Mondrian Conformal Predictive Distributions

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 Computing, Jönköping University, Sweden

Tuwe Löfström TUWE.LOFSTROM@JU.SE
Dept. of Computing, Jönköping University, Sweden

Editor: Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin and Khuong An Nguyen

Abstract

The distributions output by a standard (non-normalized) conformal predictive system all have the same shape but differ in location, while a normalized conformal predictive system outputs distributions that differ also in shape, through rescaling. An approach to further increasing the flexibility of the framework is proposed, called *Mondrian conformal predictive distributions*, which are (standard or normalized) conformal predictive distributions formed from multiple Mondrian categories. The effectiveness of the approach is demonstrated with an application to regression forests. By forming categories through binning of the predictions, it is shown that for this model class, the use of Mondrian conformal predictive distributions significantly outperforms the use of both standard and normalized conformal predictive distributions with respect to the continuous-ranked probability score. It is further shown that the use of Mondrian conformal predictive distributions results in as tight prediction intervals as produced by normalized conformal regressors, while improving upon the point predictions of the underlying regression forest.

Keywords: Conformal predictive systems, Conformal predictive distributions, Mondrian conformal predictive distributions, Continuous ranked probability score

1. Introduction

In contrast to a standard regression model, which outputs point predictions, and a conformal regressor, which outputs prediction intervals, a conformal predictive system for regression outputs cumulative probability distributions over the possible target values (Vovk et al., 2020). Following the terminology used in conjunction with conformal regressors (Johansson et al., 2014a), one may distinguish between *standard* and *normalized* conformal predictive systems, where the latter employ a quality estimate (σ_i) that is specific to each object ($\boldsymbol{x}_i$) for which a prediction is to be made, while the former use the same $\sigma_i = c$ for all objects, where c is some positive constant. As a consequence, distributions output by a standard conformal predictive system may differ only in their location and not in their shape. The distributions output by a normalized conformal predictive system may differ also in shape, as they are rescaled along the target value dimension using the quality estimates; the distribution for a high-quality prediction will be scaled down along this dimension, resulting in a steeper increase of the cumulative probability distribution close to the point prediction, while the distribution of a low-quality prediction will be upscaled, resulting in a more uniform increase over the range of possible target values.

Conformal predictive systems can be seen as generalizations of both standard and conformal regressors, as they can easily be constrained to produce point predictions and prediction intervals, respectively. The latter can be obtained from a conformal predictive distribution through considering the lower and upper percentiles of interest, e.g., using the 2.5th and 97.5th percentiles for a 95% level of confidence. Conformal predictive systems can also be used for calibrating point predictions, e.g., by using the 50th percentile (median) or the mean of the distribution, rather than the prediction of the underlying model. The latter may be useful if the underlying model systematically under- or overestimates the target values, since using the median (or mean) of the distribution may correct for this by pushing the prediction upwards or downwards.

However, when the residuals (differences between actual and predicted values) are heteroscedastic, e.g., their distribution is not independent of the actual predictions, neither standard nor normalized predictive distributions may be very effective tools for calibrating the predictions. To see this, consider an underlying model that has a tendency to overestimate low actual values and underestimate high actual values. This means that the residuals for low predicted values will be negative on average, while they will be positive on average for high predicted values. A standard (non-normalized) conformal predictive distribution cannot correct for this, as the shape of the predictive distributions will be the same for all predictions. This means that the distribution can suggest moving the predictions at most in one direction, since the median (or mean) of the distribution will always be on the same side relative to the underlying prediction. Using a normalized conformal predictive distribution will not fix the problem, since the quality estimate used for scaling is always positive, and hence the direction of the correction suggested by the normalized conformal predictive distribution will be the same as for the standard distribution.

In order to overcome the above problem, an alternative approach to producing conformal predictive distributions is proposed, called *Mondrian conformal predictive distributions*. The approach borrows the idea of Mondrian conformal prediction (Vovk et al., 2005), which originally was proposed for allowing to control the error levels of objects falling into *a priori* defined categories. For example, by defining categories according to the class labels, one can guarantee the same error level for all classes. More recently, Boström and Johansson (2020) proposed *Mondrian conformal regressors*, where this idea was used to handle two problems of normalized conformal regressors; the prediction intervals may be several times larger (or smaller) than the largest (smallest) previously observed error, and the sizes of the intervals become less uniform with less informative quality (difficulty) estimates. By forming categories through binning of the quality estimates, it was shown that the size of the intervals is bounded by the size of the largest observed error and that non-informative quality estimates result in more uniformly sized intervals.

