Signed-Error Conformal Regression

Henrik Linusson[1], Ulf Johansson[1], and Tuve Löfström[1]

School of Business and Informatics, University of Borås, Borås, Sweden*
{henrik.linusson, ulf.johansson, tuve.lofstrom}@hb.se

Abstract. This paper suggests a modification of the Conformal Prediction framework for regression that will strengthen the associated guarantee of validity. We motivate the need for this modification and argue that our conformal regressors are more closely tied to the actual error distribution of the underlying model, thus allowing for more natural interpretations of the prediction intervals. In the experimentation, we provide an empirical comparison of our conformal regressors to traditional conformal regressors and show that the proposed modification results in more robust two-tailed predictions, and more efficient one-tailed predictions.

Keywords: Conformal Prediction, prediction intervals, regression

1 Introduction

Conformal Prediction (CP) [1] is a framework for producing reliable confidence measures associated with the predictions of an underlying classification or regression model. Given a confidence level $\delta \in (0, 1)$, a conformal predictor outputs prediction regions that, in the long run, contain the true target value with a probability of at least $1 - \delta$. Unlike Bayesian models, CP does not rely on any knowledge of the *a priori* distribution of the problem space; and, compared to the PAC learning framework, CP is much more resilient to noise in the data.

Clearly, the motivation for using CP is the fact that the resulting prediction regions are guaranteed to be valid. With this in mind, it is vital to fully understand what validity means in a CP context. Existing literature (e.g. [1]) provides a thorough explanation of how the validity concept relates to conformal classification, but leaves something to be desired regarding conformal regression.

In this paper, we identify an inherent but non-obvious weakness associated with the most common type of inductive conformal regressor — conformal regressors where the nonconformity score is based on the absolute error of a predictive regression model (Abs. Error CP Regression, or AECPR). Specifically we show

* This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (IIS11-0053) and the Knowledge Foundation through the project Big Data Analytics by Online Ensemble Learning (20120192). The final publication is available at link.springer.com: http://link.springer.com/chapter/10.1007/978-3-319-06608-0_19

that when the underlying model has a skewed error distribution, AECPR produces *unbalanced* prediction intervals — prediction intervals with no guarantee regarding the distribution of errors above and below the prediction interval — and argue that this limits the expressiveness of AECPR models. We suggest to instead produce two one-tailed conformal predictors, one for the low end of the prediction interval and one for the high end. The only modification needed is to use a nonconformity score based on the signed error of the underlying model. Once we have these two conformal predictors, they can either be used to output valid one-tailed predictions, or combined to create two-tailed prediction intervals that exhibit a stronger guarantee of validity than standard AECPR prediction intervals. In addition, we show that the suggested approach is more robust, i.e., less sensitive to outliers. In particular when δ is small, AECPR may be seriously affected by outliers, resulting in very conservative (large) prediction intervals.

2 Background

CP was originally introduced in [2], and further developed in [3], as a transductive approach for associating classification predictions from Support Vector Machine models with a measure of confidence. Vovk, Gammerman & Shafer provide a comprehensive guide on conformal classification in [1], and Shafer & Vovk provide an abridged tutorial in [4]. Since its introduction, CP has been frequently applied to predictive modeling and used in combination with several different classification and regression algorithms, including Ridge Regression [5] k-Nearest Neighbors [6], Artificial Neural Networks [7, 8] and Evolutionary Algorithms [9].

In [10] and [5], Papadopoulos proposes a modified version of CP based on inductive inference called Inductive Conformal Prediction (ICP). In ICP, only one predictive model is generated, thus avoiding the relative computational inefficiency of (transductive) conformal predictors.

Conformal predictors have been applied to a number of problems where confidence in the predictions is of concern, including prediction of space weather parameters [8], estimation of software project effort [11], early diagnostics of ovarian and breast cancers [12], diagnosis of acute abdominal pain [13] and assessment of stroke risk [14].

2.1 Conformal Prediction

Given a set of training examples $Z = ((x_1, y_1), ...(x_l, y_l))$, and a previously unseen input pattern x_j, the general idea behind CP is to consider each possible target value $\tilde{y}$ and determine the likelihood of observing $(x_j, \tilde{y})$ in Z.

To measure the likelihood of observing $(x_j, \tilde{y})$ in Z, a conformal predictor first assigns a *nonconformity score* $\alpha_i^{\tilde{y}}$ to each instance in the extended set $\hat{Z} = Z \cup \{(x_j, \tilde{y})\}$. This nonconformity score is a measure of the strangeness of each instance $(x_i, y_i) \in \hat{Z}$ compared to the rest of the set, and is, in a predictive modeling scenario, often based on the predictions from a model generated using a traditional machine learning algorithm, referred to as the *underlying model* of

the conformal predictor. The underlying model is trained using $\hat{Z}$ as training data, and the nonconformity score for an instance $(x_i, y_i) \in \hat{Z}$ is defined as the level of disagreement (according to some error measure) between the prediction of the underlying model $\hat{y}_i$ and the true label y_i.

