Mondrian Conformal Regressors

Henrik Boström BOSTROMH@KTH.SE
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden

Ulf Johansson ULF.JOHANSSON@JU.SE
Dept. of Computer Science and Informatics, Jönköping University, Sweden

Editor: Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov and Giovanni Cherubin

Abstract

Standard (non-normalized) conformal regressors produce intervals that are of identical size and hence non-informative in the sense that they provide no information about the uncertainty at the instance level. A common approach to handle this limitation is to normalize the produced interval using a difficulty estimate, which results in larger intervals for instances judged to be more difficult and smaller intervals for instances judged to be easier. A problem with this approach is identified; when the difficulty estimation function provides little or no information about the true error at the instance level, one would expect the predicted intervals to be more similar in size compared to when using a more accurate difficulty estimation function. However, experiments on both synthetic and real-world datasets show the opposite. Moreover, the intervals produced by normalized conformal regressors may be several times larger than the largest previously observed prediction error, which clearly is counter-intuitive. To alleviate these problems, we propose *Mondrian conformal regressors*, which partition the calibration instances into a number of categories, before generating one prediction interval for each category, using a standard conformal regressor. Here, binning of the difficulty estimates is employed for the categorization. In contrast to normalized conformal regressors, Mondrian conformal regressors can never produce intervals that are larger than twice the largest observed error. The experiments verify that the resulting regressors are valid and as efficient as when using normalization, while being significantly more efficient than the standard variant. Most importantly, the experiments show that Mondrian conformal regressors, in contrast to normalized conformal regressors, have the desired property that the variance of the size of the predicted intervals correlates positively with the accuracy of the function that is used to estimate difficulty.

Keywords: Conformal regression, normalization, Mondrian conformal predictors.

1. Introduction

Conformal predictive regression is an established technique for producing valid regressors using a number of different machine learning techniques, e.g., ridge regression (Papadopoulos et al., 2002), neural networks (Papadopoulos and Haralambous, 2010), kNN (Papadopoulos et al., 2011), and random forests (Johansson et al., 2014a; Boström et al., 2017). In addition to being valid, conformal regressors are often regarded as more informative than

their underlying models since they output prediction intervals instead of point predictions. Specifically, the size of the prediction interval is a vital indication of the uncertainty in that prediction. In the standard inductive setting, however, the size of the prediction intervals is identical for all test instances, i.e., there is no information available about the uncertainty of the predictions on the instance level. To obtain more specific predictions, a procedure called *normalization* has been used extensively. In normalized conformal regressors, the non-conformity function contains a component measuring the difficulty of individual test instances, resulting in individualized intervals for each prediction. Previous studies, see e.g., (Papadopoulos and Haralambous, 2011; Johansson et al., 2014a) show that normalization also results in tighter intervals on average, at least for the significance levels typically used. While normalization has these nice properties, few if any studies have investigated the results of utilizing it in more detail.

In this work, we will highlight two problems with normalized conformal regressors, which may lead to that the resulting intervals are of questionable utility; i) the produced intervals do not accurately reflect the degree to which the difficulty estimate actually provides information about the true error, and ii) the resulting intervals may be unreasonably large or small. We discuss these problems in more detail below.

When we have little or no information about the uncertainty at the instance level, i.e., when the difficulty estimate has a low correlation with the true error, one may argue that the size of the predicted intervals should be similar for all instances, as we have no or little support for claiming that some predictions are more uncertain than others. Conversely, the higher the correlation between the difficulty and the prediction error, the larger should the variance of the prediction intervals be, if the interval sizes should reflect the uncertainty. We will show that on the contrary, the more random (and hence less informative) the difficulty function is, the more varies the size of the intervals produced by normalized conformal regressors. This will be illustrated with results from experiments with both synthetic and real-world data.

Moreover, as will be shown in Section 2, the intervals produced by normalized conformal regressors may, at least in theory, be several times larger (or smaller) than what have been previously observed, which means that we in such cases directly can conclude that they with high probability are either too conservative or even invalid. We will see also from experiments with both synthetic and real-world datasets, using state-of-the-art conformal regressors, that this problem may be manifested in reality and not only in theory.

In order to overcome the above two problems, an alternative approach to producing prediction intervals is proposed, which falls in between of the standard approach, which produces fixed-size, non-informative intervals (independently of whether or not we can accurately estimate the difficulty), and normalized conformal regressors, which may generate unique intervals for each prediction. The alternative approach borrows the idea of Mondrian conformal prediction (Vovk et al., 2005), which to the best of our knowledge has been applied to conformal classification only, to form categories based on the difficulty estimate. In this study, we consider the straightforward formation of categories by binning the difficulty estimates. A prediction interval is then formed for each category using a standard (non-normalized) conformal regressor.

We will compare the novel type of model, which we refer to as *Mondrian Conformal Regressors*, to both standard (non-normalized) and normalized conformal regressors, and

demonstrate that the novel model addresses the above problems successfully, without significantly sacrificing efficiency.

In the next section, we briefly describe conformal regressors and provide an extreme-case analysis of the size of the produced intervals. In Section 3, we introduce the alternative Mondrian-based approach; Mondrian Conformal Regressors. In Section 4, we present results from comparing the novel approach to normalized (and standard) conformal regressors on both synthetic and real-world datasets. Finally, in Section 5, we discuss the main findings and outline directions for future work.

2. Preliminaries

Conformal prediction was originally developed for the transductive case (Gammerman et al., 1998), requiring re-training of the underlying model for each new instance to be predicted, something which often is computationally infeasible. Inductive conformal prediction (ICP) was proposed as a computationally less costly approach (Papadopoulos et al., 2002), requiring only one underlying model to be generated, at the cost of having to set aside part of the training examples for calibration, which leaves less examples to use for model building.

An inductive conformal predictor relies on a nonconformity function A to assess the strangeness of a label y_i to assign to some object $\boldsymbol{x}_i$, with respect to some underlying model h. For regression problems, the absolute error is a common choice for defining nonconformity:

$$A\left(\boldsymbol{x}_i, y_i, h\right) = \left|y_i - h\left(\boldsymbol{x}_i\right)\right|, \tag{1}$$

A *standard (inductive) conformal regressor* is constructed as follows:

1. Divide the training data Z_{tr} into two disjoint subsets:
   the proper training set Z_t and the calibration set Z_c.
2. Train the underlying model h using Z_t.
3. Use Eq. 1 to measure the nonconformity of the examples in Z_c, producing a list of calibration scores $S = \alpha_1, ..., \alpha_q$ where $q = |Z_c|$ and S is sorted in descending order.

A valid prediction interval at the confidence level $1-\epsilon$ for a test instance $\boldsymbol{x}_{l+1}$ is obtained from a standard conformal regressor by:

1. Make a prediction $h(\boldsymbol{x}_{l+1})$.
2. Find the calibration score α_p where $p = \lfloor \epsilon(q+1) \rfloor$.
3. The prediction interval for $\boldsymbol{x}_{l+1}$ is

$$\hat{Y}^{\epsilon}_{l+1} = h\left(\boldsymbol{x}_{l+1}\right) \pm \alpha_p, \tag{2}$$

The probability that the underlying model h will make an absolute error larger than α_p is ϵ. Note that all predictions output by a standard conformal regressor for a certain confidence level are of the same size, namely $2\alpha_p$. In order to produce a more informative conformal regressor, *normalization* can be added. Here, the size of the prediction intervals

vary based on the estimated difficulty of the test examples, with easier instances having tighter intervals.

