Mondrian Predictive Systems for Censored Data

Henrik Boström BOSTROMH@KTH.SE
KTH Royal Institute of Technology, Sweden

Henrik Linusson HENRIK@EKKONO.AI
Ekkono Solutions AB, Sweden
KTH Royal Institute of Technology, Sweden

Anders Vesterberg ANDERS.VESTERBERG@SCANIA.COM
Scania CV AB, Sweden

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

Conformal predictive systems output predictions in the form of well-calibrated cumulative distribution functions (conformal predictive distributions). In this paper, we apply conformal predictive systems to the problem of time-to-event prediction, where the conformal predictive distribution for a test object may be used to obtain the expected time until an event occurs, as well as p-values for an event to take place earlier (or later) than some specified time points. Specifically, we target *right-censored* time-to-event prediction tasks, i.e., situations in which the true time-to-event for a particular training example may be unknown due to observation of the example ending before any event occurs. By leveraging the Kaplan-Meier estimator, we develop a procedure for constructing Mondrian predictive systems that are able to produce well-calibrated cumulative distribution functions for right-censored time-to-event prediction tasks. We show that the proposed procedure is guaranteed to produce conservatively valid predictive distributions, and provide empirical support using simulated censoring on benchmark data. The proposed approach is contrasted with established techniques for survival analysis, including random survival forests and censored quantile regression forests, using both synthetic and non-synthetic censoring.

Keywords: Conformal predictive systems, Mondrian predictive systems, right-censored data, time-to-event prediction, survival analysis

1. Introduction

In contrast to conformal regressors (Vovk et al., 2005; Papadopoulos et al., 2011; Lei et al., 2018; Boström and Johansson, 2020), that for a test object output a prediction interval which will include the true target label with some specified probability (confidence level), conformal predictive systems (Vovk et al., 2020; Boström et al., 2021; Vovk, 2022) output a cumulative distribution function for the target label. Conformal predictive systems are clearly more general than conformal regressors in that they allow for extracting multiple prediction intervals for different confidence levels as well as one-sided intervals. The latter is important when we want to guarantee (at some level of confidence) that the target does not exceed (or is less than) a certain threshold value. One-sided intervals also allow for calibrating the predictions of the underlying model, e.g., by extracting medians from the cumulative distribution functions. A central property of conformal predictive systems is that they output cumulative distribution functions that are well-calibrated, i.e., the p-values for the true targets are uniformly distributed. From a practical viewpoint, the predictions should also be efficient, e.g., extracted intervals should be as tight as possible.

An area of application in which predicting with confidence is particularly important is time-to-event (TTE) prediction, i.e., forecasting when an event of interest, e.g., vehicle breakdown, will occur. Having access to a well-calibrated probability for the event to occur within a certain time frame allows for decision-making based on cost-benefit calculations, e.g., balancing the cost of undertaking early maintenance against the cost of a failure on the road. In case we have obtained a sample of data from some fixed but unknown underlying distribution, for which the true time-to-event is known, conformal predictive systems can be readily applied. However, in many cases, the events may not (yet) have occurred for all of the examples, which means that the time-to-event is not always known, e.g., breakdown may have occurred only for some but not all vehicles in a fleet. Quick fixes include removing examples for which the event has not occurred or adjusting such examples by imputing target values, in both cases violating the assumption that the examples are sampled IID. Hence, after applying such fixes, the conformal predictive systems are no longer guaranteed to output well-calibrated cumulative distributions for the true targets.

Even if the event times are not known for all cases, we still often have access to the latest known time point before which an event has not (yet) occurred, e.g., the most recent time point at which communication with a vehicle in a fleet occurred. In such cases, each example can be represented by three components; the object x, a time point y, and an indicator e of whether an event occurred at the time point ($e = 1$) or if this is the latest known time point at which an event has not yet occurred, i.e., the time point represents a right-censored value ($e = 0$). This is the standard scenario within the area of survival analysis (Klein and Moeschberger, 2003), for which many different algorithms have been proposed over the years, including the Kaplan-Meier estimator (Kaplan and Meier, 1958), the Cox model (Cox, 1972), Random Survival Forests (Ishwaran et al., 2008) and more recently, Censored Quantile Regression Forests (Li and Bradic, 2020). These techniques often come with asymptotic guarantees under various assumptions, the most common being that there for each object x exists an actual event time e and an independent censoring time c, such that $y = t$ and $e = 1$, if $c \geq t$, and $y = c$ and $e = 0$, otherwise. To the best of our knowledge, the techniques have, however, not been shown to produce well-calibrated distributions from finite samples.

In this work, we propose a modification of conformal predictive systems to produce well-calibrated predictions also for right-censored data, i.e., the p-value for the true event time is uniformly distributed. This modification is obtained through generating Mondrian predictive systems (Boström et al., 2021; Vovk, 2022), where the predictive distribution output for each category is formed using a Kaplan-Meier estimator.