The nonconformity score $\alpha_j^{\tilde{y}}$ is compared to the nonconformity scores of all other instances in $\hat{Z}$ to determine how unusual $(x_j, \tilde{y})$ is according to the nonconformity measure used. Specifically, we calculate the p-value of $\tilde{y}$ using

$$p(\tilde{y}) = \frac{\#\left\{z_i \in \hat{Z} \mid a_i \geq a_j^{\tilde{y}}\right\}}{l+1} \quad . \tag{1}$$

A key property of conformal prediction is that if $p(\tilde{y})$ is below some threshold δ, the likelihood of $\tilde{y}$ being the true label for x_j is at most δ if $\hat{Z}$ is i.i.d. If we select δ to be very low, e.g., 0.05, we can thus conclude that if $p(\tilde{y}) < 0.05$ it is at most 5% likely that $\tilde{y}$ is the true label for x_j. These p-values are calculated for each tentative label $\tilde{y}$, and the conformal predictor outputs a prediction region containing each label $\tilde{y}$ for which $p(\tilde{y}) > \delta$; i.e., a set of labels that contains the true label of x_j with probability $1-\delta$. Given that we already know the probability of any prediction region containing the true output for a test instance x_j, the goal in CP is not to maximize this probability, but rather to minimize the size of the prediction regions. In essence, CP performs a form of hypothesis testing. For each label $\tilde{y}$, we want to reject the null hypothesis that $(x_j, \tilde{y})$ is conforming with Z, and for every $\tilde{y}$ we are able to reject, we reduce the size of the prediction region, thus increasing the *efficiency* of the conformal predictor.

Since $(x_j, \tilde{y})$ is included in the training data for the underlying model, the model needs to be retrained for each tentative label $\tilde{y}$; as such, this form of CP suffers from a rather poor computational complexity. ICP, as described in the next subsection, solves this problem by dividing the data set into two disjunct subsets: a proper training set and a calibration set.

2.2 Inductive Conformal Prediction

An ICP needs to be trained only once, using the following scheme:

1. Divide the training set $Z = \{(x_1, y_1), ...(x_l, y_l)\}$ into two disjoint subsets:
   - a proper training set $Z' = \{(x_1, y_1), ..., (x_m, y_m)\}$ and
   - a calibration set $Z'' = \{(x_{m+1}, y_{m+1}), ..., (x_{m+q}, y_{m+q})\}$
2. Train the underlying model h_Z using Z' as training data.
3. For each calibration instance $(x_i, y_i) \in Z''$:
   - let h_Z predict the output value for x_i so that $\hat{y}_i = h_Z(x_i)$ and
   - calculate the nonconformity score a_i using the nonconformity function.

For a novel (test) instance we simply supply the input pattern x_j to the underlying model and calculate $\alpha_j^{\tilde{y}}$ using our nonconformity function. The p-value

of each tentative label $\tilde{y}$ is then calculated by comparing $\alpha_j^{\tilde{y}}$ to the nonconformity scores of the calibration set:

$$p(\tilde{y}) = \frac{\#\left\{z_i \in Z'' \mid a_i \geq a_j^{\tilde{y}}\right\} + 1}{q+1}, \tag{2}$$

where q is the size of the calibration set. If $p(\tilde{y}) < \delta$, it is at most $\delta\%$ likely that $\tilde{y}$ is the true output of x_j, and $\tilde{y}$ is thus excluded from the prediction region.

2.3 Inductive Conformal Regression

In regression it is not possible to consider every possible output value $\tilde{y}$, so we cannot explicity calculate the p-value for each and every $\tilde{y}$. Instead a conformal regressor must effectively work in reverse. First, the size of the $(1-\delta)$-percentile nonconformity score, $\alpha_{s(\delta)}$, is determined; second, the nonconformity function is used to calculate the magnitude of error that would result in x_j being given a nonconformity score at most $\alpha_{s(\delta)}$; i.e., the conformal regressor determines the largest error that would be committed by the underlying model when predicting y_j with probability $1-\delta$. To perform conformal regression, we first define a nonconformity function, typically using the absolute error, see e.g., [5–8]:

$$\alpha_i = |y_i - \hat{y}_i| \quad . \tag{3}$$

Then, given a significance level δ and a set of calibration scores S, the goal is to find $\alpha_j^{\hat{y}}$ such that $P(\alpha_j^{\hat{y}} > \alpha_i \in S) \leq \delta$; i.e., the largest nonconformity score — and, due to the definition of (3), also the largest absolute error — with probability $1-\delta$. To do this, we simply sort the calibration scores in a descending order, and define the prediction interval as

$$\hat{Y}_j^{\delta} = (\hat{y}_j - \alpha_{s(\delta)}, \hat{y}_j + \alpha_{s(\delta)}), \tag{4}$$

