3 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 restrict 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

$$\Gamma_{opt}^\varepsilon = \Gamma_\nu^\varepsilon, \quad \nu = \operatorname*{argmin}_k |\Gamma_k^\varepsilon|. \tag{12}$$

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

Algorithm 1: Lasso Conformal Predictor

Step 1: Normalize the dataset $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}))$ such that for any $j \in \{1, \ldots, p\}$, we have $\sum_{i=1}^n x_{i,j} = 0$, $n^{-1} \sum_{i=1}^n x_{i,j}^2 = 1$ and $\sum_{i=1}^n y_i = 0$. Run the LASSO modification of the LARS algorithm on this normalized dataset

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 (7). 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 (9)–(10) to U^k

end

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

Add $l_i^k = u_i^k$ (9)–(10) 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

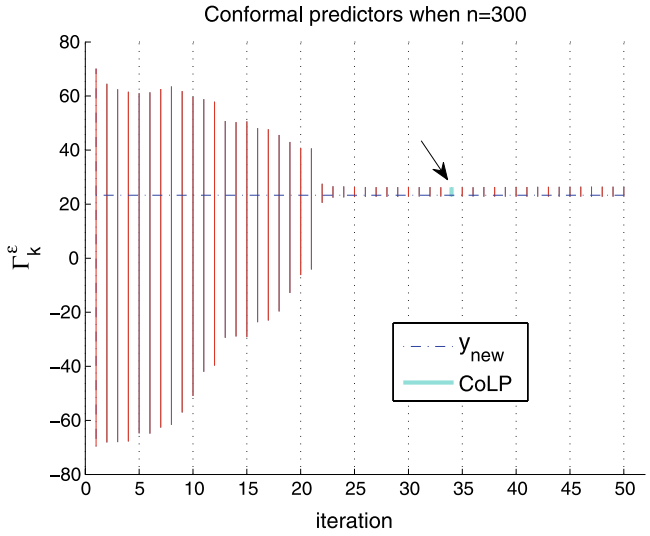
$$\Gamma_k^\varepsilon = \cup_{j:\frac{N_j}{n} > \varepsilon} [y_{(j)}, y_{(j+1)}]$$

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

We provide here a three-step algorithm which enables us to easily construct the CoLP. After a convenient normalization, we start in **Step 1** by applying the LASSO adaptation 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 2, for each λ_k, we can construct the sets S_i^k for $i = 1, \ldots, n$ defined by (8). 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. 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 3, 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 Fig. 1 (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 of variables (i.e., $n > p$) but may be not appropriated when $p \geq n$. Figure 1(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 Fig. 1(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 Sect. 7. Let us add that we only illustrated the validity of the conformal predictors in Fig. 1 (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 Sect. 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 the paper by Garrigues and El Ghaoui (2008) and more recently by Langford et al. (2009) and Shalev-Shwartz and Tewari (2009). The analysis of such on-line versions is the object of work under progress.



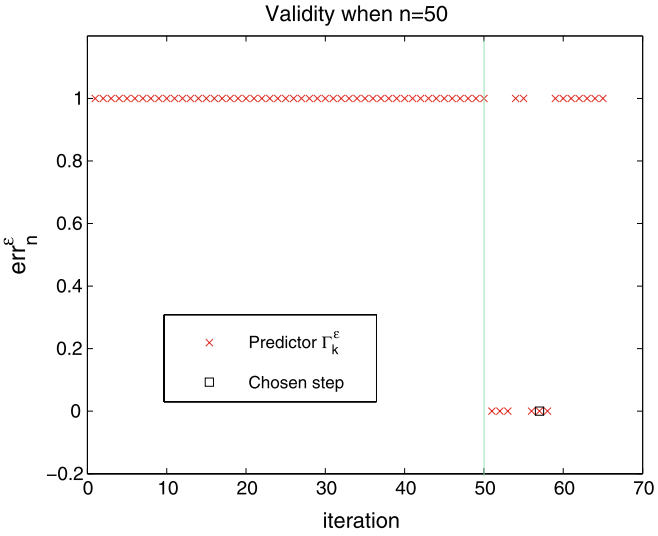


Fig. 1 (Color online) *Left*: Conformal predictors Γ_k^ε evolution through the iterations of the LASSO modification of the LARS algorithm when $n = 300$ (the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}). The CoLP is drawn in a large cyan line. It corresponds to the 34-th iteration and is marked by the arrow. The *horizontal dashed 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

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 methods as the Elastic-Net (Zou and Hastie 2005) and the Smooth-Lasso (Hebiri 2008). As for CoLP (Sect. 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 (Rosset and Zhu 2007) (to reduce computational effort by using a modification of the LARS algorithm).