In the next section, we describe the approach to constructing Mondrian predictive systems from right-censored data and present a theoretical analysis showing the validity of the approach. In Section 3, we empirically investigate the validity by considering a dataset for which censoring has been introduced synthetically and also compare the performance of the new approach to the standard Kaplan-Meier estimator, Random Survival Forests and Censored Quantile Regression Forests. The methods are also evaluated on the original dataset, i.e., with non-synthetic censoring. Finally, in Section 4, we summarize the main findings of this study and outline directions for future work.

2. Generating Conformal Predictive Systems from Censored Data

In this section, we describe our proposed method for generating predictive distributions by combining the conformal predictive system procedure with the Kaplan-Meier estimator. In Section 2.2, we provide a theoretical analysis of the proposed procedure, showing that we expect it to generate valid, i.e., well-calibrated, predictive distributions for test objects with unknown labels.

2.1. The proposed approach

Our approach for constructing a conformal predictive system for censored data is, at its core, based on the split conformal predictive system introduced by Vovk et al. (2020); specifically, we rely on the Mondrian version of split conformal predictive systems described by Boström et al. (2021). For a detailed explanation of ordinary split conformal predictive systems unsuitable for censored training data, we refer the reader to these earlier works.

To create a Mondrian split conformal predictive system that takes into account censoring, we first assume that we are given two sets of information: a set of historical data $Z \subset \boldsymbol{Z}$, which contains an arbitrary number of observations for which the true target value y is censored; and, a map $e : Z \to \{0, 1\}$ such that $e(\boldsymbol{z}_i) = 1$ exactly when $\boldsymbol{z}_i \in Z$ is an event, i.e., a non-censored observation. Given these two sets, we construct a Mondrian conformal predictive system as follows:

1. The data set Z is divided into two disjoint subsets: a proper training set Z^t and a calibration set Z^c, such that $|Z^c| = q$. We denote $Z^c = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_q, y_q)\}$. Analogously, we divide e into e^t and e^c.

2. A predictive regression model h is trained on the proper training set Z^t. It is worth noting that this stage can voluntarily be restricted to the subset $\left\{\boldsymbol{z}_i \in Z^t : e(\boldsymbol{z}_i) = 1\right\}$ (or any other subset of Z^t) without loss of validity.

3. The predictive model h is applied to the calibration set Z^c, giving a set of predictions $\hat{Y}^c = \{\hat{y}_i = h(\boldsymbol{x}_i) : (\boldsymbol{x}_i, y_i) \in Z^c\}$. Without loss of generality, we can reorder the elements of $\hat{Y}$ meaning that $\forall \hat{y}_a, \hat{y}_{a+1} \in \hat{Y}^c : \hat{y}_a \leq y_{a+1}$. We analogously reorder e^c to maintain congruence with $\hat{Y}^c$.

4. $\hat{Y}^c$ is divided into $p << q$ equal-depth bins $\hat{Y}^c_{\kappa_1}, \dots, \hat{Y}^c_{\kappa_p}$, where p is a user-specified parameter. When $p \nmid q$, we let each of the first $q \mod p$ bins contain one additional element. It should be noted that order is preserved through this binning, such that $\forall \hat{y}_a \in \hat{Y}^c_{\kappa_n} : \forall \hat{y}_b \in \hat{Y}^c_{\kappa_{n+1}} : \hat{y}_a \leq \hat{y}_b$. The greatest elements $\tau_1, \dots, \tau_{p-1} = \max(\hat{Y}^c_{\kappa_1}), \dots, \max(\hat{Y}^c_{\kappa_{p-1}})$ give the threshold values that determine the bin widths. The set $K = \{\kappa_1, \dots, \kappa_p\}$ is the Mondrian category space, and we let the map $C : Z \to K$ be the Mondrian categorization function that maps each object $\boldsymbol{z}_i \in \boldsymbol{Z}$ to its appropriate category (bin) κ_m determined by the prediction $h(\boldsymbol{x}_i)$ and the threshold values $\tau_1, \dots, \tau_{p-1}$.

5. For each category κ_n, we use the Kaplan-Meier estimator to compute a survival function S_{κ_n}
$$S_{\kappa_n}(t) = \prod_{k=0}^{t} 1 - \frac{\left|\left\{y_j \in Y^c_{\kappa_n} : e(\boldsymbol{z}_j) = 1 \wedge y_j = y_k\right\}\right|}{\left|\left\{y_j \in Y^c_{\kappa_n} : y_j \geq y_k\right\}\right| + 1}, \tag{1}$$
where $\forall y_j \in Y^c_{\kappa_n} : y_j > y_0$. Note that we are constructing the survival function over the set of true labels $Y^c_{\kappa_n}$ rather than the predicted labels $\hat{Y}^c_{\kappa_n}$. Here, the role of the predicted labels $\hat{Y}^c_{\kappa_n}$ is merely to determine the membership of the category κ_n.