A *normalized (inductive) conformal regressor* modifies the standard conformal regressor by employing the following nonconformity function, given an object $\boldsymbol{x}_i$, label y_i and underlying model h:

$$A\left(\boldsymbol{x}_i, y_i, h\right) = \frac{\left|y_i - h\left(\boldsymbol{x}_i\right)\right|}{\sigma_i + \beta}, \tag{3}$$

where σ_i is a difficulty estimate of $\boldsymbol{x}_i$ and β is a parameter.

The prediction interval output by a normalized conformal regressor is

$$\hat{Y}^{\epsilon}_{l+1} = h\left(\boldsymbol{x}_{l+1}\right) \pm \alpha_p \left(\sigma_{l+1} + \beta\right) . \tag{4}$$

Normalization is supposed to provide two benefits: (i) the sizes of the prediction intervals give information on a per-instance basis and (ii) the prediction intervals should be tighter on average, i.e., the model is more efficient. However, we now show that the intervals produced by normalized conformal regressors may be unreasonably large or small. To see this, consider the largest possible interval that may be produced by a normalized conformal regressor:

$$2\frac{e_{max}}{\sigma_{min} + \beta}(\sigma_{max} + \beta) \tag{5}$$

where e_{max} is the largest observed absolute error, σ_{min} and σ_{max} are the smallest and largest observed difficulty estimates.

Assuming a normalized σ, i.e., $\sigma_{min} = 0$ and $\sigma_{max} = 1$, then the largest possible interval becomes:

$$2\frac{e_{max}}{\beta}(1 + \beta) = 2(\frac{e_{max}}{\beta} + e_{max}) = 2e_{max}(1 + \frac{1}{\beta}) \tag{6}$$

Similarly, the smallest possible interval that can be produced is:

$$2\frac{e_{min}}{\sigma_{max} + \beta}(\sigma_{min} + \beta) \tag{7}$$

Under the same assumptions as above:

$$2\frac{e_{min}}{1 + \beta}\beta = 2e_{min}\frac{\beta}{1 + \beta} \tag{8}$$

Hence, depending on β, the largest possible interval may be several times larger than the largest previously observed error (e_{max}), and the smallest possible interval may be many times smaller than the smallest previously observed interval (e_{min}).

3. Mondrian Conformal Regressors

In Mondrian conformal predictors (Vovk et al., 2005), the available calibration instances are somehow divided into different categories, and then a valid conformal predictor is built for each category. The most common Mondrian conformal predictor is probably the *class-conditional conformal predictor* (Shi et al., 2013), where the categories represent the possible class labels, thus providing guarantees for each label, i.e., the errors will be evenly distributed over the classes. The problem space can also be divided w.r.t. to the feature space, e.g., for tree models, a very natural division is to regard each leaf (path) as a separate category, resulting in that each such leaf is independently valid, see e.g., (Johansson et al., 2014b). Until now, however, Mondrian conformal prediction has, to the best of our knowledge, only been applied to classification and not to regression.

A *Mondrian (inductive) conformal regressor* is constructed as follows:

1. Divide the training data Z_{tr} into two disjoint subsets:
   the proper training set Z_t and the calibration set $Z_c = \{(\boldsymbol{x_1}, y_1), \ldots, (\boldsymbol{x_q}, y_q))\}$.
2. Train the underlying model h using Z_t.
3. Divide Z_c into k disjoint subsets $Z_{c1}, \ldots, Z_{ck}$, according to a Mondrian taxonomy κ with categories $\kappa_1, \ldots, \kappa_k$
4. Use Eq. 1 to measure the nonconformity of the examples in Z_{ci}, for each $i = 1, \ldots, k$, producing a list of calibration scores $S_i = \alpha_{i1}, ..., \alpha_{iq_i}$ where $q_i = |Z_{ci}|$ and S_i is sorted in descending order.

A valid prediction interval at the confidence level $1-\epsilon$ for a test instance $\boldsymbol{x}_{l+1}$ is obtained from a Mondrian conformal regressor by:

1. Make a prediction $h(\boldsymbol{x}_{l+1})$.
2. Find the category $\kappa_i \in \{\kappa_1, \ldots, \kappa_k\}$ for $\boldsymbol{x}_{l+1}$
3. Find the calibration score α_{ip} where $p = \lfloor \epsilon(q_i + 1) \rfloor$.
4. The prediction interval for $\boldsymbol{x}_{l+1}$ is

$$\hat{Y}^{\epsilon}_{l+1} = h\left(\boldsymbol{x}_{l+1}\right) \pm \alpha_{ip} \tag{9}$$

It should be noted that in contrast to Mondrian conformal predictors for finite label sets, i.e., Mondrian conformal classifiers, we will here not consider Mondrian taxonomies that take the true labels of the calibration instances ($\{y_1, \ldots, y_q\}$) into account when assigning the categories. The reason for this is that when forming the predictions (step 2 above), we do not have access to the true value (y_{l+1}) and it is not obvious how to test an infinite number of values for this, although we do not exclude this possibility in the future. Another constraint on the Mondrian taxonomy is that each category must contain a sufficient number of instances to allow for finding a calibration score (step 3 above), which is dependent on

the chosen confidence level. For example, a confidence level of 0.95 requires that at least 19 instances are included in each category.

There are many different possibilities for forming categories, e.g., based on properties of the objects, but in this work we will focus on forming categories based on the difficulty scores $\{\sigma_1, \ldots, \sigma_q\}$, as considered also by normalized conformal regressors. Again, this might be approached in many different ways, but we settle for a very straightforward approach; equal-sized binning of the difficulty estimates, which means that approximately the same number of instances will fall into each bin (category). The number of bins (categories) is hence a parameter of the approach, and it should be chosen in a way such that the above constraint is satisfied; the size of the calibration set and the degree of confidence put limits on the the number of bins that can be used.

If the Mondrian taxonomy maps only to one category (bin), the resulting Mondrian conformal regressor is identical to the standard conformal regressor, while by increasing the number of bins, the Mondrian conformal regressor will approach the normalized conformal regressor. However, as discussed earlier. some important differences to these existing two approaches still remain. We investigate this further in the next section.

4. Experiments

Here, we present experimental results from applying the various conformal regressors to both synthetic and real-world datasets.

4.1. Synthetic data

The actual (y_i) and the predicted (h_i) values for the synthetic calibration and test datasets are generated in the following way, where the number of instances was set to 10 000 for both the calibration and test sets:

$$y_i \sim \mathcal{N}(0, 1) \tag{10}$$

$$n_i \sim \mathcal{N}(0, 1) \tag{11}$$

$$u_i \sim \mathcal{U}(0, 1) \tag{12}$$

$$h_i = y_i + n_i \cdot u_i \tag{13}$$

In Fig. 1, the predicted vs. actual values for the calibration set are plotted (left) together with the cumulative distribution of the calibration intervals, i.e., twice the absolute residuals (right). The 95th percentile of these intervals is 2.51, which corresponds to the length of the intervals predicted by the standard approach for a confidence level of 0.95.