To the best of our knowledge, the Mondrian approach has however not been applied to conformal predictive systems. In this paper, we propose one such approach, which uses the categories to form multiple (standard or normalized) conformal predictive distributions. For the example above, we may consider using, say two, categories based on the predictions of the underlying model. For the first of these, which corresponds to low-valued predictions, the conformal predictive distribution can correct for the systematic overestimations, while for the second category, which corresponds to high-valued predictions, the conformal predictive distribution can correct for the systematic underestimations. It should be noted that the prediction intervals output by conformal regressors do not directly allow for such corrections, as they are centered around the point predictions and provide no information on whether the true target can be expected to be higher or lower than the point prediction. In contrast to the previously proposed Mondrian conformal regressors, the categories for Mondrian conformal predictive distributions are here proposed to be formed using the predictions and not the quality estimates. The latter estimates may however still be used within the categories to form normalized conformal predictive distributions.

We will present results from a large-scale empirical evaluation of Mondrian conformal predictive systems, using regression forests as the underlying model. We will consider Mondrian distributions formed from categories corresponding to bins of the predicted values, where both standard and normalized conformal predictive distributions will be generated for the categories. The resulting Mondrian conformal predictive systems will be compared to standard and normalized conformal predictive systems, respectively, with respect to continuous ranked probability score. They will also be compared to standard and normalized conformal regressors on the tasks of generating efficient prediction intervals and accurate point predictions.

In the next section, we briefly describe conformal predictive distributions. In Section 3, we introduce the alternative Mondrian-based approach; Mondrian conformal predictive distributions. In Section 4, we first illustrate the approach and then present results from comparing the novel approach to standard and normalized conformal predictive systems, as well as to normalized (and standard) conformal regressors on a set of 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 object 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. A conformal predictive system utilizing the same idea is called a *split conformal predictive system* (Vovk et al., 2020).

A split conformal predictive system relies on an isotonic split conformity measure A to calculate a cumulative probability with respect to a label y_i, given some object $\boldsymbol{x}_i$ and underlying model h.

Given a training set Z_{tr}, and a test object x, a *conformal predictive distribution* is constructed by a *(split) conformal predictive system* as follows:

1. Divide the training sequence Z_{tr} into two disjoint subsets; the proper training set Z_t and the calibration set $Z_c = \{(x_1, y_1), \ldots, (x_q, y_q)\}$
2. Train the underlying model h using Z_t
3. Let $\hat{y}_i = h(x_i)$, for $i \in 1, ..., q$
4. Obtain quality estimates for all the calibration examples $\hat{\sigma}_i$, for $i \in 1, ..., q$
5. Make a point prediction for the test object $\hat{y} = h(x)$ and estimate its quality $\hat{\sigma}$
6. Produce a list of calibration scores using

   $$C_i = \hat{y} + \frac{\hat{\sigma}}{\hat{\sigma}_i}(y_i - \hat{y}_i)$$

   for $i \in 1, ..., q$
7. Sort $C_1, ..., C_q$ in increasing order, resulting in $C_{(1)}, ..., C_{(q)}$
8. Set $C_{(0)} = -\infty$ and $C_{(q+1)} = \infty$
9. Let $\tau \in U(0, 1)$
10. Return the predictive distribution:

    $$Q(y) = \begin{cases} \frac{n+\tau}{q+1} & \text{if } y \in \left(C_{(n)}, C_{(n+1)}\right) \text{ for } n \in \{0, ..., q\} \\ \frac{n'-1+(n''-n'+2)\tau}{q+1} & \text{if } y = C_{(n)} \text{ for } n \in \{1, ..., q\} \end{cases}$$

    where $n' = \min\{m | C_{(m)} = C_{(n)}\}$ and $n'' = \max\{m | C_{(m)} = C_{(n)}\}$

Note that for a sequence of test examples, only step $5 - 10$ above need to be repeated. We adopt a similar terminology as for conformal regressors, and refer to conformal predictive systems as *standard conformal predictive systems*, in case $\hat{\sigma}_i = \hat{\sigma} = c$, for some constant $c > 0$, i.e., the quality estimate is independent of the predicted object. We refer to conformal predictive systems that are not standard as *normalized conformal predictive systems*. Approaches to estimating the quality include training a separate model to predict the size of the error, e.g., using kNN or ANN as in (Johansson et al., 2014a); others exploit properties of the underlying model, e.g., using disagreement (variance) of the trees in a random forest (Boström et al., 2017).

Prediction intervals for a chosen significance level ϵ can be defined by

$$[C_{Q''_{.\epsilon/2}}, C_{Q'_{.100-\epsilon/2}}] \tag{1}$$

where $Q''_{.p} = \max\{m | Q_{(m)} < p/100\}$ and $Q'_{.p} = \min\{m | Q_{(m)} > p/100\}$. A point prediction can similarly be defined by $C_{Q'_{.50}}$.