When deciding the output for a test object $\boldsymbol{x}_{q+1}$, we first obtain the prediction $\hat{y}_{q+1} = h(\boldsymbol{x}_{q+1})$ and use that prediction to obtain the appropriate category $\kappa_l = C(\boldsymbol{z}_{q+1})$. Finally the survival function S_{κ_l} is output as the survival function of $\boldsymbol{x}_{q+1}$.

2.2. Theoretical analysis

As with any conformal prediction procedure, validity of a conformal predictive system means that the p-values for the true target labels are uniformly distributed. Generally, such p-values are computed based on a so-called nonconformity measure that is usually represented by an error function on the predictions made by h. For conformal predictive systems, the following nonconformity measure is a common choice:

$$\alpha_i = \frac{y_i - h(\boldsymbol{x}_i)}{\sigma_i}, \tag{2}$$

where σ_i is an estimate of the quality (difficulty) of the prediction.

Using this nonconformity measure, we can obtain a set of nonconformity scores for the calibration set, i.e., $A = \left\{\frac{y_j - h(\boldsymbol{x}_j)}{\sigma_j} : (\boldsymbol{x}_j, y_j) \in Z^c\right\} = \{\alpha_1, \ldots, \alpha_q\}$. Given a test object $\boldsymbol{x}_{q+1}$, we can postulate a label $\tilde{y}$, compute a nonconformity score $\alpha^{\tilde{y}}_{q+1}$ and finally compute a p-value for the postulated label by:

$$p^{\tilde{y}}_{q+1} = \frac{\left|\left\{\alpha_j \in A : \alpha_j > \alpha^{\tilde{y}}_{q+1}\right\}\right| + \theta\left|\left\{\alpha_j \in A : \alpha_j = \alpha^{\tilde{y}}_{q+1}\right\}\right| + 1}{q+1}, \tag{3}$$

where $\theta \sim U[0,1]$ is used to uniformly randomly break ties. Importantly, $p^{\tilde{y}}_{q+1} \sim U[0,1]$ whenever $\tilde{y} = y_{q+1}$, allowing for an exact false rejection rate when $p^{\tilde{y}}_{q+1}$ is used to reject the hypothesized label $\tilde{y}$.

In our construction of a Mondrian predictive system for censored data, the nonconformity measure is implicit. For each Mondrian category, we (again, implicitly) let $h(\boldsymbol{x}_i) = 0$ and $\sigma_i = 1$ for any $\boldsymbol{x}_i$, which means that the nonconformity measure becomes

$$\alpha_i = y_i\,. \tag{4}$$

We let $h(\boldsymbol{x}_i)$ and σ_i be constant in order to avoid violating the assumption made by the Kaplan-Meier estimator, i.e., the modeled survival-time variable (here α) and the censoring time are generated independently. If we place no constraints on h and σ, then α_i may no longer be independent of the censoring time, even if independence holds with respect to y_i.

To see this, consider a model h that has been fitted using data with right-censored labels, i.e., the labels are either true event times or censoring times. Due to the right-censoring, true event times are less frequently observed for instances with low censoring times, and consequently the model can be expected to more frequently underestimate the true target for such instances. Hence, the residuals $(y_i - h(\boldsymbol{x}_i))$ can in such a case be expected to correlate with the censoring time. If the model on the other hand is trained using non-censored data only, the absolute residuals for the true targets for censored data can be expected to be larger, again leading to a possible correlation between the size of the residuals and censoring time. However, even if a model is found such that the residuals indeed are independent of the censoring time, there may still be a correlation between the latter and the difficulty estimate, again leading to the assumption of independence being violated. Keeping both $h(\boldsymbol{x}_i)$ and σ_i constant is a rather crude way of avoiding the problem, at some obvious cost of flexibility. We will later discuss relaxing these constraints.

Further, we somewhat loosen the requirement on the generated p-values, requiring them to allow for conservative rather than exact false rejection rates. In effect, this means that we expect to generate p-values $\bar{p}^{\tilde{y}}_{q+1}$ such that $\bar{p}^{\tilde{y}}_{q+1} \geq p^{\tilde{y}}_{q+1}$. In an ordinary conformal predictive system, this behavior can be enforced by simply not breaking ties, i.e.,

$$\bar{p}^{\tilde{y}}_{q+1} = \frac{\left|\left\{\alpha_j \in A : \alpha_j \geq \alpha^{\tilde{y}}_{q+1}\right\}\right| + 1}{q+1}. \tag{5}$$

Our main goal, then, is to illustrate how we are able to transform this conservative p-value generating method into a still-valid procedure that uses the Kaplan-Meier estimator. Specifically, we show that the p-values generated by our proposed procedure are bounded from below by a process generating uniformly distributed p-values, resulting in conservative validity. As a first step, we simply flip the sign of the comparison in the numerator and restate the p-value generating function as

$$\bar{p}^{\tilde{y}}_{q+1} = 1 - \frac{\left|\left\{\alpha_j \in A : \alpha_j < \alpha^{\tilde{y}}_{q+1}\right\}\right|}{q+1}. \tag{6}$$