The above formulation allows us to investigate difficulty functions that are either uniformly or normally distributed, i.e., using either $|n_i|$ or u_i as an estimate of the difficulty. Moreover, we can experiment with various amounts of randomness; with some probability the difficulty estimate is replaced with a random value, sampled from either the normal or uniform distribution (depending on which difficulty function we use, and using the absolute value for the normal distribution). Three levels of randomness are investigated here; 0.0, 0.5 and 1.0 (corresponding to that none, about half and all difficulty estimates are replaced by random values). The choice of randomness clearly affects the correlation between the difficulty estimate and the true error; see Fig. 2(a) where the difficulty vs. residuals are

plotted for the three levels of randomness, using the normally distributed difficulty function. The corresponding plots for the uniform difficulty function are shown in Fig. 2(b). One may observe that complete randomness leads to non-correlated difficulty estimates, while when there is no replacement of difficulty estimates (randomness = 0), the correlation with the true error is quite high (but not perfect), where the use of a normal difficulty function gives a higher correlation compared to using the uniform difficulty function.

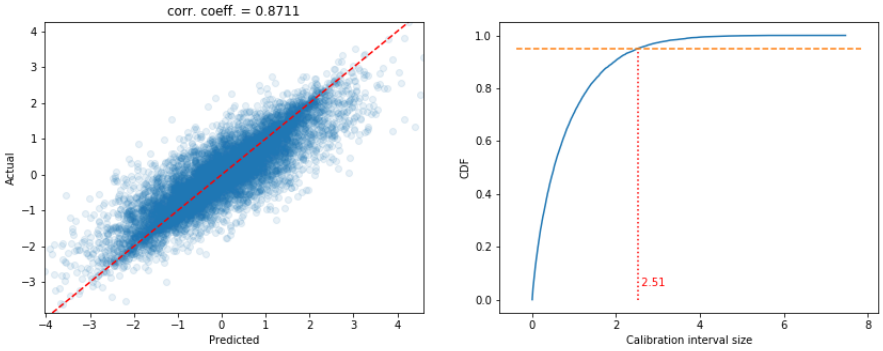
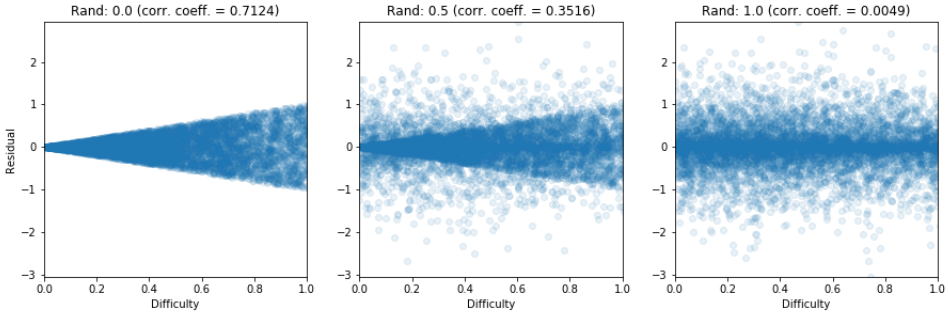
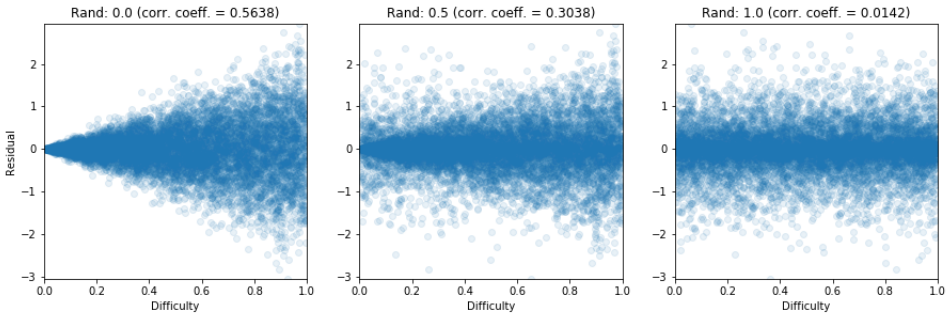


Figure 1: Predicted vs. actual values



(a) Residual vs. normal difficulty



(b) Residual vs. uniform difficulty

Figure 2: Residual vs. difficulty

We now present results for normalized conformal regression, when calibrating and testing on the above datasets, using three different values for β; 0.01, 0.1 and 1.0. In Fig. 3, the results for when using normally distributed difficulty values are shown. One may note that the median predicted interval size for normalized regression are sometimes higher than for the standard (non-normalized) approach, depending on the correlation between the difficulty function and the true error; for higher degrees of randomness, and in particular for lower values of β, the median efficiency of the normalized approach exceeds that of the standard approach.

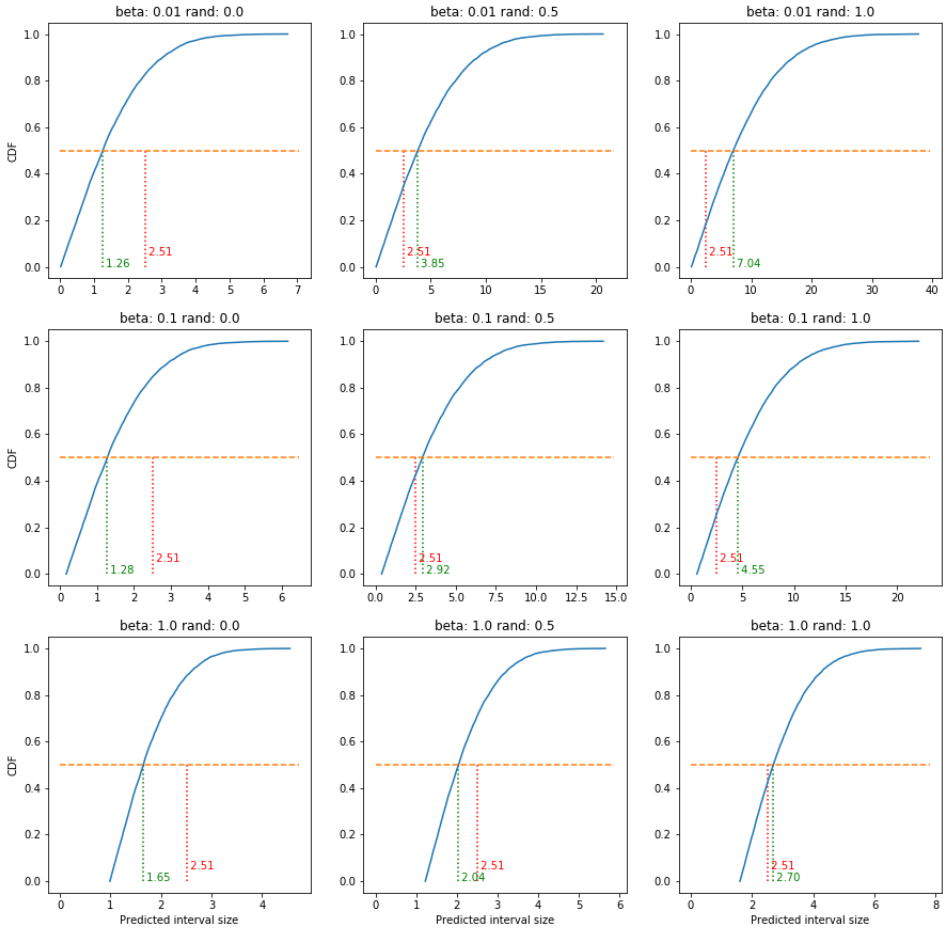


Figure 3: Cumulative interval sizes for the normal difficulty function

In Fig. 4, results are shown for when using uniformly distributed difficulty values. One may note the completely different distribution of interval sizes compared to when using the normally distributed difficulty function, here resulting in that the cumulative distribution increases linearly with the size of the predicted intervals. It can again be observed that the median efficiency is worse for the normalized approach, for lower values of β and higher degrees of randomness.