We use the same notation as in Sect. 3 for the LASSO estimator. Set $\hat{\beta}(\mathbf{x}, \mathbf{y})$ 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 $\hat{\beta}(\mathbf{x}, \mathbf{y})$ (the components of s can equal zero if the corresponding regression coefficients equal zero). On the other hand, using the notation in Sect. 4, we set $\hat{\mu} = \widetilde{\mathbf{x}}\hat{\beta}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ where this time $\hat{\beta}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is based on the augmented dataset $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$.

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

$$\hat{\mu} = U(\widetilde{\mathbf{x}}, s)\widetilde{\mathbf{y}} + V(\widetilde{\mathbf{x}}, s), \tag{13}$$

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 (**Step 1** in Algorithm 1). Nevertheless, let us first mention that the functions $U(\cdot)$ and $V(\cdot)$ in the CoLP construction induce the dataset $\widetilde{\mathbf{x}}_k$ instead of $\widetilde{\mathbf{x}}$, where k is one step of the LASSO modification of the LARS algorithm. Then these functions map into a smaller space. For this reason, we will note for such iterative methods $u(\cdot)$ and $v(\cdot)$ instead of $U(\cdot)$ and $V(\cdot)$, but we mention that U and V can easily be reconstituted using u and v by adding, if necessary zeros to the proper places. Now let $\hat{\beta}$ be the Elastic-Net estimator (Zou and Hastie 2005). Based on the dataset $(\mathbf{x}, \mathbf{y})$, this estimator is defined by

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

where $\lambda \geq 0$ and $\nu \geq 0$ are two tuning parameters. We note s its sign vector. Similarly to the LASSO we use a modification (the Elastic-Net here) of the LARS algorithm to iteratively compute $K > 0$ solutions $\bar{\beta}_1, \ldots, \bar{\beta}_K$ (analogues to (5)), which easily provide K solutions of the above minimization problem by setting to zero the components corresponding to non active variables. Note that for each $k \in \{1, \ldots, K\}$, the vector $\bar{\beta}_k$ is a $|\mathcal{A}_k|$-dimensional vector given

by

$$\bar{\beta}_k = \mathbf{x}_k(\mathbf{x}_k'\mathbf{x}_k + \nu_k\mathbf{I}_k)^{-1}\mathbf{x}_k'\mathbf{y} - \lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k,$$

where s_k is the sign vector at the k-th step of the Elastic-Net modification of the LARS algorithm and λ_k and ν_k are the tuning parameters evaluated at this step. Finally, following the same reasoning as for the LASSO, Assumption 1 holds with the functions $u(\widetilde{\mathbf{x}}_k, s_k) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \nu_k\mathbf{I}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}_k, s_k) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$, where $\mathbf{I}_k$ is the $|\mathcal{A}_k| \times |\mathcal{A}_k|$ identity matrix. In the same way, we can also define the Conformal Smooth Lasso Predictor (CoSmoLaP) based on a Smooth-Lasso modification of the LARS algorithm (Hebiri 2008). Here $u(\widetilde{\mathbf{x}}_k, s_k) = \widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k + \nu_k\mathbf{J}_k)^{-1}\widetilde{\mathbf{x}}_k'$ and $v(\widetilde{\mathbf{x}}_k, s_k) = -\lambda_k\widetilde{\mathbf{x}}_k(\widetilde{\mathbf{x}}_k'\widetilde{\mathbf{x}}_k)^{-1}s_k$. 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 correlation between successive variables.