We note that the numerator becomes the rank of the test object's nonconformity score $\alpha^{\tilde{y}}_{q+1}$ given the set of nonconformity scores $A^+ = A \cup \{\alpha^{\tilde{y}}_{q+1}\}$, and the quotient represents the probability $P(A \leq \alpha_j)$. If we ensure that the calibration scores in A are ordered such that $\alpha_1, \ldots, \alpha_n$, we can refactor this expression further by first evaluating the probability mass of each calibration score $\alpha_k \in A$ and subsequently computing the cumulative probability of terms with a rank lower than the test object's nonconformity score. Thus, we can compute the p-value as

$$\bar{p}^{\tilde{y}}_{q+1} = \prod_{k=0}^{\pi(\alpha^{\tilde{y}}_{q+1}, A^+)} 1 - \frac{|\{\alpha_j \in A : \alpha_j = \alpha_k\}|}{|\{\alpha_j \in A : \alpha_j \geq \alpha_k\}| + 1}, \tag{7}$$

where $\forall \alpha_j \in A : \alpha_j > \alpha_0$ and $\pi(\alpha^{\tilde{y}}_{q+1}, A^+)$ is the rank of $\alpha^{\tilde{y}}_{q+1}$ relative to A^+. This produces, for each test object, a p-value identical to that given by Equation (6). It is worth noting that the numerator of the fraction is always at least 1 and, in the case where all calibration scores are unique, the numerator is always exactly 1. Finally, in order to obtain the Kaplan-Meier estimator, we simply adjust the numerator of the fraction as

$$\bar{p}^{\tilde{y}}_{q+1} = \prod_{k=0}^{\pi(\alpha^{\tilde{y}}_{q+1}, A^+)} 1 - \frac{|\{\alpha_j \in A : e(\boldsymbol{z}_j) = 1 \wedge \alpha_j = \alpha_k\}|}{|\{\alpha_j \in A : \alpha_j \geq \alpha_k\}| + 1}, \tag{8}$$

where, again, $\forall \alpha_j \in A : \alpha_j > \alpha_0$ and $\pi(\alpha^{\tilde{y}}_{q+1}, A^+)$ is the rank of $\alpha^{\tilde{y}}_{q+1}$. Put in terms commonly used to describe the Kaplan-Meier estimator, the numerator is the number of (non-censored) events that occur at time k and the denominator is the number of objects at risk at time k.

Comparing Equations (7) and (8), we note that the latter is bounded from below by the former; since $\{\alpha_j \in A : e(\boldsymbol{z}_j) = 1\} \subseteq A$, the quotient in Equation (8) is smaller than that in Equation (7), resulting in the multiplied terms, and the final p-value, being larger. Since the computed p-values can only grow, thus reducing the (true and false) rejection rate, a conformal predictive system based on a Kaplan-Meier estimator constructed in this way should be considered conservatively valid.

3. Empirical investigation

In this section, we empirically investigate the predictive distributions output by the proposed adaptation of Mondrian predictive systems as well as by a selection of approaches for survival analysis; the standard Kaplan-Meier estimator, random survival forests, and censored quantile regression forests. We first describe the experimental setup and then present the experimental results.

3.1. Experimental setup

In order to investigate whether the predictive distributions are well-calibrated or not, we need to have access to the true target labels, which by definition, is not the case for real-world censored data. We opt for using a real-world dataset and introducing censoring synthetically to allow for such an investigation. The learning algorithms will have access to the true target labels only for non-censored data during training as usual but will be requested to provide p-values for the true target labels for the test set. In principle, any regression dataset could be used for this purpose, but we will here consider a publicly available survival dataset concerning the relationship between serum free light chain (FLC) and mortality.[1]

The FLC dataset consists of 7874 instances and the following eight features, which will be used for predicting time-to-event (time to death); age, sex, the calendar year in which a blood sample was obtained, kappa portion of the FLC, lambda portion of the FLC, the FLC group for the subject, serum creatinine, and if the subject had been diagnosed with monoclonal gammopathy. Missing values for the creatinine level have been replaced by the mean of the non-missing values. Days from enrollment will here be used as the

[1] For details about the origin of the dataset, see `https://stat.ethz.ch/R-manual/R-devel/library/survival/html/flchain.html`. For this study, the dataset was obtained via the Python package `SurvSet` (Drysdale, 2022).

target, independently of whether the event took place or not, and synthetic censoring times are obtained by randomly permuting these targets, resulting in approximately half of the observations being censored, i.e., the target is less than the censoring time.

It should be stressed that for the purpose of this investigation, the original censored data cannot be directly used, as it will not allow for determining if the p-values for true target labels are uniformly distributed or not. In the original dataset, we only have access to the true labels for a subset of the test instances, and since this subset cannot be assumed to be randomly selected, we should not expect the p-values to be uniformly distributed for this group. The experiment will provide an illustration that the p-values for the true labels may indeed be distributed differently for the censored and non-censored observations, respectively.