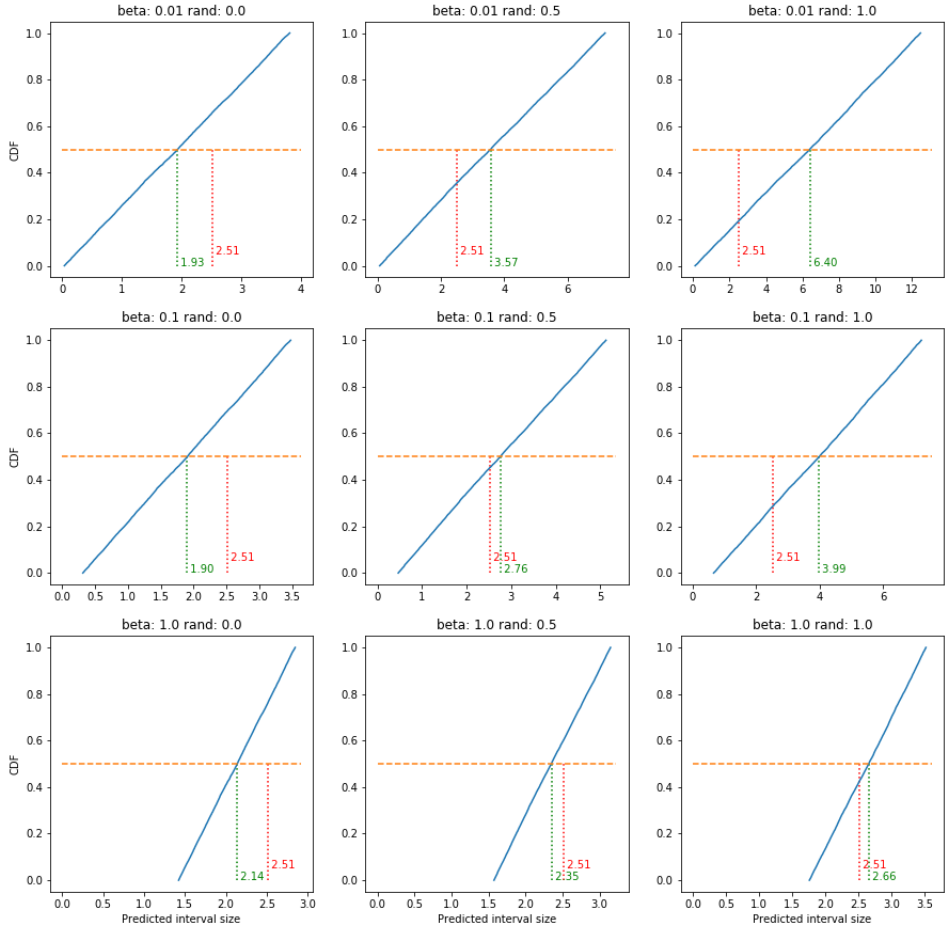


Figure 4: Cumulative interval sizes for the uniform difficulty function

In Fig. 5 and Fig. 6, predicted vs. actual (twice the observed absolute error) interval sizes are shown for when using normally and uniformly distributed difficulty values, respectively. One may note that the actual intervals are larger than the predicted ones in approximately 5% of the cases, as expected given the confidence level of 0.95, independently of the value for β and randomness level. One may also see that, independently of the employed difficulty function, the variance of the predicted interval size increases with the randomness (for any fixed β), hence demonstrating that the spread of the interval sizes indeed increases, as the correlation between the difficulty and true error decreases. In the extreme case, when there is no correlation between the difficulty and true error, the predicted intervals may be several times larger than the largest observed actual interval; more than five times larger for normally distributed difficulty estimates and two times for uniformly distributed difficulty estimates.

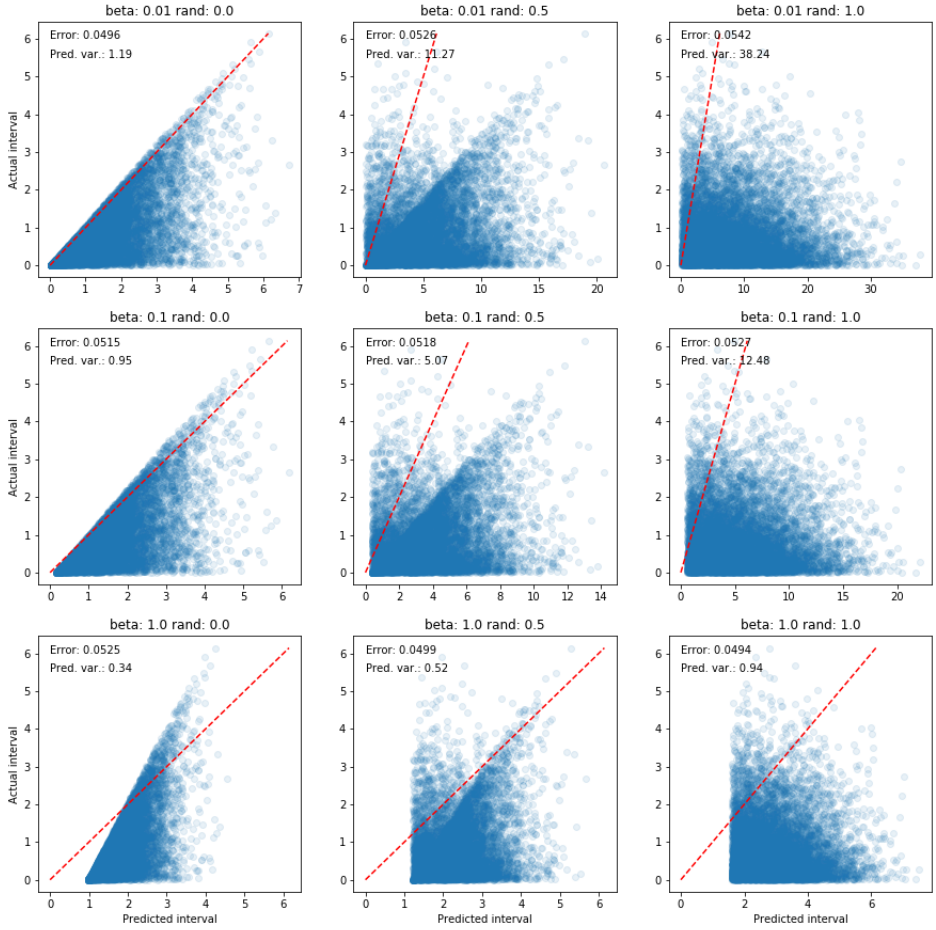


Figure 5: Predicted vs. actual interval sizes using normal difficulty

We now present results for Mondrian conformal regressors (MCR), using various number of bins (20, 50 and 100 bins); see Fig. 7 for results when using normally distributed difficulty values and Fig. 8 for when using uniformly distributed difficulty values, with the same levels of randomness as before. In contrast to the above results for normalized conformal regressors, the median predicted interval sizes are generally lower than for the standard approach, with the former approaching the latter with a higher degree of randomness in the difficulty function. Note that the efficiency is not very much affected by the number of bins, but clearly deteriorates as the quality of the difficulty function degrades, as expected.