where $s(\delta) = \lfloor \delta(q+1) \rfloor$, i.e., the index of the $(1-\delta)$-percentile in the sorted list of nonconformity scores. Since the underlying model's error is at most $\alpha_{s(\delta)}$ with probability $1-\delta$, the resulting interval covers the true target y_j with probability $1-\delta$. Note that when using (3) and (4) the conformal regressor will, for any specific significance level δ, always produce prediction intervals of the same size for every x_j; i.e., it does not consider the difficulty of a certain instance x_j. Papadopoulos et al. [5] suggest that prediction intervals can be *normalized* using some estimation of the difficulty of each instance, e.g., by using a separate model for estimating the error of the underlying predictor. In this paper, we will not consider normalized nonconformity scores, but leave them for future work.

3 Method

AECPR will, as described in the Background, always produce 'symmetrical' prediction intervals where the underlying model's prediction is the center of

the interval, and the distance from the interval's center to either boundary is equal to $\alpha_{s(\delta)}$, i.e., the absolute error from the calibration set associated with the significance level $1 - \delta$. If the errors of the underlying model are symmetrically distributed — i.e., the underlying model is equally likely to underestimate and overestimate the true output — AECPR will always yield optimal interval boundaries in the sense that neither boundary is overly optimistic nor overly pessimistic in relation to the error distribution of the model (as estimated on the calibration set). However, when the error distribution of the underlying model is skewed, it is possible for one of the boundaries to become overly optimistic, while the other becomes overly pessimistic, simply because the errors committed in one direction will influence the nonconformity scores, and consequently the prediction intervals, in both directions. Figure 1 shows an example of a skewed error distribution, where the AECPR nonconformity scores of the underlying model (a neural network) fail to capture the model's tendency to produce smaller negative errors (overestimation) and larger positive errors (underestimation).

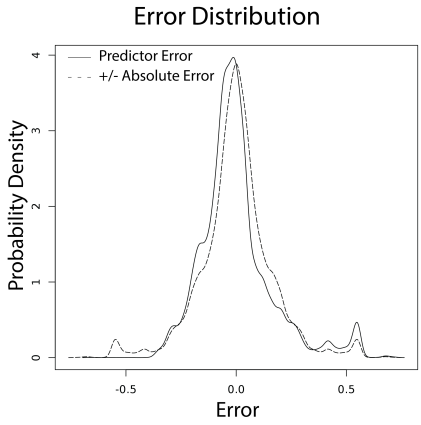

Fig. 1. Error distribution of an ANN on the Boston Housing data set. The signed error (solid line) approximates the skewed distribution of the ANN's error rate, while the absolute error (dashed line) assumes a symmetrical distribution when mirrored onto the negative range as per equations (3) and (4).

AECPR is proven valid [5], and thus there is no need to suspect that the mismatch between absolute error nonconformity scores and skewed error distributions would lead to invalid prediction intervals; however, it is necessary to be very specific about what validity means in the context of AECPR. Given that AECPR operates on the magnitude of the underlying model's error, the validity applies to the magnitude of errors and nothing else; i.e., AECPR guarantees that the absolute error of the underlying model is no larger than $\alpha_{s(\delta)}$ with probability $1 - \delta$. However, without considering the underlying model's tendency to commit

positive or negative errors, AECPR cannot provide information regarding how δ is distributed above and below the prediction interval.

Without this information, it is not possible for a user of AECPR to distinctly assess the validity of the prediction boundaries. To illustrate, consider a 95%-confidence prediction interval on the form $(-1, 1)$. If asked to assess the likelihood of the true value y_j being greater than 1, one is easily tempted to assume a probability of 2.5%, since intuitively, the probability of y_j being greater than the interval's upper boundary should be about the same as the probability of y_j being less than the interval's lower boundary. The true answer however, is that AECPR can only guarantee that there is at most a 5% probability of y_j being less than -1, and at most a 5% probability of y_j being greater than 1, since we have no information of the probabilities of the true value being higher or lower than the prediction interval's upper and lower boundaries respectively. In the following subsections, we expand on this argument, and propose a straightforward method for producing prediction intervals that possess a stronger guarantee of validity for the individual boundaries than for the full interval, while maintaining efficiency.

3.1 Validity of Interval Boundaries

If $\hat{Y} = (\hat{Y}_{low}, \hat{Y}_{high})$ is a valid prediction region at $\delta = d$, the one-tailed predictions $(-\infty, \hat{Y}_{high})$ and $(\hat{Y}_{low}, +\infty)$ must also be valid prediction regions at $\delta = d$, as they both cover at least the same error probability mass covered by $\hat{Y}$. However, without any knowledge of how the error probability d is distributed above and below $\hat{Y}$ it is not possible to assume that one specific one-tailed prediction is valid at any $\delta < d$. Hence, we can, in fact, only be confident in the one-tailed predictions with probability $1-d$, i.e., if $\hat{Y} = (\hat{Y}_{low}, \hat{Y}_{high})$ is valid at $\delta = d$, then $(-\infty, \hat{Y}_{high})$ and $(\hat{Y}_{low}, +\infty)$ are valid at $\delta = d$, but may be invalid at $\delta < d$.