As for CoLP, we can define the nonconformity score $\alpha = (\alpha_1, \ldots, \alpha_n)'$ of an expected label y associated to the estimator $\hat{\mu}$ as follows:

$$\begin{aligned}\alpha(y) &:= |\widetilde{\mathbf{y}} - \hat{\mu}| = |\,(\mathbf{I} - U(\widetilde{\mathbf{x}}, s))\,\widetilde{\mathbf{y}} - V(\widetilde{\mathbf{x}}, s)|\\ &= |A + B\,y|,\end{aligned}$$

where $A = (a_1, \ldots, a_n)'$ and $B = (b_1, \ldots, b_n)'$ with

$$\begin{cases} A := (\mathbf{I} - U(\widetilde{\mathbf{x}}, s))\,(y_1, \ldots, y_{n-1}, 0)' - V(\widetilde{\mathbf{x}}, s),\\ B := (\mathbf{I} - U(\widetilde{\mathbf{x}}, s))\,(0, \ldots, 0, 1)', \end{cases}$$

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.

To evaluate the computational cost of the proposed algorithm, three main points should be taken into account. 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 Sect. 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 question of the choice of the grid and the window of this grid.

On the other hand, we saw that the LARS algorithm allows 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 by Vovk et al. (2005, Chaps. 2.3 and 4.1). It turns out that sparse conformal predictors and in particular the CoLP require computation time $\mathcal{O}(n^2)$ which can be further reduced to $\mathcal{O}(n\log(n))$.

7 Experimental results

In this section we present the experimental performance of the Sparse Conformal Predictors (SCP) with respect to their validity, their accuracy and also their selection power. As a benchmark, we use the CoRP[2] for its validity and accuracy and the original LASSO and Elastic-Net estimators for their selection[3] power.

Three SCPs are considered: the Conformal Lasso Predictor (CoLP was introduced in Sects. 4 and 5) and the Conformal Elastic Net Predictor (CENeP was described in Sect. 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 Sect. 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 Synthetic data

We consider four simulated datasets from the linear regression model

$$y = \mathbf{X}'\beta + \sigma\xi,$$

with $\beta \in \mathbb{R}^{50}$, and

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

[2]We construct the CoRP associated to same tuning parameters as the CoLP (i.e., the transition points λ_k observed in Sect. 5). Note that the performance would not be altered as conformal predictors according to this method are almost embedded and changes sensitively while the tuning parameter varies (Vovk et al. 2005, p. 39).