Finally, in Fig. 9 and Fig. 10, predicted vs. actual interval sizes are shown for MCR when using normally and uniformly distributed difficulty values, respectively. One may note that the actual intervals are again larger than the predicted ones in approximately 5% of the cases, as expected, independently of the number of bins and randomness level. One may also see that, independently of the employed difficulty function, the variance of the predicted interval size decreases with the randomness (for any fixed number of bins), hence demonstrating that the size of the predicted intervals indeed becomes more uniform, as the difficulty function becomes less informative, as opposed to when using normalized

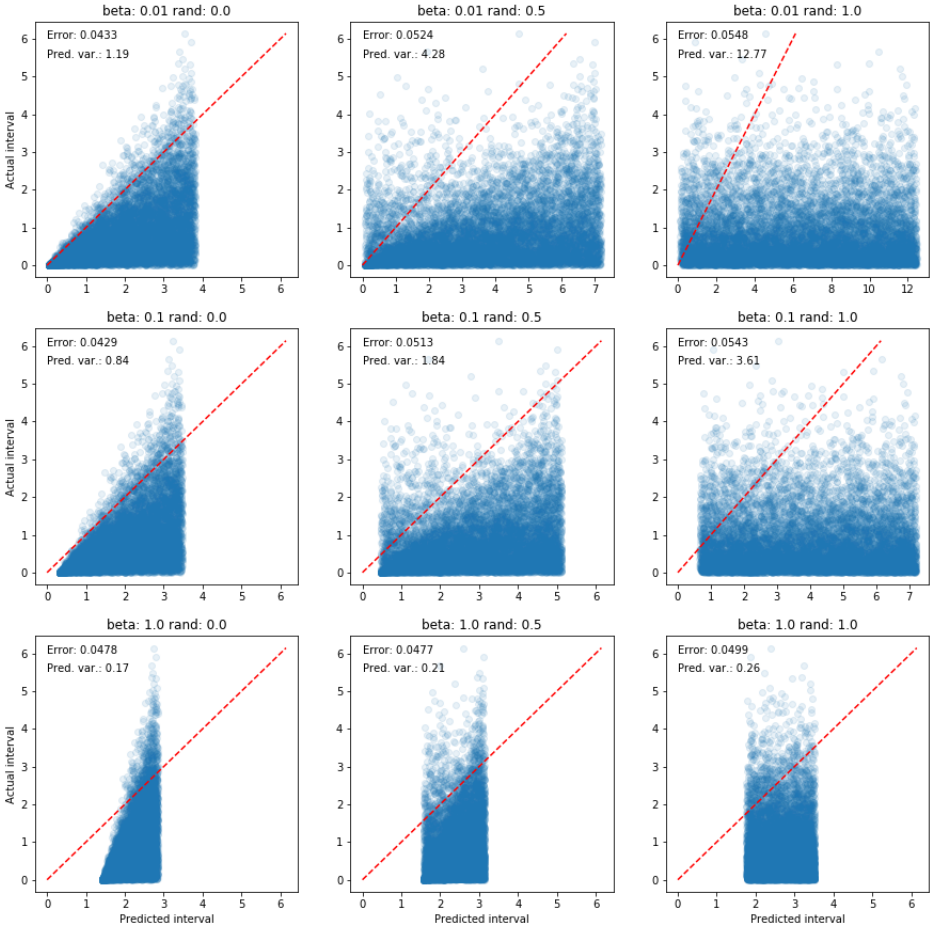


Figure 6: Predicted vs. actual interval sizes using uniform difficulty

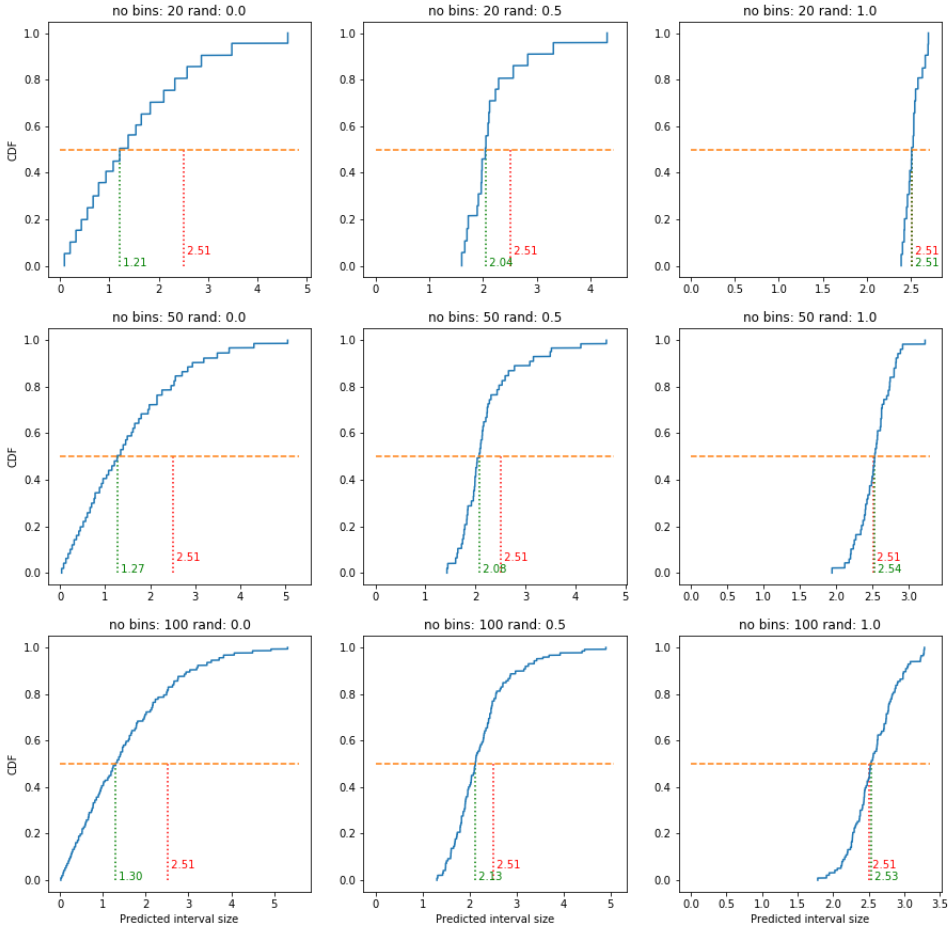


Figure 7: Cumulative interval sizes for MCR with the normal difficulty function

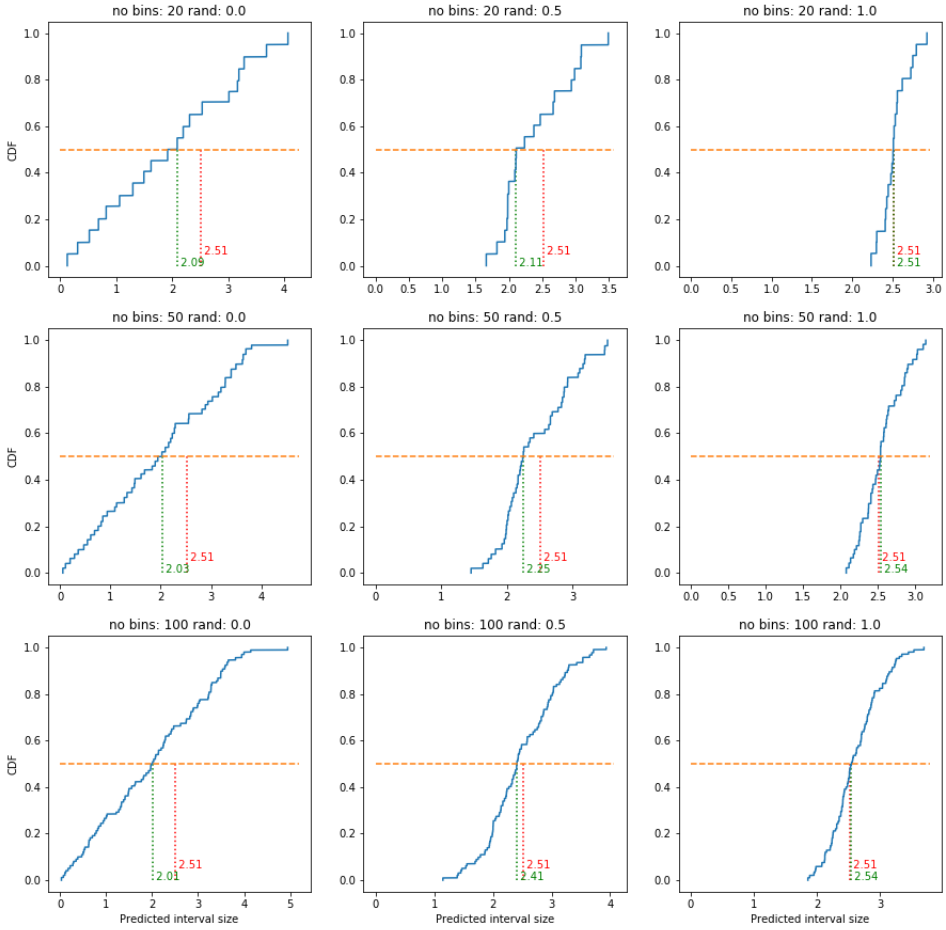


Figure 8: Cumulative interval sizes for MCR with the uniform difficulty function

conformal regression. One may also observe that the predicted intervals are never larger than the largest observed actual interval.