On the other hand, if $(-\infty, \hat{Y}_{high})$ and $(\hat{Y}_{low}, +\infty)$ are both known to be valid at $\delta = \frac{d}{2}$, we know by definition that either of these one-tailed predictions will be incorrect with probability at most $\frac{d}{2}$. Thus, if the two one-tailed predictions are combined into a prediction interval $\hat{Y}_c = (\hat{Y}_{low}, \hat{Y}_{high})$, we can guarantee that $\hat{Y}_c$ will be wrong in a specific direction with probability at most $\frac{d}{2}$, and in total with probability at most d. Now, we are able to express not only a confidence $1 - d$ for the interval, but also a greater confidence $1 - \frac{d}{2}$ for the individual boundaries. Hence, if $(-\infty, \hat{Y}_{high})$ and $(\hat{Y}_{low}, +\infty)$ are valid at $\delta = \frac{d}{2}$, then $\hat{Y}_c = (\hat{Y}_{low}, \hat{Y}_{high})$ must be valid at $\delta = d$. This follows from the fact that when two one-tailed predictions are combined into a two-tailed interval, the probability of the resulting interval being wrong is the sum of the probabilities of the boundaries being wrong.

Using AECPR, prediction intervals with boundaries guaranteed at $\frac{\delta}{2}$ can be constructed simply by creating a prediction region $\hat{Y}_j^{0.5\delta}$, and outputting it as a combined interval $\hat{Y}_{c,j}^{\delta}$. This is of course rather impractical — as $|\hat{Y}_j^{0.5\delta}| \geq |\hat{Y}_j^{\delta}|$, we're not only 'increasing' the guarantee of validity for the individual boundaries in $\hat{Y}_{c,j}^{\delta}$ compared to $\hat{Y}_j^{\delta}$, we're also effectively guaranteeing that our prediction

interval is unnecessarily large! We would much rather output a combined interval such that $|\hat{Y}^{\delta}_{c,j}| \approx |\hat{Y}^{\delta}_{j}|$. To accomplish this, we are required to reduce the size of the predicted boundaries $\hat{Y}_{low}$ and $\hat{Y}_{high}$. We note the following: if the underlying model's predictions tend to underestimate the true output values, we are only required to adjust the upper boundary of the prediction intervals for them to remain valid; similarly, if the underlying model's predictions tend to overestimate, we are only required to adjust the lower boundary. Hence when predicting $\hat{Y}_{low}$ we only need to consider the negative errors made by the underlying predictor; and, when predicting $\hat{Y}_{high}$, we only need to consider the positive errors made by the underlying predictor. That is, we can construct the interval boundaries by considering only about half of the errors committed by the underlying model, thus reducing the expected size of each boundary while maintaining validity in the one-tailed predictions. This follows directly from the semantics of validity we are applying to the lower and upper boundary predictions — in both cases, we are only interested in guaranteeing validity in a single direction, i.e., for the upper boundary, we want to guarantee only that the probability of y_j being greater than $\hat{Y}_{high}$ is at most $1 - \frac{\delta}{2}$, and for the lower boundary we want to guarantee that the probability of y_j being less than $\hat{Y}_{low}$ is at most $1 - \frac{\delta}{2}$. Thus, if we are interested in predicting the upper boundary of y_j, we can define the nonconformity measure as $\alpha_i = y_i - \hat{y}_i$, i.e., a nonconformity function that returns larger nonconformity scores with larger positive errors, and define the prediction interval as $(-\infty, \hat{y}_j + \alpha_{s(\frac{\delta}{2})})$. In this case, we guarantee that the true value $y_j \leq \hat{y}_j + \alpha_{s(\frac{\delta}{2})}$ with probability $1 - \frac{\delta}{2}$. Conversely, we can define the nonconformity score such that it increases with larger negative errors, i.e., $\alpha_i = \hat{y}_i - y_i$, output the prediction interval $(\hat{y}_j - \alpha_{s(\frac{\delta}{2})}, +\infty)$, and guarantee that $y_j \geq \hat{y}_j - \alpha_{s(\frac{\delta}{2})}$ with probability $1 - \frac{\delta}{2}$.

From this point, we are able to do one of two things: we can output either $(-\infty, \hat{Y}_{high})$ or $(\hat{Y}_{low}, +\infty)$ as the prediction with confidence $1 - \frac{\delta}{2}$; or, we can combine the two one-tailed predictions, and guarantee the boundaries at $1 - \frac{\delta}{2}$, and the interval at $1 - \delta$.