We will also investigate a scenario where we want to make time-to-event predictions with confidence; we request each model to provide a time point, such that an event will not occur before that with high probability. This could be relevant in situations where we need to take some action before an adverse event occurs, e.g., undertake maintenance before a machine breaks. We will consider three confidence levels; 90%, 95% and 99%, respectively, and we report the mean and median time points as well as the error rates, i.e., the relative frequency of events occurring before the output time point. Ideally, a model should output as late time points as possible, e.g., to not constrain a scheduler or request components to be changed earlier than necessary, while at the same time making sure that the error rate does not exceed what is allowed by the specified confidence level.

In addition to presenting results from data with synthetic censoring, we will also report results for the original (non-synthetic) censored labels, i.e., where only the event or the censoring time is known for each instance, according to what is indicated for the original FLC dataset. As discussed above, there is no point in investigating the distribution of p-values in this case, since the true labels are not known for all test instances, but instead, we compare the models with respect to the ranking performance, as measured by Harrell's concordance index (C-index) (Harrell et al., 1982).

In the experiment, the FLC dataset is randomly split into two parts; 75% for training and 25% for testing.

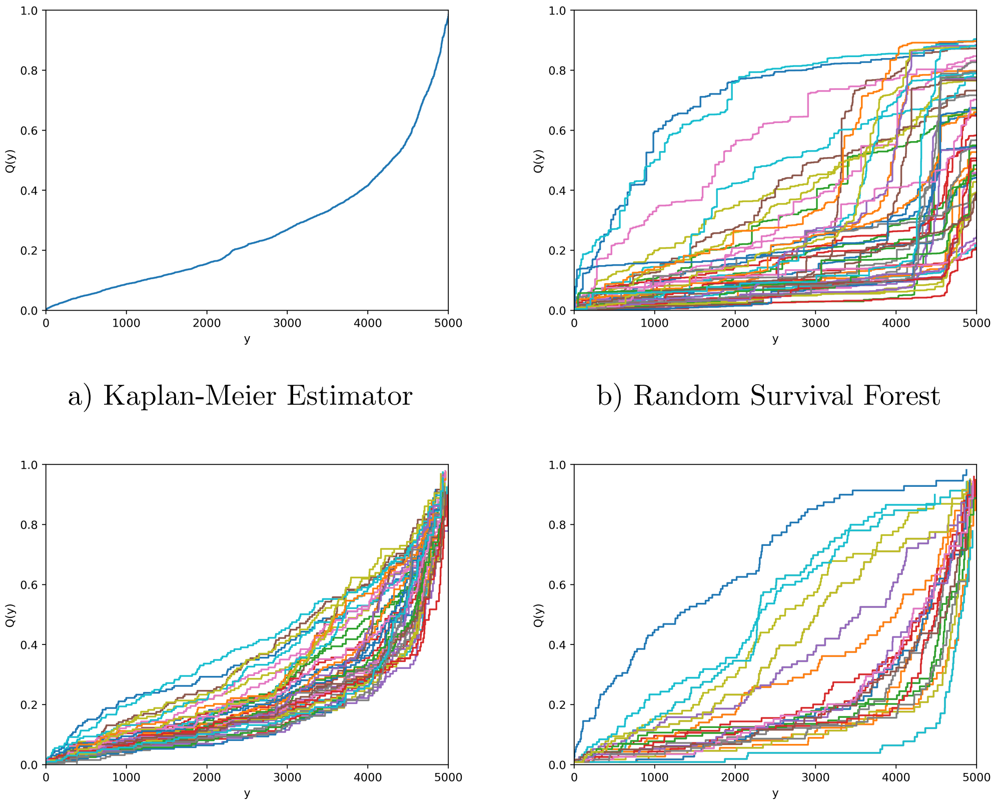
We will compare the following approaches:

- The Kaplan-Meier estimator (KME) (Kaplan and Meier, 1958); no hyperparameters
- Random Survival Forests (RSF)[2] (Ishwaran et al., 2008); default settings for all hyperparameters except for `n_estimators = 500`
- Censored Quantile Regression Forests (QRF)[3] (Li and Bradic, 2020); default settings for the individual regression trees, `n_estimators = 500` and `k = 200` (no. of neighbors)
- The proposed approach, employing Conformal Predictive Systems (CPS)[4]; the underlying model is a `RandomForestRegressor` generated with default settings for all

[2] As implemented in the Python package `sksurv` (Pölsterl, 2020).

[3] Using a re-implementation of quantile regression forests (Meinshausen and Ridgeway, 2006) using `DecisionTreeRegressor` of the Python package `sklearn` (Pedregosa et al., 2011).