The above way of forming prediction intervals contrasts to (standard or normalized) conformal regressors by not enforcing that the underlying prediction is placed in the middle of the interval. In fact, when using conformal predictive systems to generate prediction intervals, it may very well be the case that the underlying prediction falls outside the prediction interval.

3. Mondrian Conformal Predictive Distributions

In Mondrian conformal predictors (Vovk et al., 2005), the available calibration examples 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). Recently, Mondrian conformal regressors were proposed in (Boström and Johansson, 2020). Until now, however, Mondrian conformal prediction has, to the best of our knowledge, not been applied to conformal predictive distributions.

A *(split) Mondrian conformal predictive system* produces a Mondrian conformal predictive distribution in the following way, given a training sequence Z_{tr}, and a test object x:

1. Divide the training sequence Z_{tr} into two disjoint subsets; the proper training set Z_t and the calibration set $Z_c = (x_1, y_1), \ldots, (x_q, y_q)$
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. For each category κ_j, let $\hat{y}_{ji} = h(x_{ji})$ and let $\hat{\sigma}_{ji}$ be the corresponding quality estimate, for each $x_{ji} \in Z_{cj}$
5. Make a point prediction for the test object $\hat{y} = h(x)$ and estimate its quality $\hat{\sigma}$
6. Identify which category κ_j the test object belongs to and produce a list of calibration scores using

$$C_{ji} = \hat{y} + \frac{\hat{\sigma}}{\hat{\sigma}_{ji}} (y_{ji} - \hat{y}_{ji})$$

   for $i \in 1, ..., q$, where $q = |Z_{cj}|$
7. Sort $C_{j1}, ..., C_{jq}$ in increasing order, resulting in $C_{j(1)}, ..., C_{j(q)}$
8. Set $C_{j(0)} = -\infty$ and $C_{j(q+1)} = \infty$.
9. Let $\tau \in U(0, 1)$

10. Return the predictive distribution:

$$Q(y) = \begin{cases} \frac{n+\tau}{q+1} & \text{if } y \in \left(C_{j(n)}, C_{j(n+1)}\right) \text{ for } n \in \{0, ..., q\} \\ \frac{n'-1+(n''-n'+2)\tau}{q+1} & \text{if } y = C_{j(n)} \text{ for } n \in \{1, ..., q\} \end{cases}$$

where $n' = \min\{m|C_{j(m)} = C_{j(n)}\}$ and $n'' = \max\{m|C_{j(m)} = C_{j(n)}\}$

For Mondrian conformal classification, the possible class labels are often used to define the categories, while for Mondrian conformal regression (Boström and Johansson, 2020), the categories are defined using the quality (difficulty) estimate σ. We here propose to form the categories of the Mondrian taxonomy through binning of the predictions, using equal-sized bins, similar to what has been proposed for classification problems in the context of Venn prediction (Zhou et al., 2014). To handle the special case when the number of identical predictions is larger than the bin size, we assume that a very small random number ξ is added to each prediction, which allows for that approximately the same number of examples 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 size of each partition of the calibration set is sufficiently large, e.g., to allow for prediction intervals with specified level of confidence to be extracted.