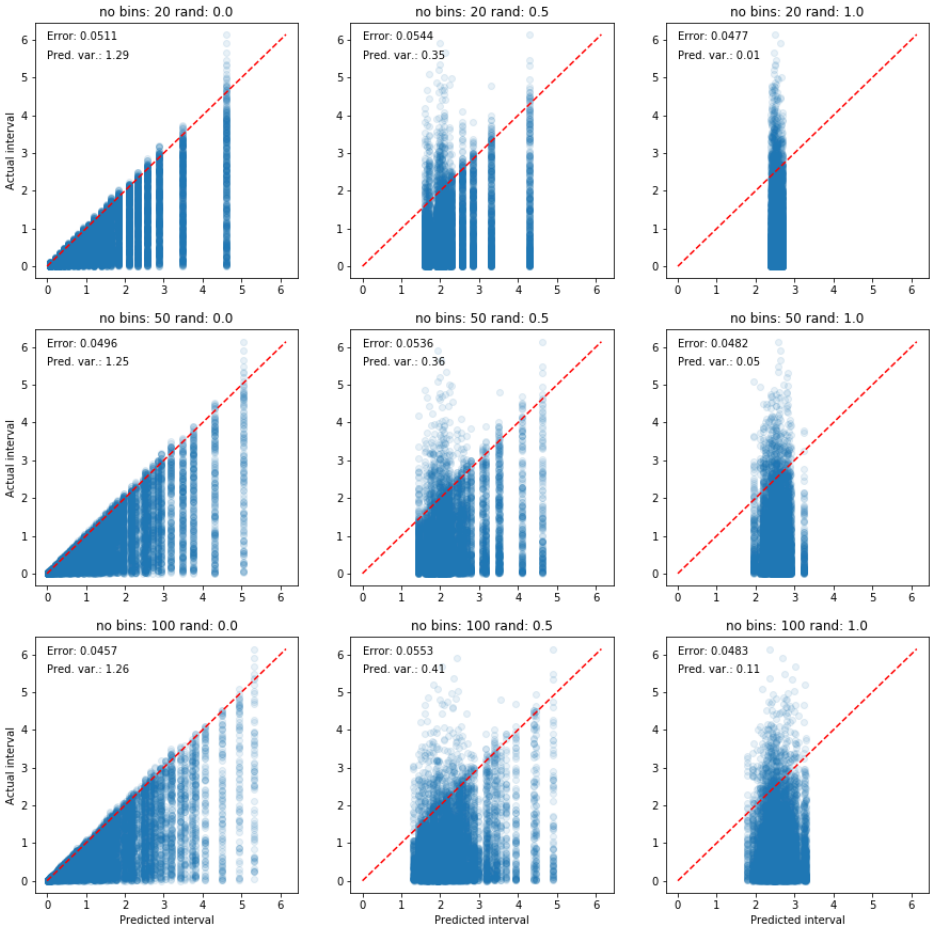


Figure 9: Predicted vs. actual interval sizes for MCR using normal difficulty

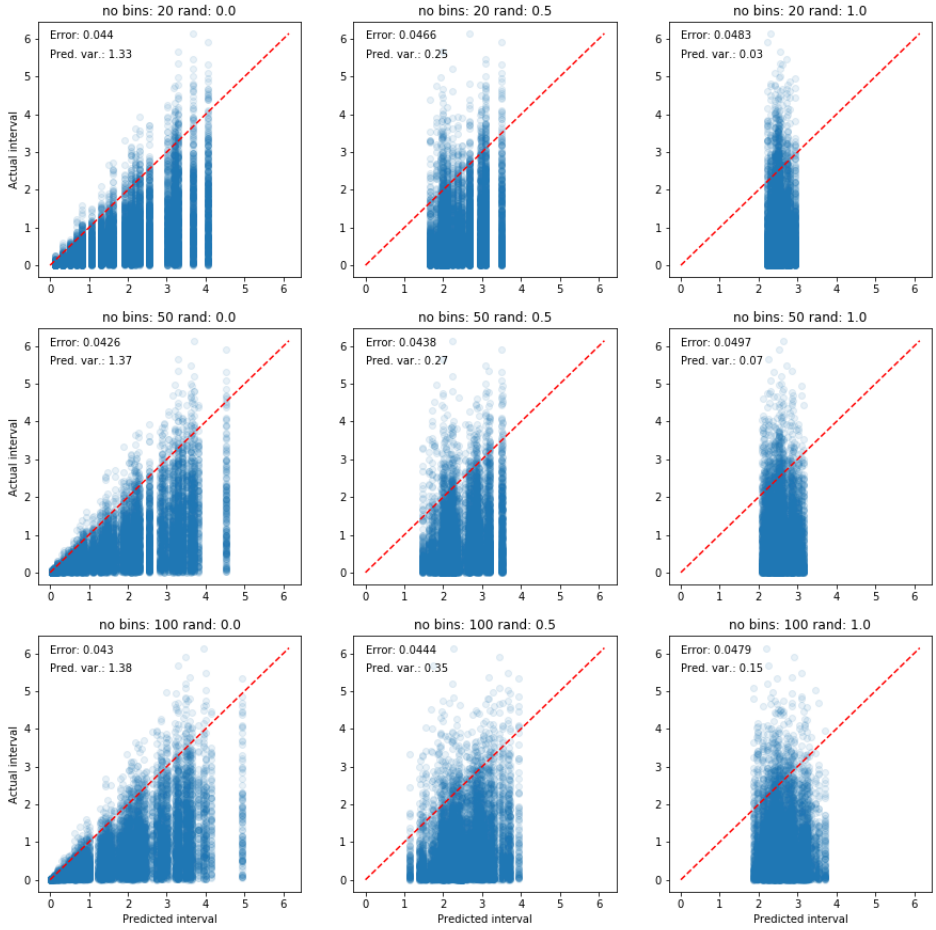


Figure 10: Predicted vs. actual interval sizes for MCR using uniform difficulty

4.2. Real-world datasets

4.2.1. Experimental setup

In the experiments with real-world datasets, we consider the same 33 datasets as used in e.g., (Johansson et al., 2014a; Boström et al., 2017); for characteristics of these datasets, see Table 1. We use random forests with 500 binary regression trees as the underlying model, with 1/3 of the available features randomly selected for evaluation at each node during tree generation. We employ 10-fold cross validation and investigate the output of the resulting conformal regressors for the confidence level 0.95. Since the underlying model employs bagging, we have here opted to employ out-of-bag calibration, rather than dividing the training data into a proper training set and a calibration set, leaving all training instances for both generating the underlying model and obtaining calibration scores. As proposed in (Boström et al., 2017), the variance of the predictions of the individual trees in the forest, was employed as an estimate of the difficulty for both normalized regression and Mondrian conformal regression. For the former, we employed a β of 0.01, as also used in (Boström et al., 2017). For the latter, a small random number was added to each difficulty estimate to resolve ties during binning, and the minimum bin size was set to 99, resulting in that

Table 1: Real-world datasets

Dataset	#Instances	#Features	Dataset	#Instances	#Features
abalone	4177	8	kin8nh	8192	8
anacalt	4052	7	kin8nm	8192	8
bank8fh	8192	8	laser	993	4
bank8fm	8192	8	mg	1385	6
bank8nh	8192	8	mortage	1048	15
bank8nm	8192	8	plastic	1649	2
boston	506	13	puma8fh	8192	8
comp	8192	12	puma8fm	8192	8
concreate	1030	8	puma8nh	8192	8
cooling	768	8	puma8nm	8192	8
deltaA	7129	5	quakes	2178	3
deltaE	9517	6	stock	950	9
friedm	1200	5	treasury	1048	15
heating	768	8	wineRed	1599	11
istanbul	536	7	wineWhite	4898	11
kin8fh	8192	8	wizmir	1461	9
kin8fm	8192	8			

a various number of bins were produced for the different datasets (the number of resulting bins for each dataset is reported in Table. 2 below).

4.2.2. Results

Table. 2 below shows the empirical error rates and the number of bins used by the Mondrian approach, averaged over the ten folds. From these results, it is evident that all setups produce valid conformal regressors. As seen in the last column, the number of bins varies a lot over the different datasets; from only four to 86.

Table 2: Empirical error rates. $\epsilon = 0.05$