[4] Implemented using the Python package `crepes` (Boström, 2022).



a) Kaplan-Meier Estimator

b) Random Survival Forest

c) Quantile Regression Forest

d) Conformal Predictive System

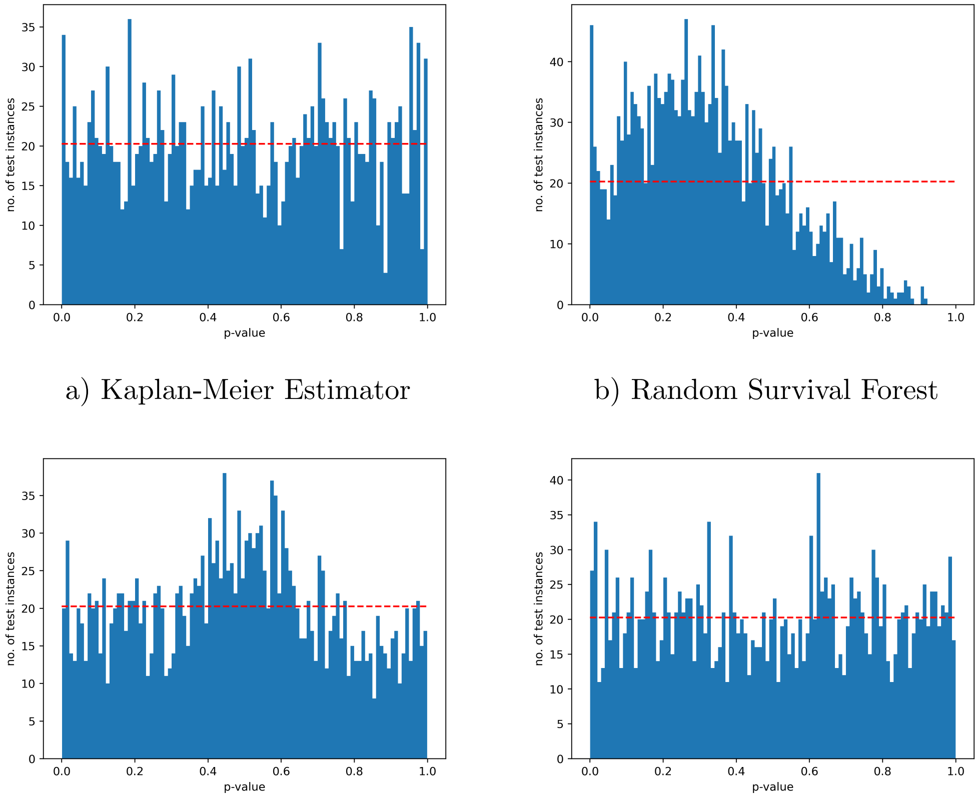
Figure 1: Cumulative distributions for 50 test instances

hyperparameters except for `n_estimators = 500`, the training set is randomly split into a proper training set and a calibration set of equal size, Mondrian categories are formed by binning predictions for the calibration set into 25 bins.

3.2. Experimental results

We start out by illustrating the predictive distributions produced by the different approaches; in Fig. 1, we show the distributions output by each approach for the same 50 random test instances. As can be seen in Fig. 1a, the Kaplan-Meier estimator outputs the same distribution for all test instances. The shape of the distributions produced by the censored quantile regression forest, displayed in Fig. 1c, are more similar to the former than to the distributions output by the random survival forest (Fig. 1b) and the Mondrian conformal predictive system (Fig. 1d). Note that the latter will only output up to 25 unique predictive distributions; one for each Mondrian category.

In Fig. 2, histograms for the output p-values for the true target labels for the test instances are presented. The red dashed line in each plot shows the expected number of observations in each bin if the p-values are uniformly distributed. The most striking pattern