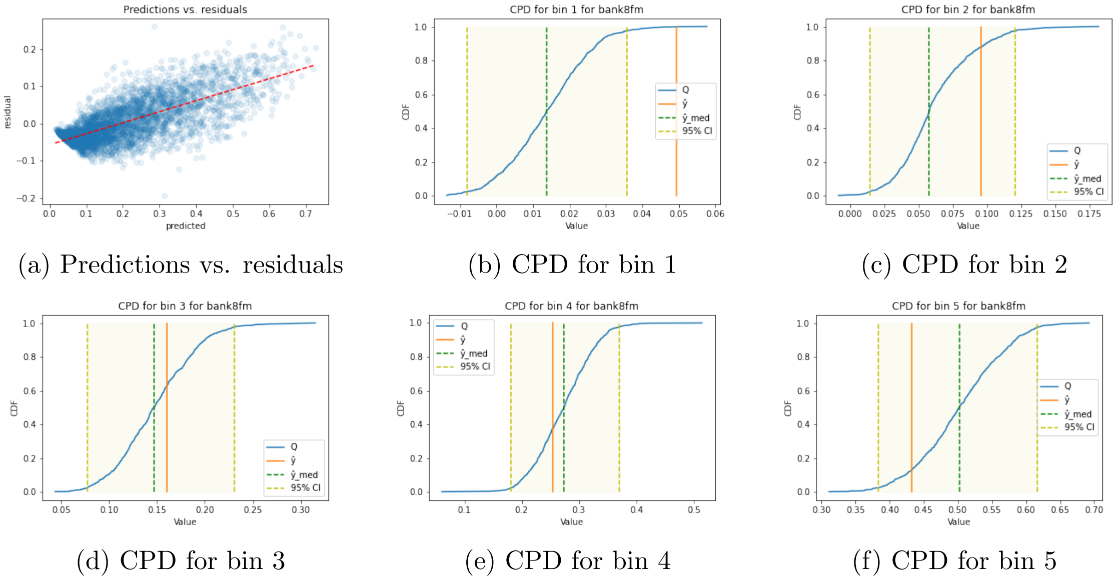
4. Experiments

First, we illustrate the proposed approach using the bank8fm dataset. After that, we present results from using different strategies to forming predictive systems on real-world datasets.

4.1. Illustration using the bank8fm dataset

In this demonstration, half of the data is used as a proper training set and the other half for calibration. The underlying model is a regression forest with 500 trees. Fig. 1a plots the predictions vs. the residuals. Interestingly enough, there is a clear trend showing that the residuals are larger for the higher predictions. With this in mind, we divide the predictions into five bins, with an equal number of examples in each bin, and then form a standard (i.e., non-normalized) conformal predictive distribution (CPD), given the mean prediction of each bin. Fig. 1b – Fig. 1f, show the CPD for each bin, specifically comparing the mean prediction of each bin (indicated by an orange solid line) to the median of the CPD (indicated by a green dashed line).

If we would consider using the medians of the CPDs as point predictions, instead of the original mean predictions, we can see that they would be substantially modified. For the first three bins, the predictions would be lowered, but for the last two, they would instead be increased. In particular, in the first bin, the prediction from the underlying model is actually outside of the 95%-confidence interval (indicated by the yellow dashed lines), see Fig. 1b, showing the potential of a CPD to improve predictions by calibration.



(a) Predictions vs. residuals

(b) CPD for bin 1

(c) CPD for bin 2

(d) CPD for bin 3

(e) CPD for bin 4

(f) CPD for bin 5

Figure 1: Illustration using the bank8fm dataset

4.2. Real-world datasets

4.2.1. Experimental setup

In the experimentation, we utilize random forests with 500 binary regression trees as the underlying models. Here, 1/3 of the available features were randomly selected for consideration in each split optimisation during the generation. Following the findings in (Werner et al., 2020), we have opted for using out-of-bag-calibration, thus making all training examples available for both training the underlying model and obtaining calibration scores. As a quality estimate, the variance of the predictions from the individual trees in the forest was used in all normalised setups, as originally proposed for conformal regressors in (Boström et al., 2017) and also investigated for conformal predictive systems in (Werner et al., 2020).

For the evaluation, we compare the suggested Mondrian approaches to both conformal regression setups and standard predictive conformal distributions. All-in-all, six different setups are evaluated:

- CR: Standard conformal regressor as described in, for instance, (Boström et al., 2017), i.e., using the absolute error as the non-conformity measure, and utilizing out-of-bag estimates for calibration.
- CRn: Normalized conformal regressor using the variance of the predictions from the individual trees as the difficulty function, as described in (Boström et al., 2017).
- CPS: Standard conformal predictive system as described in Section 2.
- CPSn: Normalized conformal predictive system as described in Section 2.
- MCPS: Mondrian conformal predictive system using five bins as described in Section 3.

Table 1: Real-world datasets

Dataset	#Examples	#Features	Dataset	#Examples	#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
concrete	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			

- MCPSn: Normalized Mondrian conformal predictive system using five bins as described in Section 3.

All-in-all 33 publicly available data sets, previously used in e.g., (Johansson et al., 2014a; Boström et al., 2017; Boström and Johansson, 2020; Werner et al., 2020), were employed; for characteristics of these data sets, see Table 1. For the actual evaluation, standard 10-fold cross-validation was used.

4.2.2. Results

We first look at the the quality of the predictions from the four variants of conformal predictive systems using the loss function *continuous ranked probability score* (CRPS) (Vovk et al., 2020) in Table 2. Here, the two Mondrian approaches clearly outperform both standard and normalized conformal predictive systems. This is confirmed by a Friedman test (Friedman, 1937), followed by a Nemenyi post-hoc test (Nemenyi, 1963), showing the differences to be significant at α=0.05, see Fig. 2. This is of course a very strong result for the suggested Mondrian conformal predictive system and consequently a key finding of this study.