3.2 Signed Error CPR

To approximate the (potentially skewed) error distribution of the underlying model, we propose a nonconformity measure based on the signed error of the model; i.e., we define the nonconformity score as

$$\alpha_i = y_i - \hat{y}_i \ . \tag{5}$$

Just as in AECPR, we sort α in a descending order — note though, that while the sorted α in AECPR contains the absolute errors from largest to smallest, the sorted α in SECPR ranges from maximum positive error to maximum negative error. The prediction interval for a novel instance x_j is then formulated as

$$\hat{Y}^{\delta}_{j} = (\hat{y}_j + \alpha_{low(\frac{\delta}{2})}, \hat{y}_j + \alpha_{high(\frac{\delta}{2})}), \tag{6}$$

where $high(\frac{\delta}{2}) = \lfloor \frac{\delta}{2}(q+1) \rfloor$ and $low(\frac{\delta}{2}) = \lfloor (1-\frac{\delta}{2})(q+1) \rfloor + 1$. In effect, SECPR performs two simultaneous conformal predictions — one for each boundary of the interval. The boundaries are predicted with confidence $1 - \frac{\delta}{2}$ and, when combined, form a prediction interval with confidence $1 - \delta$.

3.3 Evaluation

The two methods (AECPR and SECPR) were evaluated on 33 publicly available data sets from the UCI [15] and Delve [16] repositories. Before experimentation all output values were scaled to $[0, 1]$, only to enhance interpretability in efficiency comparisons — with the outputs scaled to $[0, 1]$, the size of a prediction interval expresses the fraction of possible outputs covered by the interval. For each data set, 100 random sub-sampling tests were performed; in each iteration, a randomized subsample (20%) of the data set was used as the test set, and remaining instances were used for training and calibration. The calibration set size was defined as a function of the training set size: $|Z''| = 100 \left\lfloor \frac{|Z|}{100} \right\rfloor \times 0.1 + 199$.

Standard multilayer perceptron neural networks with $\left\lceil \sqrt{k} \right\rceil + 1$ hidden nodes were used as underlying models for the conformal predictors, where k is the number of input features of each data set. In each iteration, both AECPR and SECPR predictions were calculated from the same model and calibration instances.

4 Results

Given any $\delta \in (0, 1)$, an ICR should produce valid prediction intervals — in reality however, predictions with low confidence are rarely of interest. Thus, we choose to show the empirical validity and evaluate the efficiency of AECPR and SECPR at three commonly used confidence levels: 99%, 95% and 90%.

4.1 Validity of Intervals

As illustrated in Table 1, both methods produce prediction intervals that cover the true targets of the test set at or very close to the predefined significance levels; thus, in terms of interval coverage both methods are, as expected, valid.

4.2 Validity of Boundaries

Here, we take a closer look at the coverage of the lower and upper boundaries of the intervals produced by AECPR and SECPR. Specifically, we expect AECPR and SECPR to have a clear difference in coverage of their one-tailed predictions $(\hat{Y}_{low}, +\infty)$ and $(-\infty, \hat{Y}_{high})$. We expect SECPR's boundaries to be valid at $1 - \frac{\delta}{2}$, and AECPR boundary coverage to vary between $1 - \frac{\delta}{2}$ and $1 - \delta$.

Table 2 reveals that AECPR's interval boundaries show only a small deviance from $1 - \frac{\delta}{2}$-validity on average (across all data sets). However, more often than not, one of the boundaries is overly optimistic and the other overly pessimistic to

Table 1. Mean coverage (portion of predictions that coincide with the true output of the test instances) for AECPR and SECPR at $\delta \in \{0.01, 0.05, 0.10\}$ on the 33 sets.

	99%		95%		90%	
	Abs	Sign	Abs	Sign	Abs	Sign
abalone	.990	.990	.950	.949	.902	.902
anacalt	1.00	1.00	.943	.997	.920	.944
bank8fh	.989	.989	.950	.949	.900	.901
bank8fm	.990	.991	.950	.951	.902	.903
bank8nh	.989	.990	.951	.949	.901	.901
bank8nm	.990	.991	.950	.951	.900	.901
boston	.996	.992	.948	.958	.897	.897
comp	.992	.993	.952	.962	.904	.923
concrete	.989	.993	.947	.954	.898	.904
cooling	.990	.990	.949	.949	.898	.903
deltaA	.991	.991	.950	.951	.901	.902
deltaE	.990	.991	.951	.951	.901	.902
friedm	.990	.994	.948	.951	.902	.906
heating	.990	.991	.948	.950	.895	.897
istanbul	.991	.991	.952	.955	.903	.903
kin8fh	.990	.990	.952	.951	.902	.902
kin8fm	.990	.990	.951	.951	.901	.903
kin8nh	.990	.990	.950	.950	.901	.902
kin8nm	.990	.990	.950	.950	.900	.901
laser	.990	.991	.949	.948	.898	.901
mg	.991	.994	.951	.954	.905	.911
mortage	.990	.994	.950	.952	.906	.907
plastic	.999	1.00	.997	.997	.994	.993
puma8fh	.990	.990	.949	.950	.900	.902
puma8fm	.990	.990	.950	.950	.901	.902
puma8nh	.990	.990	.951	.951	.902	.903
puma8nm	.990	.990	.950	.950	.901	.902
quakes	.992	.995	.955	.970	.908	.934
stock	.990	.990	.948	.948	.899	.896
treasury	.991	.993	.952	.952	.903	.910
wineR	.991	.994	.956	.957	.910	.913
wineW	.993	.993	.954	.963	.911	.913
wizmir	.991	.994	.949	.954	.904	.908
mean	.991	.992	.952	.955	.905	.909
min	.989	.989	.943	.948	.895	.896