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

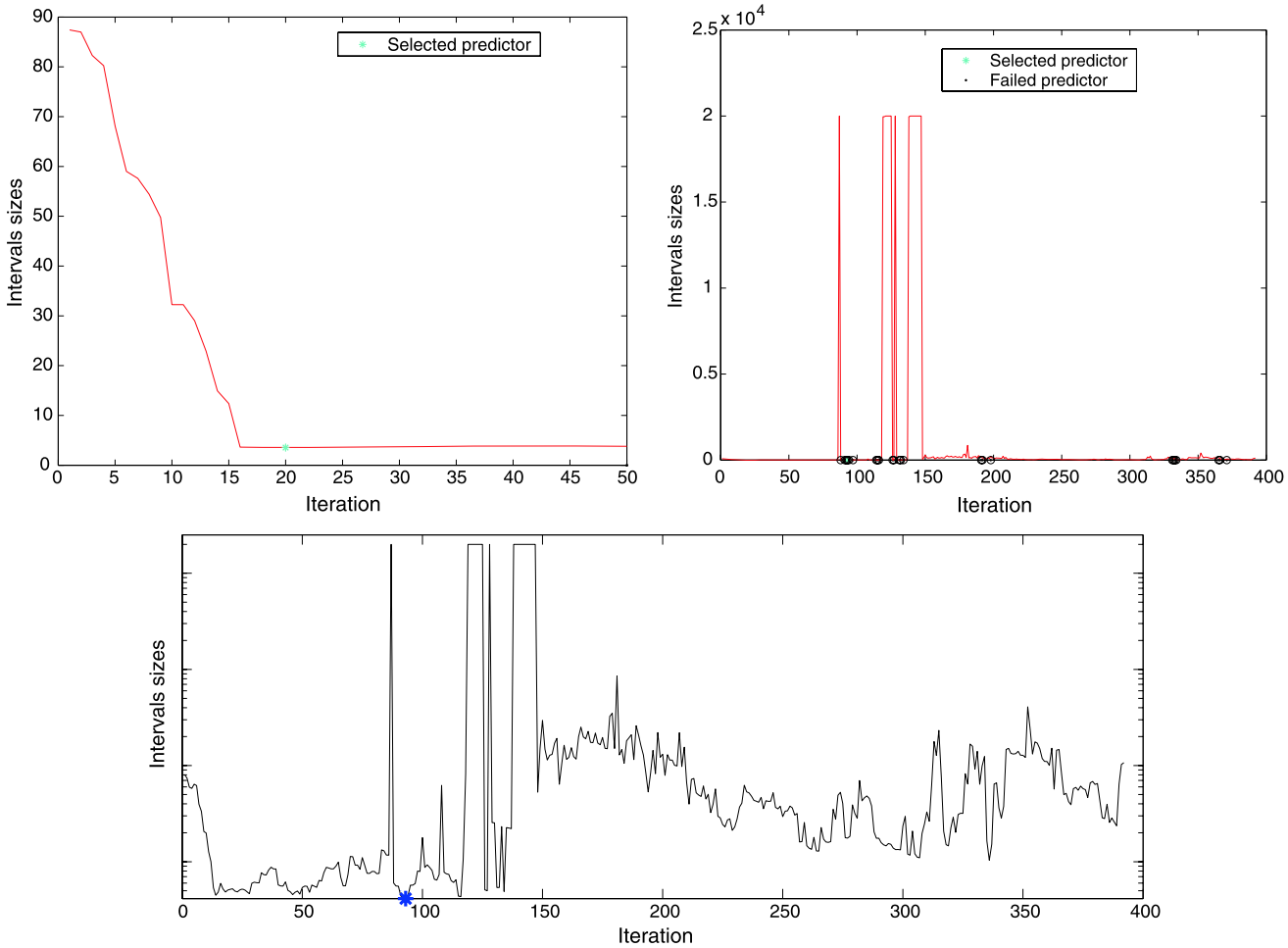


Fig. 2 (Color online) Analysis of conformal predictors length (y-axis) through the LASSO modification of the LARS algorithm iterations (x-axis: the first iteration corresponds to λ_{max} and the last one corresponds to λ_{min}) in Example (c)[300/1] (*top left*) and in Example (c)[50/1] (*top right*). 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 associated to the same Example (c)[50/1] displayed in the *top right panel*

Hence $p = 50$ through the simulations. Noise level σ and the sample size n are left 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 correlation 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.* 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: $\Sigma_{j,j} = 1$ for every $j \in \{1,\ldots,p\}$; for $j \neq k$, $\Sigma_{j,k} \approx 1$ (actually $\Sigma_{j,k} = \frac{1}{1+0.01}$, due to an extra noise variable) when (j,k) belongs to $\{1,\ldots,5\}^2$, $\{6,\ldots,10\}^2$ and $\{11,\ldots,15\}^2$ and zero otherwise.

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

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_k^ε, $k = 1,\ldots,K$ obtained at the end of **Step 2** in Algorithm 1 described in Sect. 5. We recall that each of these predictors is associated to an iteration of a modification of the LARS algorithm, that is the transition points λ_k, $k = 1,\ldots,K$. Figure 2 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_k^εs sensitively changes from

Table 1 Mean lengths [with precision ±95%] of the CoRP, CoLP, CoRLaP, CENeP, the Early-Stopped CoLP and the 2-PN CoLP based on 500 replications

EXAMPLE	σ	CORP	COLP	CORLAP	CENEP
(a)[300/σ]	1	3.7 ± 0.1	3.2 ± 0.1	3.1 ± 0.1	3.2 ± 0.1
(a)[50/σ]	3	13.4 ± 0.3	7.4 ± 0.3	4.7 ± 0.3	7.1 ± 0.3
(b)[300/σ]	1	8.5 ± 0.1	3.4 ± 0.1	3.3 ± 0.1	3.4 ± 0.1
(b)[50/σ]	1	20.3 ± 0.1	3.9 ± 0.1	2.3 ± 0.1	3.7 ± 0.1
(b)[20/σ]	1	101.2 ± 0.1	52.5 ± 0.1	17.2 ± 0.1	37.6 ± 0.1
(c)[300/σ]	1	3.9 ± 0.1	3.4 ± 0.1	3.2 ± 0.1	3.3 ± 0.1
(c)[300/σ]	3	11.0 ± 0.3	10.1 ± 0.3	9.6 ± 0.3	9.7 ± 0.3
(c)[300/σ]	10	34.3 ± 0.9	33.0 ± 0.9	31.8 ± 0.9	32.2 ± 0.9
(d)[300/σ]	10	286.5 ± 0.9	70.2 ± 0.9	36.0 ± 0.9	54.0 ± 0.9