a) Kaplan-Meier Estimator

b) Random Survival Forest

c) Quantile Regression Forest

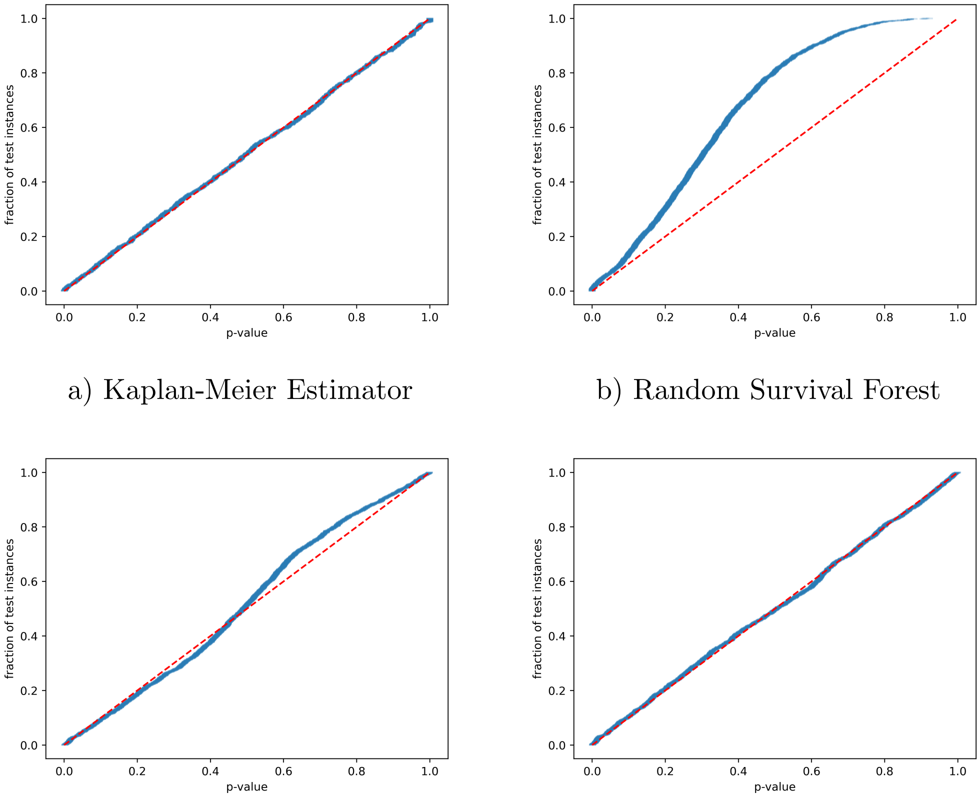
d) Conformal Predictive System

Figure 2: Distribution of p-values

here is that the p-values output by the random survival forest (Fig. 2b) are clearly not uniform; far too many low p-values are output. Also, the p-values output by the censored quantile regression forest (Fig. 2b) exhibit a pattern; the more frequent p-values are located between 0.4 and 0.6.

Another view of the p-value distributions is given in Fig. 3, where cumulative distributions instead are presented. The red dashed line in each plot again shows what should be expected from a uniform distribution. It is clear that the p-values output by the random survival forest and censored quantile regression forest (Fig. 2b and c, respectively) are non-uniform. In contrast, when inspecting the graphs in Fig. 2a and d, both the Kaplan-Meier estimator and the proposed approach appear to be valid, i.e., they produce uniformly distributed p-values for the true target labels.

To provide more quantitative support for the above conclusions, we applied the Kolmogorov-Smirnov test to each set of p-values, with the null hypothesis that they come from a uniform distribution. As can be seen in the last column of Table 1, we can safely reject the null hypotheses for the random survival forests and censored quantile regression forests (the corresponding p-values are underlined in the table), while this cannot be done for the two other approaches, for any reasonable level of confidence.



a) Kaplan-Meier Estimator

b) Random Survival Forest

c) Quantile Regression Forest

d) Conformal Predictive System

Figure 3: Cumulative distribution of p-values

Fig. 4 illustrates that a uniform distribution of p-values for the true targets does not necessarily imply that the p-values are uniformly distributed for the censored and non-censored instances, respectively. In Fig. 4a, the p-values for the true targets of the non-censored instances are shown for the Kaplan-Meier estimator, indicating that the non-censored instances generally receive lower p-values as compared to the censored test instances (Fig. 4b).

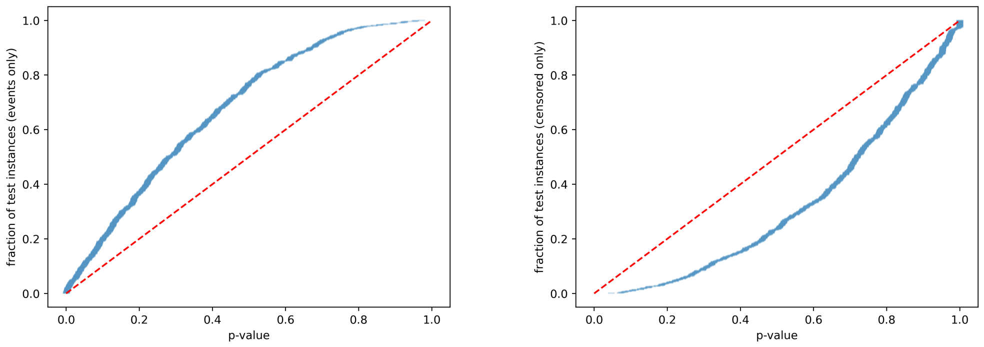
In the second and third columns of Table 1, the training and testing (inference) times (in seconds) are shown for the different approaches.[5] These results indicate that the computational cost can be quite substantial for the random survival forests even for limited-sized datasets, in contrast to the other approaches.

The results for time-to-event prediction with confidence are shown in Tables 2a-c. We can see that although the random survival forest provides the least constraining time points (indicated with bold font), which is desirable, the error rates are clearly higher than what is requested (an error rate exceeding the specified rate with more than 10% is indicated by an underlined value in the tables). The error rates of the other three methods are

[5] The experiment was executed on an HP Z-book 15, with an i9-11950H CPU (8 cores) and 32 GiB primary memory, running Ubuntu 22.04.