compensate, and the reason the two interval boundaries appear valid on average is the simple fact that they are alternatingly optimistic and pessimistic. We also note that, in the worst cases (e.g. the treasury data set), one of the boundaries is valid only at $1 - \delta$. In contrast, SECPR interval boundaries show coverage at or very near $1 - \frac{\delta}{2}$-validity in both the average and worst cases.

To further illustrate, we let the D_p^δ be the average coverage of the boundary (low or high) that has the highest average coverage for data set D (the most pessimistic boundary), and we let D_o^δ be the average coverage of the boundary with the lowest average coverage for D (the most optimistic boundary). E.g., for abalone, AECPR has $D_o^{99} = 0.991$ and $D_p^{99} = 0.999$; and SECPR has $D_o^{99} = 0.995$ and $D_p^{99} = 0.995$. We then calculate the mean coverage $\bar{D}_o^\delta$ and $\bar{D}_p^\delta$ for each δ, and for both AECPR and SECPR (Table 3).

In Table 3 we can clearly see that, as we argued in the Method, AECPR intervals show a tendency towards having an overly pessimistic boundary and an overly optimistic boundary, in such a way that the error probability δ above and below the interval is unevenly distributed. Furthermore, we are not given any information regarding the distribution of δ, and thus as noted in the Method and supported by the empirical evidence, AECPR can only guarantee the validity of its interval boundaries at $1 - \delta$.

SECPR intervals can guarantee the interval boundaries at $1 - \frac{\delta}{2}$; so, this statement is supported by the empirical evidence. We also note that SECPR tends to produce balanced intervals — i.e., intervals where δ is evenly distributed above and below the prediction intervals. More importantly, even in the cases where the intervals are not perfectly balanced, we have already defined the boundaries to be valid at $1 - \frac{\delta}{2}$.

Table 2. Mean low/high coverage for AECPR and SECPR at $\delta \in \{0.01, 0.05, 0.10\}$.

	99%				95%				90%			
	Abs		Sign		Abs		Sign		Abs		Sign	
	low	high	low	high	low	high	low	high	low	high	low	high
abalone	.999	.991	.995	.995	.989	.961	.974	.976	.968	.934	.949	.953
anacalt	1.00	1.00	1.00	1.00	.943	1.00	.997	1.00	.920	1.00	.944	1.00
bank8fh	.999	.990	.995	.995	.986	.964	.975	.974	.967	.933	.950	.951
bank8fm	.997	.993	.995	.995	.981	.970	.975	.976	.958	.944	.952	.951
bank8nh	1.00	.990	.995	.995	.995	.956	.974	.975	.981	.920	.950	.951
bank8nm	.998	.992	.995	.995	.983	.967	.976	.974	.959	.941	.951	.950
boston	.999	.997	.994	.998	.993	.955	.975	.983	.967	.930	.950	.947
comp	.995	.997	.997	.996	.960	.992	.979	.983	.919	.985	.953	.970
concrete	1.00	.989	.997	.996	.994	.953	.977	.976	.972	.927	.954	.950
cooling	1.00	.990	.996	.994	.997	.952	.976	.973	.984	.914	.954	.949
deltaA	.995	.996	.996	.996	.971	.978	.976	.975	.946	.956	.952	.949
deltaE	.994	.996	.996	.995	.975	.976	.976	.976	.953	.949	.951	.951
friedm	.995	.995	.996	.997	.978	.970	.974	.976	.956	.946	.951	.955
heating	1.00	.990	.996	.995	.998	.951	.976	.974	.982	.913	.949	.947
istanbul	.995	.996	.996	.994	.976	.976	.977	.978	.947	.955	.952	.951
kin8fh	.993	.997	.995	.995	.972	.980	.976	.975	.948	.954	.952	.951
kin8fm	.994	.996	.995	.995	.971	.979	.975	.976	.946	.955	.951	.952
kin8nh	.995	.995	.995	.994	.972	.979	.975	.975	.946	.954	.950	.951
kin8nm	.996	.994	.995	.995	.974	.975	.975	.975	.949	.952	.951	.950
laser	.995	.995	.994	.997	.973	.976	.973	.975	.947	.952	.950	.950
mg	.996	.995	.996	.998	.978	.972	.977	.977	.952	.953	.955	.956
mortage	1.00	.990	.998	.997	.998	.952	.976	.976	.995	.911	.955	.953
plastic	1.00	.999	1.00	1.00	1.00	.997	.998	.998	.998	.995	.996	.997
puma8fh	.995	.995	.995	.995	.972	.977	.974	.975	.946	.955	.950	.951
puma8fm	.994	.996	.995	.995	.973	.977	.975	.975	.950	.951	.952	.950
puma8nh	.993	.997	.995	.995	.972	.980	.976	.975	.947	.955	.951	.952
puma8nm	.994	.995	.995	.995	.974	.976	.976	.975	.949	.952	.952	.950
quakes	1.00	.992	.998	.997	1.00	.956	.991	.980	.998	.910	.979	.955
stock	.996	.994	.995	.994	.978	.970	.974	.974	.957	.942	.948	.948
treasury	1.00	.991	.996	.997	.999	.953	.973	.978	.993	.911	.956	.954
wineR	.994	.997	.997	.997	.976	.980	.977	.980	.952	.958	.955	.959
wineW	.996	.996	.996	.997	.984	.970	.978	.985	.961	.949	.955	.957
wizmir	.994	.997	.997	.997	.974	.976	.976	.978	.954	.950	.954	.954
mean	.997	.994	.996	.996	.981	.971	.977	.978	.960	.946	.954	.955
min	.989		.994		.943		.973		.910		.944	