EXAMPLE	σ	CORP	COLP	STOPPED-COLP	2-PN-COLP
(a)[50/σ]	3	13.1 ± 0.3	7.2 ± 0.3	9.1 ± 0.3	9.5 ± 0.3
(b)[50/σ]	1	20.7 ± 0.1	3.9 ± 0.1	5.5 ± 0.1	5.9 ± 0.1
(b)[50/σ]	10	55.3 ± 0.9	32.1 ± 0.9	46.2 ± 0.9	48.7 ± 0.9
(c)[20/σ]	3	28.3 ± 0.3	7.44 ± 0.3	13.2 ± 0.3	14.1 ± 0.3
(d)[20/σ]	10	233.0 ± 0.9	115.3 ± 0.9	164.1 ± 0.9	170.2 ± 0.9

one iteration to the following and that the larger predictor has a reasonable length compared 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 Fig. 1(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 cannot 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 strong 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 Fig. 2(right), the selected predictor at iteration 93 is not valid). This aspect can also be observed in Fig. 1(right) (the graph corresponds to Example (b)[50/1]) where the selected CoLP at iteration 57 is not valid. Similar strong variations of the corresponding predictors lengths would have been observed after iteration 49 if we have provided a graph as Fig. 2 (right).

Now let us compare the accuracy of the final conformal predictors obtained at the end of **Step 3** in Algorithm 1 while using the different methods or different values for the setting parameters. Table 1 sums up the obtained results. First of all, an important remark is that all Sparse Conformal Predictors (CoLP, CoRLaP, CENeP, ...) are more accurate than the CoRP. Indeed, the length of the SCPs are most of the time more than twice smaller than the CoRP one. However when we treat problems with both small level of noise and big sample size, it happens that the gain of accuracy is limited as can be seen in Example (a)[300/1] and Example (c)[300/1]. In such situations, one should all the same mention that all provided conformal predictors are accurate. Through these observations we conclude that SCPs exploit favorably the sparsity in order to improve the accuracy of conformal predictors. Comparing the accuracy of the CoLP, the CoRLaP and the CENeP, it turns out that the CoRLaP is the more accurate SCP, whereas the CoLP is the less accurate one. In the other hand, let us now consider the influence of the setting parameters on the accuracy. It seems to be clear that the smaller the sample size is or the higher the noise level is, the larger the length of the conformal predictors is (see Example (b)[n/1] and Example (c)[300/σ] respectively). Noise level and sample size seems to be the more influential parameters on the accuracy of the predictors. Finally except for the case where the model is not sparse (Example (d)[n/σ]), one can observe that the sparsity is not a crucial parameter on the accuracy. This can be illustrated through the obtained results on Example (a)[300/1], Example (b)[300/1] (for which the dataset is built with the same correlation matrix as in Example (a)) and Example (c)[300/1].

Validity. Now, we consider the validity of the selected predictors (cf. **Step 3** in Algorithm 1). As shown in Table 2, we observe that variations on the noise level, the variables correlations and the sparsity of the model do not

Table 2 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	0.899 ± 0.019	0.886 ± 0.020	0.854 ± 0.022	0.882 ± 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	0.895 ± 0.019	0.889 ± 0.020	0.892 ± 0.019
(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	0.854 ± 0.022	0.624 ± 0.030	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