Turning to the efficiency results in Table. 3, we see from the mean interval sizes that the normalized versions produce the most informative predictions. Specifically, the Mondrian normalized version obtained the tightest intervals, on average. Looking at the results of the statistical testing in Fig. 3, the only significant difference is, however, that the standard conformal regressor and standard conformal predictive system are significantly worse than the other four alternatives.

Considering median interval sizes instead, the above results are confirmed; see Table. 4 and Fig. 4.

Table 2: CRPS

	CPS	CPSn	MCPS	MCPSn
abalone	.040	.039	.039	.039
anacalt	.007	.007	.007	.006
bank8fh	.053	.053	.048	.048
bank8fm	.030	.030	.021	.020
bank8nh	.056	.056	.053	.053
bank8nm	.025	.025	.020	.019
boston	.036	.034	.033	.033
comp	.015	.015	.014	.014
concrete	.036	.035	.033	.032
cooling	.021	.020	.019	.018
deltaA	.020	.019	.019	.019
deltaE	.029	.029	.029	.028
friedm	.042	.042	.031	.031
heating	.009	.009	.009	.009
istanbul	.044	.044	.044	.044
kin8fh	.042	.042	.040	.039
kin8fm	.025	.025	.018	.018
kin8nh	.072	.071	.068	.068
kin8nm	.060	.059	.051	.050
laser	.012	.011	.011	.010
mg	.039	.035	.038	.034
mortage	.004	.004	.003	.003
plastic	.094	.096	.093	.096
puma8fh	.082	.081	.078	.078
puma8fm	.042	.042	.034	.034
puma8nh	.081	.080	.075	.074
puma8nm	.052	.052	.039	.038
quakes	.095	.096	.091	.094
stock	.013	.013	.013	.013
treasury	.004	.004	.004	.004
wineRed	.060	.058	.054	.053
wineWhite	.052	.050	.048	.046
wizmir	.011	.011	.011	.011
Mean	**.039**	**.039**	**.036**	**.036**
Ranks	**3.70**	**3.09**	**2.00**	**1.21**

Table 3: Efficiency (mean)

	CR	CRn	CPS	CPSn	MCPS	MCPSn
abalone	.320	.283	.325	.284	.291	.288
anacalt	.074	.050	.076	.052	.087	.066
bank8fh	.392	.341	.368	.353	.349	.352
bank8fm	.228	.176	.210	.176	.144	.140
bank8nh	.461	.413	.432	.414	.403	.412
bank8nm	.238	.150	.234	.151	.146	.134
boston	.294	.267	.282	.265	.285	.294
comp	.114	.107	.116	.108	.108	.106
concrete	.274	.246	.274	.246	.281	.246
cooling	.187	.142	.190	.142	.140	.126
deltaA	.155	.145	.155	.146	.155	.145
deltaE	.215	.214	.215	.214	.217	.214
friedm	.299	.315	.301	.316	.239	.247
heating	.074	.068	.070	.066	.073	.064
istanbul	.321	.333	.314	.334	.369	.379
kin8fh	.297	.290	.303	.291	.280	.278
kin8fm	.183	.177	.184	.177	.131	.126
kin8nh	.492	.482	.491	.481	.477	.470
kin8nm	.417	.395	.414	.394	.372	.347
laser	.090	.066	.092	.066	.117	.063
mg	.358	.210	.358	.211	.349	.221
mortage	.037	.033	.036	.033	.030	.029
plastic	.657	.803	.679	.788	.678	.833
puma8fh	.562	.522	.560	.521	.543	.534
puma8fm	.281	.273	.280	.275	.245	.241
puma8nh	.554	.525	.552	.523	.540	.534
puma8nm	.330	.319	.331	.318	.293	.296
quakes	.709	.848	.693	.820	.633	.757
stock	.095	.089	.095	.090	.098	.090
treasury	.043	.038	.043	.038	.040	.036
wineRed	.500	.450	.498	.456	.495	.464
wineWhite	.417	.371	.418	.374	.400	.369
wizmir	.080	.078	.079	.077	.079	.079
Mean	**.295**	**.279**	**.293**	**.279**	**.275**	**.272**
Ranks	**4.97**	**2.58**	**4.82**	**3.00**	**3.33**	**2.30**