4.3 Efficiency

We can choose to compare the interval sizes of AECPR and SECPR (Table 4) based on two different criteria: either we compare intervals that share the same guarantee of validity for the full interval; or, we compare intervals that share the same guarantee of validity for the individual boundaries. That is, we either compare the 99% SECPR intervals to the 99% AECPR intervals, and so on, and remember that SECPR provides a stronger guarantee of validity for the boundaries than does AECPR; or, we compare the 90% SECPR intervals to the 95% AECPR intervals, and so on, and remember that the two methods in this case provide the same guarantee for the one-tailed prediction boundaries.

First, we consider predictions that share the same guarantee of validity for the full interval. Here, we note that on average SECPR produces tighter intervals than AECPR at the 99% confidence level, due to SECPR taking into account the sign of the underlying model's error. If the underlying model commits large errors in only one direction, only one of the interval boundaries predicted by SECPR is affected, while both boundaries are affected in AECPR. Thus, for data sets

Table 3. Mean optimistic and pessimistic boundary coverage of AECPR and SECPR.

	$\bar{D}_o^{99}$	$\bar{D}_p^{99}$	$\bar{D}_o^{95}$	$\bar{D}_p^{95}$	$\bar{D}_o^{90}$	$\bar{D}_p^{90}$
AECPR	.994	.997	.967	.985	.939	.966
SECPR	.996	.996	.977	.979	.952	.957

Table 4. Mean-median interval sizes for AECPR and SECPR at $\delta \in \{0.01, 0.05, 0.10\}$.

	99%		95%		90%	
	Abs	Sign	Abs	Sign	Abs	Sign
abalone	.528	.486	.324	.322	.238	.248
anacalt	1.77	1.00	1.18	.985	.728	.707
bank8fh	.543	.540	.379	.371	.295	.297
bank8fm	.266	.255	.191	.186	.152	.153
bank8nh	.793	.720	.452	.440	.327	.338
bank8nm	.390	.374	.224	.221	.158	.162
boston	1.11	.969	.712	.707	.488	.554
comp	.790	.685	.419	.405	.286	.303
concrete	1.00	.933	.753	.780	.646	.673
cooling	1.07	.910	.787	.779	.655	.665
deltaA	.269	.283	.162	.164	.126	.127
deltaE	.317	.333	.214	.215	.175	.175
friedm	.289	.317	.210	.213	.175	.179
heating	1.04	.952	.904	.840	.776	.752
istanbul	.503	.542	.338	.344	.274	.274
kin8fh	.387	.383	.286	.287	.238	.239
kin8fm	.163	.162	.119	.120	.099	.100
kin8nh	.628	.631	.482	.482	.405	.407
kin8nm	.549	.551	.395	.396	.321	.323
laser	.625	.605	.291	.281	.208	.215
mg	.787	.837	.513	.529	.401	.411
mortage	1.15	.915	.895	.791	.663	.721
plastic	.993	.976	.983	.973	.975	.968
puma8fh	.765	.768	.570	.571	.470	.472
puma8fm	.359	.361	.266	.266	.223	.224
puma8nh	.762	.756	.552	.554	.448	.451
puma8nm	.369	.365	.253	.255	.205	.207
quakes	1.08	.871	.709	.639	.492	.527
stock	.748	.733	.597	.604	.530	.530
treasury	1.13	.889	.807	.763	.556	.649
wineR	.841	1.10	.556	.567	.448	.449
wineW	.741	.756	.546	.552	.403	.394
wizmir	.507	.558	.418	.424	.383	.384
mean	.705	.652	.500	.486	.393	.402
std dev	.349	.260	.266	.242	.212	.216