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 2. As pointed in the accuracy part, one explication can be observed in Fig. 2 as the selected predictor which also is not valid (iteration 93) corresponds to an iteration in the unstable 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. Table 1 exemplifies the performance of these corrected versions of the CoLP according to their accuracy. Obviously both of the above rules provide slightly larger predictors than the original CoLP. In the same time we observe that the Early Stopping CoLP and 2 Previous Neighbors CoLP are still much more accurate than the CoRP. Note further that N Previous Neighbors rule does not alter a lot the accuracy of the Early Stopping CoLP (see Table 1). This is due to the fact that the Early Stopping rule ensures that we are in the stable zone (cf. Fig. 2(right) and Fig. 1(right)). Moreover, note that the N Previous Neighbors rule does not neither alter the selection properties of the Early Stopping CoLP as Γ_k^ε is usually constructed with more variables than Γ_j^ε when $j < k$. Finally Table 2 sums up the performance 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 Fig. 1(right) where we observe that in Example (b)[50/1], the early-stopped CoLP is valid whereas the original CoLP is not.

Selection. Here, we are concerned by the selection ability of Sparse Conformal Predictors. First of all, 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 cannot hope to get a predictor which contains only the true variables. Figure 3 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 to the superiority of the CoLP on the LASSO in term of variable selection. A similar conclusion can be

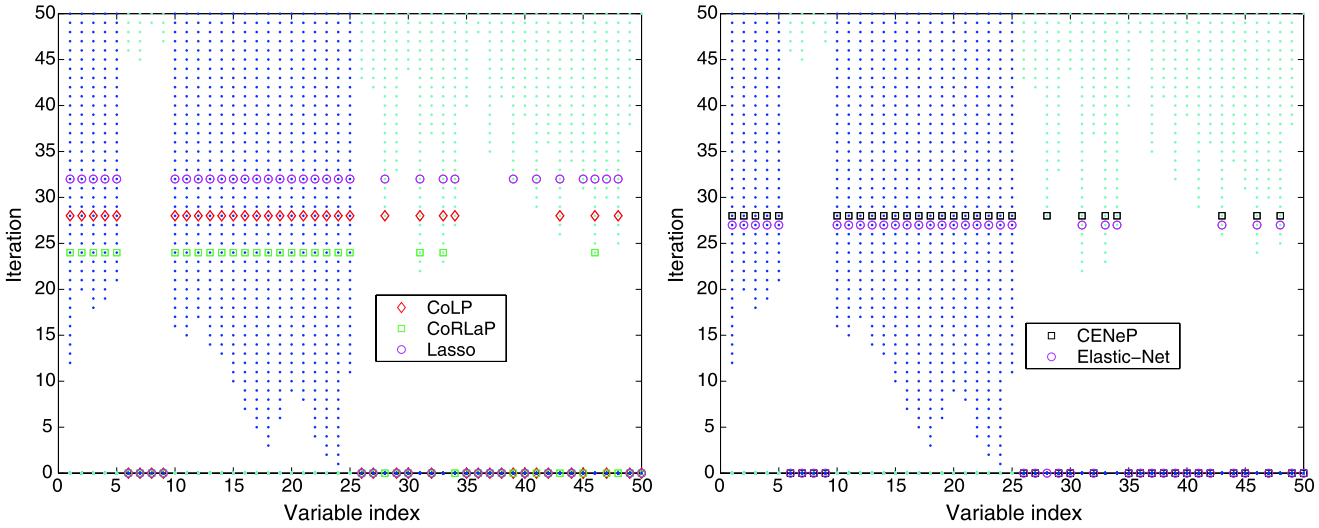


Fig. 3 (Color online) Variable selection analysis for the CoLP, the CoRLaP and the CENeP in Example (b)[300/1] (variables 1 to 5 and 10 to 25 are relevant; see variables in *dark blue* on the plot). On the *left*, we consider the CoLP and the CoRLaP selected variables (x-axis) with respect to the LASSO 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

Table 3 Selection frequency of each variable by the different SCPs on 150 random permutations of the House Boston dataset with ($p = 13$ and $n = 506$)

Variable	$X_1, X_6, X_{11} \to X_{13}$	X_2	X_3	X_4	X_5	X_7	X_8	X_9	X_{10}
CoLP	1	1	0.90	1	1	0	1	1	1
CoRLaP	1	1	0.17	1	1	0.01	1	0.99	0.99
CENeP	1	0.99	0.80	0.99	0.99	0.03	0.99	0.99	1

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/σ]). 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 apply SCPs on 150 random permutations of the House Boston dataset,[4] 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. First Table 3 displays the obtained variable selection results. 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 frequency equal to 17%. Conforming to Sect. 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. A general observation is that variable selection improves the accuracy of conformal predictors (as already seen in Table 1). Here, 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.