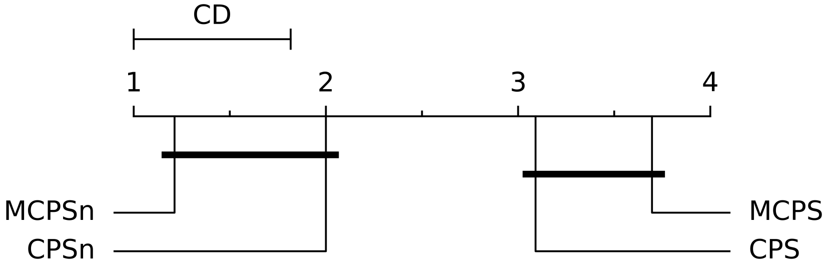


Figure 2: CRPS ranks

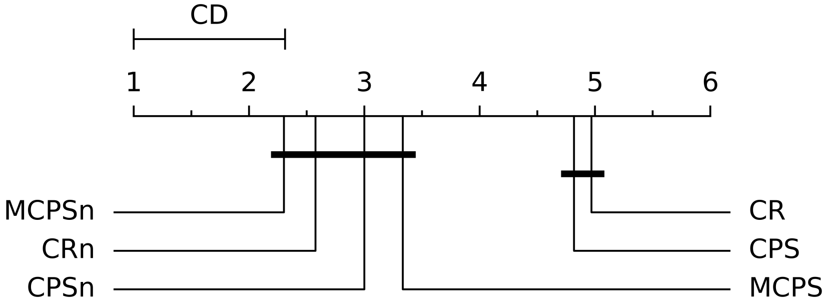


Figure 3: Efficiency ranks

Turning to the predictive performance, Table. 5 shows the mean absolute errors. Here, the two Mondrian approaches clearly outperform the other setups. In fact, statistical testing shows that the differences are significant at $\alpha = 0.05$, see Fig. 5. So, based on these results considering a large number of real-world data sets, the Mondrian approach leads to better predictors than the alternatives.

When considering mean squared error instead of mean absolute error, the differences between the setups are actually very small in absolute numbers. Still, when looking at the mean ranks over all data sets, the normalized Mondrian conformal predictive systems are the most accurate, see Table. 6. Here, however, as seen in Fig. 6, the statistical tests identify fewer significant differences.

Summarising the main experiment, we see that the novel Mondrian conformal predictive systems outperformed standard and normalized conformal predictive systems with regard to the continuous ranked probability score loss metric. In addition, the informativeness, as measured using the prediction interval sizes, was generally better for the normalized versions compared to standard conformal regressors and conformal predictive systems. Finally, also when evaluating the predictive performance in terms of mean absolute or mean-squared error, the Mondrian variants were the most accurate, thus demonstrating their inherent capability of improving the predictions of the underlying models.

Table 4: Efficiency (median)

	CR	CRn	CPS	CPSn	MCPS	MCPSn
abalone	.320	.258	.325	.259	.273	.261
anacalt	.074	.044	.076	.046	.000	.000
bank8fh	.392	.322	.368	.334	.344	.345
bank8fm	.228	.165	.210	.166	.155	.146
bank8nh	.461	.381	.432	.384	.387	.389
bank8nm	.238	.120	.234	.121	.095	.087
boston	.294	.206	.282	.208	.257	.242
comp	.114	.103	.116	.103	.105	.102
concrete	.274	.219	.274	.222	.244	.213
cooling	.187	.124	.190	.124	.134	.110
deltaA	.155	.138	.155	.140	.128	.125
deltaE	.215	.208	.215	.209	.216	.205
friedm	.299	.309	.301	.311	.234	.240
heating	.074	.064	.070	.062	.076	.063
istanbul	.321	.319	.314	.322	.371	.358
kin8fh	.297	.281	.303	.283	.272	.267
kin8fm	.183	.172	.184	.173	.118	.117
kin8nh	.492	.478	.491	.477	.489	.465
kin8nm	.417	.385	.414	.384	.369	.336
laser	.090	.055	.092	.054	.117	.050
mg	.358	.162	.358	.162	.381	.168
mortage	.037	.032	.036	.032	.023	.023
plastic	.657	.694	.679	.689	.651	.725
puma8fh	.562	.515	.560	.515	.549	.527
puma8fm	.281	.274	.280	.277	.271	.249
puma8nh	.554	.492	.552	.491	.569	.495
puma8nm	.330	.301	.331	.301	.318	.278
quakes	.709	.758	.693	.733	.635	.705
stock	.095	.084	.095	.085	.092	.086
treasury	.043	.037	.043	.037	.033	.030
wineRed	.500	.432	.498	.437	.479	.443
wineWhite	.417	.364	.418	.369	.382	.364
wizmir	.080	.075	.079	.075	.078	.078
Mean	**.295**	**.260**	**.293**	**.260**	**.268**	**.251**
Ranks	**5.06**	**2.58**	**4.97**	**2.91**	**3.39**	**2.09**