where the underlying model commits *outlier errors* — atypical errors that are of a much larger magnitude than typical errors — in only one direction, SECPR will produce tighter intervals than AECPR. It must be noted that while this applies to all significance levels, most of the outlier errors will, for lower significance levels, be excluded from the prediction intervals anyway. Consequently, at the 95% confidence level we observe a similar pattern, but the effect is much less pronounced. At 90%, the effect is all but gone, and SECPR is instead slightly less effective than AECPR on almost all data sets. A Wilcoxon signed-ranks test at $\alpha = 0.05$ shows that, in terms of efficiency, SECPR is significantly better than AECPR for 99%-confidence predictions, while AECPR is significantly better than SECPR for 90%-confidence predictions. Thus, at 90% confidence or lower, we can expect that the strengthened guarantee of validity provided by SECPR is accompanied with a small but significant decrease in efficiency.

Second, we consider one-tailed predictions, specifically at the 95%-confidence level (i.e., AECPR at 95%, SECPR at 90%); here, we can clearly see that SECPR produces tighter intervals than AECPR for all data sets. Hence, simply by ensuring that the boundaries are affected only by the relevant errors of the underlying model, we can significantly increase the efficiency of one-tailed predictions.

5 Conclusions

In this paper, we have shown that conformal regressors based on the absolute error of the underlying model produce unbalanced prediction intervals — intervals with no guarantee for the distribution of error above and below the intervals

— and that this unbalance leads to prediction intervals with weak guarantees of validity for the individual interval boundaries. To address this issue, we have proposed a straightforward approach for producing prediction intervals based on the signed error of the underlying model (SECPR) that can provide a stronger guarantee of validity for the interval boundaries. Also, we have shown that SECPR is less sensitive to outlier errors than AECPR, resulting in more efficient prediction intervals at the highest confidence levels. Finally, we show that, when expected to provide the same guarantees of boundary validity as AECPR, SECPR produces much more efficient prediction intervals than corresponding AECPR models.

References

1. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer Verlag, DE (2006)
2. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (1998) 148–155
3. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99). Volume 2. (1999) 722–726
4. Shafer, G., Vovk, V.: A tutorial on conformal prediction. The Journal of Machine Learning Research **9** (2008) 371–421
5. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Machine Learning: ECML 2002. Springer (2002) 345–356
6. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. Journal of Artificial Intelligence Research **40**(1) (2011) 815–840
7. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. Tools in artificial intelligence **18**(315-330) (2008) 2
8. Papadopoulos, H., Haralambous, H.: Reliable prediction intervals with regression neural networks. Neural Networks **24**(8) (2011) 842–851
9. Lambrou, A., Papadopoulos, H., Gammerman, A.: Reliable confidence measures for medical diagnosis with evolutionary algorithms. Information Technology in Biomedicine, IEEE Transactions on **15**(1) (2011) 93–99
10. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. Tools in Artificial Intelligence **18** (2008) 315–330
11. Papadopoulos, H., Papatheocharous, E., Andreou, A.S.: Reliable confidence intervals for software effort estimation. In: AIAI Workshops, Citeseer (2009) 211–220
12. Devetyarov, D., Nouretdinov, I., Burford, B., Camuzeaux, S., Gentry-Maharaj, A., Tiss, A., Smith, C., Luo, Z., Chervonenkis, A., Hallett, R., et al.: Conformal predictors in early diagnostics of ovarian and breast cancers. Progress in Artificial Intelligence **1**(3) (2012) 245–257
13. Papadopoulos, H., Gammerman, A., Vovk, V.: Reliable diagnosis of acute abdominal pain with conformal prediction. Engineering Intelligent Systems **17**(2) (2009) 127
14. Lambrou, A., Papadopoulos, H., Kyriacou, E., Pattichis, C.S., Pattichis, M.S., Gammerman, A., Nicolaides, A.: Assessment of stroke risk based on morphological

ultrasound image analysis with conformal prediction. In: Artificial Intelligence Applications and Innovations. Springer (2010) 146–153
15. Bache, K., Lichman, M.: UCI machine learning repository. URL http://archive.ics.uci.edu/ml (2013)
16. Rasmussen, C.E., Neal, R.M., Hinton, G., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R., Tibshirani, R.: Delve data for evaluating learning in valid experiments. URL http://www. cs. toronto. edu/ delve (1996)