	Training time	Testing time	KS-test
KME	0.00	0.02	7.70e-01
RSF	43.37	4.49	6.95e-174
QRF	1.39	0.82	2.33e-09
CPS	0.56	0.07	3.06e-01

Table 1: Run times (seconds) and p-values from testing uniformity with the KS-test



a) Non-censored instances

b) Censored instances

Figure 4: Cumulative distribution of p-values for the Kaplan-Meier Estimator

more reasonable, with both the quantile regression forest and conformal predictive systems being slightly conservative; they result in slightly fewer errors than allowed. Of the three approaches with acceptable error rates, CPS is outperforming the other two for all three confidence levels in terms of the least constraining output time point on average.

We have also investigated using the original (censored) labels for the dataset, with all other settings, including how the dataset is split into training and test sets, being the same as in the previous experiment. In Table 3, we show Harrell's concordance index (the C-index) for all approaches on the test set. This is a measure of the ranking performance, considering all comparable pairs in the test set which are correctly ordered with respect to the expected time-to-event, as estimated by each approach. Since KME predicts the same expected time-to-event for all instances, the corresponding C-index is no better than random. We can see that there is a clear order in performance with respect to this metric, with random survival forests excelling before CPS and QRF. In addition, the table also shows the average predicted time-to-event for all methods, as well as the computation time for the modified training and test sets. Most notable here is that the training time for the random survival forests is substantially reduced, which can be explained by the changed censoring rate; about 72% of the instances are censored with the original censoring, compared to about 50% for the synthetic censoring.

	Error	Mean time point	Median time point
KME	0.0153	48.0	48
RSF	0.0390	**402.0**	**186**
QRF	0.0039	12.3	5
CPS	0.0094	78.6	14

a) 99% confidence

	Error	Mean time point	Median time point
KME	0.0523	487.0	487
RSF	0.0676	**1344.6**	**1110**
QRF	0.0444	510.5	461
CPS	0.0449	710.9	532

b) 95% confidence

	Error	Mean time point	Median time point
KME	0.1027	1195.0	1195
RSF	0.1362	**2318.3**	**2165**
QRF	0.0893	1352.8	1370
CPS	0.0977	1621.7	1764

c) 90% confidence

Table 2: Empirical error rates, mean and median time points for three confidence levels

	C-index	Mean TTE	Median TTE	Training time	Testing time
KME	0.5000	4998.0	4998	0.00	0.00
RSF	**0.7234**	4554.6	4998	36.58	2.83
QRF	0.5804	4655.4	4715	1.38	0.56
CPS	0.6379	4241.7	4486	0.58	0.05

Table 3: Results for non-synthetic censoring

4. Concluding remarks

We have presented a procedure for constructing Mondrian predictive systems that are able to produce well-calibrated cumulative distribution functions for right-censored time-to-event prediction tasks, by forming a Kaplan-Meier estimator for each Mondrian category. We have shown that the proposed procedure is guaranteed to produce conservatively valid predictive distributions, and have also provided empirical support for this using simulated censoring on benchmark data. The empirical investigation showed that two established techniques for survival analysis, random survival forests, and censored quantile regression forests, do not share this property, making the former exceed specified error levels for quantile time-to-event prediction, and the latter producing overly conservative predictions, in contrast to the proposed approach.

There are several possible directions for future work. The most obvious and important direction concerns relaxing the constraints of the proposed approach, i.e., the prediction and the difficulty estimate within each Mondrian category are set to be constant. These constraints effectively lead to that the resulting Mondrian predictive system can be defined by a set of Kaplan-Meier estimators. Research in this direction includes investigating conditions under which an informative difficulty estimate can be employed while still producing well-calibrated predictive distributions. Another direction for future research is to investigate the use of conformal predictive systems in real-world scenarios where censored data is frequently occurring, e.g., to support maintenance planning based on well-calibrated predictions of time-to-event.

Acknowledgements

HB and HL were partly funded by Vinnova (RAPIDS, grant no. 2021-02522).

References

Henrik Boström. crepes: a python package for generating conformal regressors and predictive systems. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction and Applications*, volume 179 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, pages 114–133, 2020.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 24–38. PMLR, 08–10 Sep 2021.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Erik Drysdale. SurvSet: An open-source time-to-event dataset repository. *arXiv preprint arXiv:2203.03094*, 2022.

F. Harrell, R. Califf, D. Pryor, K. Lee, and R. Rosati. Evaluating the Yield of Medical Tests. *JAMA: The Journal of the American Medical Association*, 247(18):2543–2546, May 1982.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *Ann. Appl. Statist.*, 2(3):841–860, 2008.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Alexander Hanbo Li and Jelena Bradic. Censored quantile regression forest. In *International Conference on Artificial Intelligence and Statistics*, pages 2109–2119. PMLR, 2020.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Vladimir Vovk. Universal predictive systems. *Pattern Recognition*, 126:108536, 2022.

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.