Table 5: Mean absolute errors

	CR	CRn	CPS	CPSn	MCPS	MCPSn
abalone	.055	.054	.053	.053	.053	.053
anacalt	.008	.008	.008	.008	.008	.008
bank8fh	.076	.076	.074	.076	.067	.067
bank8fm	.043	.043	.042	.045	.029	.029
bank8nh	.082	.082	.075	.078	.073	.072
bank8nm	.034	.034	.031	.034	.028	.027
boston	.048	.048	.048	.048	.046	.045
comp	.020	.020	.020	.020	.020	.019
concrete	.049	.049	.049	.049	.045	.044
cooling	.026	.026	.027	.026	.026	.026
deltaA	.026	.026	.026	.026	.026	.026
deltaE	.040	.040	.040	.040	.040	.040
friedm	.058	.059	.059	.059	.043	.043
heating	.013	.013	.013	.013	.012	.012
istanbul	.060	.060	.061	.060	.061	.061
kin8fh	.059	.059	.059	.059	.056	.056
kin8fm	.035	.035	.035	.035	.025	.025
kin8nh	.102	.102	.102	.102	.097	.097
kin8nm	.085	.085	.085	.085	.072	.071
laser	.014	.014	.014	.014	.014	.013
mg	.051	.051	.051	.051	.049	.048
mortage	.005	.005	.005	.005	.005	.005
plastic	.132	.131	.129	.129	.128	.128
puma8fh	.117	.117	.117	.117	.111	.111
puma8fm	.060	.060	.060	.060	.048	.048
puma8nh	.117	.117	.117	.117	.105	.105
puma8nm	.077	.077	.078	.077	.054	.052
quakes	.137	.138	.134	.132	.129	.131
stock	.019	.019	.019	.019	.018	.018
treasury	.005	.005	.005	.005	.005	.005
wineRed	.081	.081	.081	.081	.072	.071
wineWhite	.069	.069	.069	.069	.064	.063
wizmir	.015	.015	.015	.015	.015	.015
Mean	**.055**	**.055**	**.055**	**.055**	**.050**	**.049**
Ranks	**4.64**	**4.70**	**3.88**	**4.15**	**2.15**	**1.48**

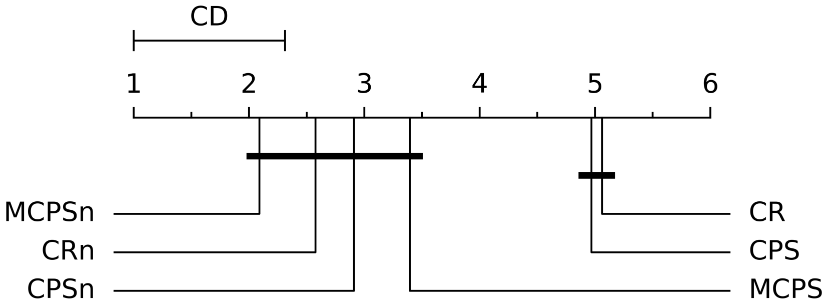


Figure 4: Median efficiency ranks

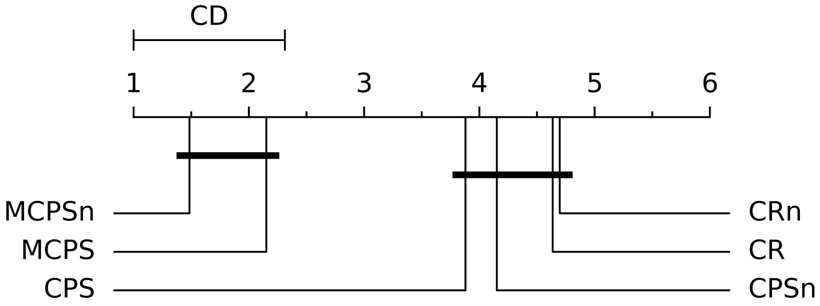


Figure 5: Mean absolute error ranks

5. Concluding remarks

We have introduced a new approach to forming conformal predictive distributions, called Mondrian conformal predictive systems, which generate one predictive distribution for each available Mondrian category. We have shown that by using the predictions of an underlying regression forest to form the categories through binning, the use of the resulting Mondrian predictive distributions will not only improve upon using standard and normalized conformal predictive distributions, but also result in as tight prediction intervals as produced by normalized conformal regressors, and even improve upon the point predictions of the underlying forest with respect to the mean absolute error.