To consider a high dimensional setting we use the same trick as in the paper by Bühlmann and Hothorn (2010). For this purpose, we look at a synthetically enlargement of the Boston Housing dataset. We add 483 additional, ineffective noise predictor variables $X_{add} \sim \mathcal{N}_{483}(0, \mathbf{I})$. The new design matrix X has then $p = 500$ columns or variables with at most 13 effective predictors. We fix this ma-

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

Table 4 Validity frequencies (with precision ±95%) and noise variables selection (variables X_{14} to X_{500}) of the CoRP, CoLP, CENeP, the Early-Stopped CoLP and the 2-PN CoLP based on the augmented Boston Housing dataset ($p = 500$ and $n = 50$)

	CoRP	CoLP	CENeP	Stopped-CoLP	2-PN-CoLP
Validity	0.93 ± 0.01	0.43 ± 0.04	0.85 ± 0.02	0.85 ± 0.02	0.93 ± 0.01
Noise	100%	20.3%	4.0%	5.9%	5.9%

trix. In the sequel, we arbitrarily chose 50 examples among the available 506 rows of this matrix X. As a consequence we have constructed two datasets: a training set with 50 instances and a test set with 456 instances. Each instance is a 500-dimensional vector. In this high dimensional setting, we apply the SCPs on 100 random permutations of the training dataset. Each time, we randomly chose one row in the test dataset to be the new pair (x_{new}, y_{new}). We study the behavior of the CoRP, the CoLP, the CENeP, the early-stopped-CoLP and of the 2-PN CoLP in such a framework.

As observed on synthetic data, all of the CENeP, the early-stopped-CoLP and the 2-PN CoLP have better performance than the original CoLP (see Table 4). We also observe that the better performance is reached by the 2-PN CoLP and the CoRP with a validity equal to 0.93 (this is better than the expected validity level). However, an important point is that the 2-PN CoLP has also the advantage of producing a sparse predictor whereas the CoRP does not.

As for the accuracy, let us remark that the lengths of the SCPs are much smaller than the CoRP length. Indeed the median lengths of the CoLP, the CENeP, the early-stopped-CoLP and the 2-PN CoLP are respectively 1.5, 8, 8 and 8, whereas the CoRP's length equals 19. This observation advocates for the use of the 2-PN CoLP.

According to the selection task, we observe in Table 4 (last line) that at most 6% of the additional noise variables are selected by the SCPs (except the CoLP which selects more than 20% of these irrelevant variables). Concerning the original variables, only X_1, X_4, X_6, X_{11}, X_{12} and X_{13} are selected. This at least confirms that the proportion of non-retail business acres per town and the proportion of owner-occupied units built prior to 1940 (X_3 and X_7 respectively) do not interfere in the value of owner-occupied homes as observed above in the case $p \leq n$.

Remark 5 For comparison, Sparse Conformal Predictors have also been applied on the same setting as above but without the 483 additional noise predictor variables. It turns out that also in this situation the 2-PN CoLP has a validity frequency equal to 0.92 which is larger than expected (0.9). The 2-PN CoLP seems to provide better performance than the CoRP and the CoLP in this dataset. Moreover, the same variables X_1, X_4, X_6, X_{11}, X_{12} have been considered as relevant.

8 Conclusion

In this paper, we introduced a new family of l_1 regularized conformal predictors termed Sparse Conformal Predictors. We then focused on LASSO and Elastic-Net versions of these Sparse Conformal Predictors and illustrated their performance in terms of accuracy, validity and variable selection. The experiments reported in the paper show that SCPs are valid and nicely exploit the sparsity of the model when the sample size is larger than 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. It turns out that a 2 Previous Neighbors rule is really attractive. Indeed, even in a high dimensional setting, it allows to achieve good performance for all of the criteria: *validity, accuracy* and *selection*.