Dataset	Std	Norm	Bin	#Bins	Dataset	Std	Norm	Bin	#Bins
abalone	.049	.051	.049	37	kin8nh	.050	.048	.052	74
anacalt	.051	.048	.019	36	kin8nm	.048	.048	.050	74
bank8fh	.050	.049	.049	74	laser	.048	.047	.059	9
bank8fm	.052	.052	.047	74	mg	.052	.051	.045	12
bank8nh	.050	.049	.048	74	mortage	.049	.046	.046	9
bank8nm	.051	.050	.048	74	plastic	.047	.050	.039	14.1
boston	.047	.057	.041	4	puma8fh	.049	.048	.049	74
comp	.050	.050	.049	74	puma8fm	.050	.048	.048	74
concreate	.050	.050	.039	9	puma8nh	.049	.050	.050	74
cooling	.049	.043	.043	6	puma8nm	.050	.050	.050	74
deltaA	.049	.050	.051	64	quakes	.052	.051	.050	19
deltaE	.049	.049	.051	86	stock	.044	.042	.039	8
friedm	.052	.052	.050	10	treasury	.054	.051	.045	9
heating	.047	.053	.043	6	wineRed	.050	.051	.049	14
istanbul	.054	.054	.052	4	wineWhite	.051	.049	.051	44
kin8fh	.050	.049	.049	74	wizmir	.053	.051	.042	13
kin8fm	.050	.050	.049	74	**Mean**	**.050**	**.050**	**.047**	**41.7**

Turning to the efficiency results in Table. 3 below, we see that both the normalized and the Mondrian versions obtained tighter intervals (using both the mean size and the median size) compared to a standard conformal regressor. This is confirmed by a Friedman test (Friedman, 1937), followed by Bergmann-Hommel's dynamic procedure (Bergmann and Hommel, 1988), showing the differences to be significant at α=0.05. There are no significant differences between the normalized and the Mondrian version though.

Looking specifically for unreasonable large intervals produced by the normalized approach, i.e., larger than twice the size of the largest observed absolute error on the calibration and test instances, such intervals exist on a majority of the datasets. While on most datasets, this applies only to a few instances, i.e., less than 1%, for a couple of datasets, the proportion is much larger. Specifically, for the *Plastic* dataset, approximately 16.6% of all intervals are unreasonably large. As follows from their construction, no such large intervals can be produced by the standard and Mondrian approaches.

The two first columns (Diff. est.) in Table. 4 below show the quality of the difficulty estimation, measured as the correlation between the estimated difficulty and the absolute error over all test instances. The following two columns (Var.) show the variance in interval sizes over the test instances. As stated above, a better difficulty estimation should lead to a larger variance in intervals, but here we see that this is the case only for the Mondrian version. In fact, the correlation between Diff. est. and Var. is negative (-0.301) for the normalized conformal regressor, but positive (0.253) for the Mondrian.