Directions for future work include investigating more sophisticated approaches to forming the categories, e.g., using clustering as an alternative to binning as proposed in (Zhou et al., 2014). Such approaches may not only use the predictions of the underlying model but also the quality estimates. Currently, these estimates have only been considered for normalizing (rescaling) the predictive distribution of each category. The proposed framework also needs to be applied and evaluated in conjunction with other types of underlying model. Alternative ways of forming the categories may be required to allow for similar performance improvements as for the considered regression forests. Another important direction for future research includes investigating the use of the distributions for decision making, possibly

Table 6: Mean squared errors

	CR	CRn	CPS	CPSn	MCPS	MCPSn
abalone	.006	.006	.006	.006	.006	.006
anacalt	.001	.001	.001	.001	.001	.001
bank8fh	.009	.009	.010	.010	.009	.009
bank8fm	.003	.003	.003	.004	.002	.002
bank8nh	.013	.013	.014	.015	.013	.013
bank8nm	.003	.003	.003	.004	.002	.002
boston	.005	.005	.005	.006	.005	.005
comp	.001	.001	.001	.001	.001	.001
concrete	.005	.005	.005	.005	.004	.004
cooling	.002	.002	.002	.002	.002	.002
deltaA	.001	.001	.001	.001	.002	.002
deltaE	.003	.003	.003	.003	.003	.003
friedm	.005	.006	.006	.006	.003	.003
heating	.000	.000	.000	.000	.000	.000
istanbul	.006	.006	.006	.006	.007	.006
kin8fh	.006	.006	.006	.006	.005	.005
kin8fm	.002	.002	.002	.002	.001	.001
kin8nh	.016	.016	.016	.016	.015	.015
kin8nm	.011	.011	.011	.011	.009	.009
laser	.001	.001	.001	.001	.001	.001
mg	.006	.006	.006	.006	.006	.006
mortage	.000	.000	.000	.000	.000	.000
plastic	.029	.029	.031	.029	.031	.030
puma8fh	.021	.021	.021	.021	.020	.021
puma8fm	.005	.005	.005	.005	.004	.004
puma8nh	.021	.021	.021	.021	.019	.019
puma8nm	.008	.008	.008	.008	.005	.005
quakes	.031	.031	.032	.032	.032	.032
stock	.001	.001	.001	.001	.001	.001
treasury	.000	.000	.000	.000	.000	.000
wineRed	.013	.013	.013	.013	.013	.013
wineWhite	.010	.010	.010	.010	.010	.010
wizmir	.000	.000	.000	.000	.000	.000
Mean	**.007**	**.007**	**.008**	**.008**	**.007**	**.007**
Ranks	**3.27**	**3.21**	**4.58**	**4.45**	**3.00**	**2.48**

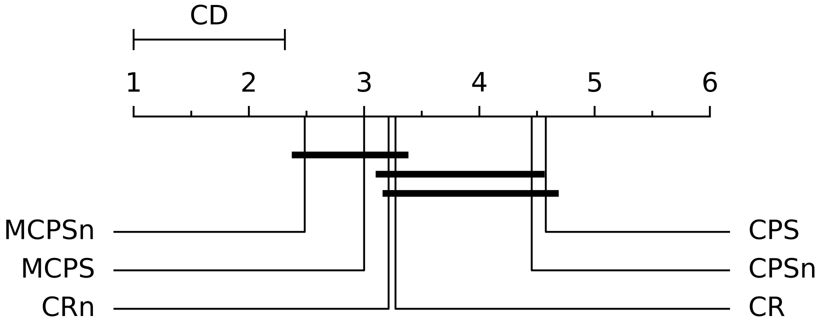


Figure 6: Mean squared error ranks

including performance metrics that are more directly connected to the utility, compared to the metrics employed in this study, e.g., CRPS.

Acknowledgements

HB was partly funded by the Swedish Foundation for Strategic Research (CDA, grant no. BD15-0006). UJ and TL were partly funded by the Swedish Knowledge Foundation (DATAKIND 20190194). The authors would like to thank the anonymous reviewers for their helpful suggestions.

References

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 114–133. PMLR, 09–11 Sep 2020.

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.

Milton 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.

Peter Björn Nemenyi. *Distribution-free multiple comparisons. PhD-thesis.* Princeton University, 1963.

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

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*, volume 1, pages 235–239. IEEE, 12 2013.

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

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Computationally efficient versions of conformal predictive distributions. *Neurocomputing*, 397:292–308, 2020.

Hugo Werner, Lars Carlsson, Ernst Ahlberg, and Henrik Boström. Evaluating different approaches to calibrating conformal predictive systems. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni N. Smirnov, Giovanni Cherubin, and Marco Christini, editors, *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 134–150. PMLR, 2020.

Chenzhe Zhou, Ilia Nouretdinov, Zhiyuan Luo, and Alexander Gammerman. SVM venn machine with k-means clustering. In Lazaros S. Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *Artificial Intelligence Applications and Innovations - AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings*, volume 437 of *IFIP Advances in Information and Communication Technology*, pages 251–260. Springer, 2014.