Several extensions of this work can be explored such as the construction of SCP with Adaptive LASSO (Zou 2006) or the adaptation of SCPs in the generalized l_1 regularized linear model, using algorithmic developments presented for instance by Park and Hastie (2007). These topics, as well as the combination of the conformal predictors with other sparsity inducing procedures, such as the exponentially weighted aggregate (Dalalyan and Tsybakov 2007, 2008) or grouped variable Lasso (Yuan and Lin 2006; Chesneau and Hebiri 2008), are interesting avenues for future research. □

Acknowledgements We would like to thank the Referees for their helpful comments and suggestions. They helps us to improve significantly this revised version of the paper. We also would like to thank Professors Arnak Dalalyan and Nicolas Vayatis for insightful comments.

References

Bühlmann, P., Hothorn, T.: Twin boosting: improved feature selection and prediction. Stat. Comput. (2010, this issue)

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

Casella, G., Berger, R.L.: Statistical Inference. Duxbury, N. Scituate (2001)

Chen, S.S., Donoho, D.L.: Atomic decomposition by basis pursuit. Technical Report (1995)

Chesneau, Ch., Hebiri, M.: Some theoretical results on the grouped variables Lasso. Math. Methods Stat. **17**, 317–326 (2008)

Dalalyan, A., Tsybakov, A.: Aggregation by exponential weighting and sharp oracle inequalities. In: Learning Theory. Lecture Notes in Comput. Sci., vol. 4539, pp. 97–111. Springer, Berlin (2007)

Dalalyan, A., Tsybakov, A.: Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. Mach. Learn. **72**, 39–61 (2008)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression—with discussion. Ann. Stat. **32**, 407–499 (2004)

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**, 302–332 (2007)

Garrigues, P., El Ghaoui, L.: An homotopy algorithm for the lasso with online observations. In: Neural Information Processing Systems (Nips), vol. 21, pp. 489–496. MIT Press, Cambridge (2008)

Györfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics. Springer, New York (2002)

Hebiri, M.: Regularization with the smooth-lasso procedure. Technical Report (2008)

Huang, C., Cheang, G.L.H., Barron, A.: Risk of penalized least squares, greedy selection and l1 penalization for flexible function libraries. Preprint (2008)

Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: An interior-point method for large-scale l1-regularized least squares. IEEE J. Sel. Top. Signal Process. **1**, 606–617 (2007)

Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356–1378 (2000)

Langford, J., Li, L., Zhang, T.: Sparse online learning via truncated gradient. J. Mach. Learn. Res. **10**, 777–801 (2009)

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

Osborne, M., Presnell, B., Turlach, B.: On the LASSO and its dual. J. Comput. Graph. Stat. **9**, 319–337 (2000a)

Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. IMA J. Numer. Anal. **20**, 389–403 (2000b)

Park, M.Y., Hastie, T.: L_1-regularization path algorithm for generalized linear models. J. R. Stat. Soc., Ser. B, Stat. Methodol. **69**, 659–677 (2007)

Rosset, S., Zhu, J.: Piecewise linear regularized solution paths. Ann. Stat. **35**, 1012–1030 (2007)

Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. **7**, 1307–1330 (1986)

Shalev-Shwartz, S., Tewari, A.: Stochastic methods for ℓ_1 regularized loss minimization. In: Proceedings of the 26th International Conference on Machine Learning. Omnipress, Montreal (2009)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B **58**, 267–288 (1996)

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

Vapnik, V.: Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, New York (1998)

Vovk, V.: Asymptotic optimality of transductive confidence machine. In: Algorithmic Learning Theory. Lecture Notes in Comput. Sci., vol. 2533, pp. 336–350. Springer, Berlin (2002a)

Vovk, V.: On-line confidence machines are well-calibrated. In: Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science, pp. 187–196. IEEE Computer Society, Los Alamitos (2002b)

Vovk, V., Gammerman, A., Saunders, C.: Machine-learning applications of algorithmic randomness. In Proceedings of the 16th International Conference on Machine Learning, pp. 444–453. ICML (1999)

Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)

Vovk, V., Nouretdinov Ilia, G., Gammerman, A.: On-line predictive linear regression. Technical Report (2007)

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

Zhao, P., Yu, B.: On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)

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

Zou, H., Hastie, T., Tibshirani, R.: On the "Degrees of Freedom" of the lasso. Ann. Stat. **35**, 2173–2192 (2007). URL citeseer.ist.psu.edu/766780.html