Table 3: Efficiency. $\epsilon = 0.05$

	Mean			Median				Mean			Median		
Dataset	Std	Norm	Bin	Std	Norm	Bin	Dataset	Std	Norm	Bin	Std	Norm	Bin
abalone	.321	.283	.302	.321	.258	.278	kin8nh	.493	.482	.486	.493	.478	.485
anacalt	.074	.051	.059	.074	.045	.000	kin8nm	.415	.396	.399	.415	.385	.389
bank8fh	.391	.343	.358	.391	.324	.362	laser	.092	.067	.091	.092	.055	.037
bank8fm	.228	.175	.183	.228	.165	.156	mg	.357	.209	.231	.357	.162	.152
bank8nh	.458	.417	.445	.458	.385	.443	mortage	.036	.033	.026	.036	.032	.017
bank8nm	.237	.149	.159	.237	.119	.097	plastic	.659	.803	.701	.659	.695	.644
boston	.287	.261	.276	.287	.202	.222	puma8fh	.562	.523	.533	.562	.515	.543
comp	.114	.108	.104	.114	.103	.100	puma8fm	.280	.274	.274	.280	.277	.279
concreate	.275	.246	.260	.275	.220	.240	puma8nh	.553	.523	.524	.553	.490	.502
cooling	.187	.141	.131	.187	.125	.115	puma8nm	.331	.319	.316	.331	.301	.320
deltaA	.156	.144	.150	.156	.138	.135	quakes	.704	.839	.742	.704	.745	.727
deltaE	.214	.214	.219	.214	.208	.214	stock	.096	.091	.096	.096	.085	.091
friedm	.298	.312	.308	.298	.309	.305	treasury	.042	.039	.032	.042	.038	.021
heating	.073	.068	.073	.073	.064	.066	wineRed	.498	.454	.449	.498	.438	.506
istanbul	.320	.334	.323	.320	.319	.321	wineWhite	.418	.371	.359	.418	.366	.397
kin8fh	.296	.291	.295	.296	.283	.278	wizmir	.079	.078	.086	.079	.076	.076
kin8fm	.183	.177	.181	.183	.172	.169	**Mean**	**.295**	**.279**	**.278**	**.295**	**.260**	**.263**
							Mean rank	**2.67**	**1.45**	**1.88**	**2.79**	**1.55**	**1.67**

Table 4: Difficulty estimation and variance in interval sizes

	Diff. est.		Var.			Diff. est.		Var.	
datasets	Norm	Bin	Norm	Bin	datasets	Norm	Bin	Norm	Bin
abalone	.358	.361	.007	.016	kin8nh	.212	.213	.008	.005
anacalt	.559	.567	.001	.025	kin8nm	.298	.298	.008	.007
bank8fh	.298	.300	.010	.008	laser	.658	.655	.002	.018
bank8fm	.483	.479	.003	.007	mg	.778	.774	.016	.029
bank8nh	.282	.282	.021	.019	mortage	.699	.699	.000	.001
bank8nm	.683	.679	.006	.025	plastic	.034	.035	.133	.036
boston	.565	.564	.031	.016	puma8fh	.277	.274	.016	.011
comp	.273	.271	.000	.002	puma8fm	.166	.165	.004	.003
concreate	.444	.449	.010	.010	puma8nh	.301	.303	.020	.011
cooling	.720	.728	.002	.008	puma8nm	.229	.231	.009	.004
deltaA	.404	.406	.000	.004	quakes	.159	.158	.160	.016
deltaE	.185	.181	.001	.002	stock	.351	.349	.000	.001
friedm	-.048	-.055	.002	.001	treasury	.547	.551	.000	.001
heating	.497	.515	.000	.002	wineRed	.403	.399	.024	.041
istanbul	.085	.091	.004	.001	wineWhite	.437	.437	.014	.034
kin8fh	.216	.219	.003	.004	wizmir	.222	.214	.000	.001
kin8fm	.262	.262	.001	.002	**Mean**	**.365**	**.365**	**.016**	**.011**

5. Concluding remarks

We have in this paper investigated conformal regressors using a difficulty estimation function to produce individualized prediction intervals. Experiments using both synthetic and real-world datasets show that such normalized conformal regressors suffer from two inherent but subtle drawbacks. First of all, the normalization can lead to unreasonably small or large intervals, in comparison to errors on the calibration set. Second, in order to improve the informativeness, better difficulty estimators should lead to more specific models, i.e., a larger variation in prediction intervals. The experimental results, however, conclusively show that this is not the case. As a solution to these problems, we suggested Mondrian conformal regressors, which bins the difficulty estimations into categories and then generates one prediction interval for each such category, using standard conformal regression. By their construction, the intervals of Mondrian conformal regressors can never be larger than twice the largest calibration set error. The experiments verify that the Mondrian variant is valid, as efficient as using normalization and significantly more efficient than the standard approach. Finally, it was shown that in contrast to when normalization is employed, for Mondrian conformal regressors, more informative difficulty estimators will indeed lead to more varied interval sizes.

Directions for future work include investigating alternative ways of forming the categories for the Mondrian conformal regressors, e.g., based on features of the objects, possibly in addition to using the difficulty estimates. Alternative binning procedures, which do not necessarily result in equal-sized bins, could also be investigated. Dynamic formation of bins, similar to Venn-Abers predictors (Vovk and Petej, 2012), is one such possibility. Previous approaches to handle small calibration sets, e.g., as suggested in (Johansson et al., 2015), are expected to be effective also for Mondrian conformal regressors, making the approach less conservative for smaller bin sizes.

Acknowledgments

HB was partly funded by the Swedish Foundation for Strategic Research (CDA, grant no. BD15-0006) and the Vinnova program for Strategic Vehicle Research and Innovation (FFI) Transport Efficiency (CODA, grant no. 2016-05143). UJ was partly funded by the Swedish Knowledge Foundation (DATAKIND 20190194).

References

Beate Bergmann and Gerhard Hommel. Improvements of general multiple test procedures for redundant systems of hypotheses. In *Multiple Hypotheses Testing*, pages 100–115. Springer, 1988.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Ann. Math. Artif. Intell.*, 81(1-2): 125–144, 2017.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of American Statistical Association*, 32:675–701, 1937.

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann, 1998.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176, 2014a. ISSN 0885-6125.

Ulf Johansson, Cecilia Sönströd, Henrik Boström, and Henrik Linusson. Regression trees for streaming data with local performance guarantees. In *IEEE International Conference on Big Data*. IEEE, 2014b.

Ulf Johansson, Ernst Ahlberg, Henrik Boström, Lars Carlsson, Henrik Linusson, and Cecilia Sönströd. Handling small calibration sets in mondrian inductive conformal regressors. In *Statistical Learning and Data Sciences*, pages 271–280. Springer, 2015.

Harris Papadopoulos and Haris Haralambous. Neural networks regression inductive conformal predictor and its application to total electron content prediction. In *Artificial Neural Networks – ICANN 2010*, volume 6352 of *Lecture Notes in Computer Science*, pages 32–41. Springer Berlin Heidelberg, 2010.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, pages 815–840, 2011.

Fan Shi, Cheng Soon Ong, and Christopher Leckie. Applications of class-conditional conformal predictor in multi-class classification. In *12th International Conference on Machine Learning and Applications*. IEEE, 2013.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., 2005.