Anytime Valid Conformal Prediction

Eszter Barabás, ELTE TTK

Supervisor: Balázs Csanád Csáji, ELTE TTK & HUN-REN SZTAKI

1. Abstract

This semester, I continue to work on the topic I began in the previous semester, exploring more advanced concepts and methods. Conformal prediction is a powerful framework for uncertainty quantification via the production of prediction sets containing the true data with a preset probability. I have already defined conformal predictors using p-values, where the theoretical foundations were based on [1]. In this report, I explore an alternative approach with e-values, known as conformal e-prediction.

2. Introduction

Conformal prediction was originally developed using p-values. Although p-values remain standard, they have been widely debated and criticized, motivating different methods.P-hacking has become widespread in research, leading to the selective reporting of results and the distortion of true effect sizes in published studies. Common practices include conducting analyses midway through experiments to decide whether to continue collecting data, deciding whether to include or not outliers post analysis or stopping data exploration once a significant p-value is achieved. However, choosing paths that produce significant results can inflate the actual Type I error rate. [6]

Although e-values have appeared relatively late in research, they have emerged as one of the most prominent alternatives. Both e-variables and p-variables are test statistics of the data, but with entirely different properties. Even though they are closely related and, in fact, exist under essentially the same conditions, a connection whose similarities we will discuss further. However, e-values have several important advantages, such as being suitable for unknown sampling schemes. If the data collection procedure is not specified, we may not be able to calculate the p-value correctly, but we are often able to calculate the e-value for a wide range of sampling schemes. E-values allow for post-hoc decision making; it is possible to select the significance level α in a data-dependent way while still maintaining nontrivial guarantees. Suppose a scientist collects some data but stays indecisive to reject a null hypothesis. Seeing this result, another scientist may run their own experiment, and their e-values can be merged by simple multiplication. However, we cannot say the same for p-values because the latter studies may depend on the p-value obtained from the previous experiment, which is also related to p-hacking. E-values remain valid in sequential use, even if we keep collecting data or stop at any time. Anytime validity is central in this report, we are going to show that e-values can be built as a supermartingale and also run some experiments. To highlight their practical relevance, e-values can be applied effectively in adaptive procedures, online learning, and sequential experiments. [2] [7]

3. E-values and p-values

In the following section I used the definitions found in [2]. We begin with a sample space Ω equipped with a σ-algebra $\mathcal{F}$, and the set $\mathcal{M}$ of all probability measures on $(\Omega, \mathcal{F})$, whose elements are the distributions.

We assume that our data $X = (X_1, \ldots, X_n)$ are described by some distribution $P_0 \in \mathcal{M}$. The variables $X_1, \ldots, X_n$ may be independent and identically distributed under P_0, but do not necessarily have to be.

Definition 1. *A hypothesis is a set of probability measures in $\mathcal{M}$. A hypothesis is* simple *if it is a singleton, such as $\{P\}$ and $\{Q\}$. Otherwise, it is* composite.

Definition 2. *An* e-variable *E for $\mathcal{P}$ is a $[0, \infty]$-valued random variable satisfying*

$$\mathbb{E}_P[E] \leq 1 \quad \text{for all } P \in \mathcal{P}.$$

An e-variable E is exact *if*

$$\mathbb{E}_P[E] = 1 \quad \textit{for all } P \in \mathcal{P}.$$

Definition 3. *A* p-variable P *for* $\mathcal{P}$ *is a* $[0,\infty)$*-valued random variable satisfying*

$$P(P \le \alpha) \le \alpha \quad \textit{for all } \alpha \in (0,1) \textit{ and all } P \in \mathcal{P}.$$

A p-variable P is exact *if*

$$P(P \le \alpha) = \alpha \quad \textit{for all } \alpha \in (0,1) \textit{ and all } P \in \mathcal{P}.$$

Proposition 1. *E-variables and p-variables are random variables, while e-values and p-values refer to their realized values after observing the data.*

Definition 4. test φ *is a* $[0,1]$*-valued random variable. A test is* binary *if its range is* $\{0,1\}$*. The type-I error (rate) of a test φ for every $\mathcal{P}$ is* $\mathbb{E}_P[\varphi]$.

A test φ has level $\alpha \in [0,1]$ for $\mathcal{P}$ if its type-I error is at most α for every $P \in \mathcal{P}$.

We denote by $\mathfrak{E} = \mathfrak{E}(\mathcal{P})$ the set of all e-variables for $\mathcal{P}$ and by $\mathfrak{U} = \mathfrak{U}(\mathcal{P})$ the set of all p-variables for $\mathcal{P}$, with $\mathcal{P}$ often omitted.

E-values may be converted into the more classical concepts of level-α tests and p-values. Markov's inequality plays a central role in this conversion, it guarantees that a test rejecting an e-value larger than $1/\alpha$ yields a level-α test.

Proposition 2 (Markov's inequality for e-values)**.** *Let E be an e-variable for $\mathcal{P}$. Then*

$$P\left(E \ge \frac{1}{\alpha}\right) \le \alpha \quad \textit{for all } P \in \mathcal{P} \textit{ and } \alpha \in (0,1]. \tag{1}$$

Hence, $1/E$ is a p-variable, $(\alpha E) \wedge 1$ is a level-α test, and $\mathbf{1}_{\{E \ge 1/\alpha\}}$ is a level-α binary test.

(1) allows for the construction of conformal sets. Conformal e-prediction refers to conformal prediction methods based on e-variables. The term conformal prediction broadly denotes the construction of conformal sets for a test point using a calibration set, regardless of the method.

Remark 1. *For an e-variable E and a p-variable P, although $1/E$ is a p-variable, the reciprocal $1/P$ is in general not an e-variable.*

4. Framework for the prediction

We will describe the setup needed for conformal e-prediction, just like we did with the standard version last semester. The following three sections were written based on [4] source. Suppose we are given a training set $z_1, \dots, z_n$ consisting of labeled objects $z_i = (x_i, y_i)$, and our goal is to predict the label of a new object x.

For each potential label y for x, we would like to have a number $f(z_1, \dots, z_n, x, y)$ reflecting the plausibility of y being the true label of x. The output for every label is thus the following:

$$y \mapsto f(z_1, \dots, z_n, x, y).$$

We can also write $f(z_1, \dots, z_n, z)$,
where $z := (x, y)$, instead of $f(z_1, \dots, z_n, x, y)$.

We use the notation $\mathcal{X}$ for the object space and $\mathcal{Y}$ for the label space (both assumed non-empty) like we did earlier. These are measurable spaces from which the objects and labels. Full observations $z = (x, y)$ are drawn from the observation space

$$\mathcal{Z} := \mathcal{X} \times \mathcal{Y}.$$

For any non-empty set $\mathcal{X}$, let

$$\mathcal{X}^+ := \bigcup_{n=1}^{\infty} \mathcal{X}^n$$

be the set of all non-empty finite sequences of elements of $\mathcal{X}$.

Definition 5. *A nonconformity e-measure is a measurable function*

$$A : \mathcal{Z}^+ \to [0,\infty)^+$$

that maps any finite sequence $(z_1, \dots, z_m)$, $m \in \{1, 2, \dots\}$, to a finite sequence $(\alpha_1, \dots, \alpha_m)$ of the same length consisting of nonnegative numbers (nonconformity scores) with average at most 1:

$$\frac{1}{m} \sum_{i=1}^{m} \alpha_i \le 1, \tag{2}$$

and satisfies the following property: for any $m \ge 2$, any permutation π of $\{1, \dots, m\}$, any $(z_1, \dots, z_m) \in \mathcal{Z}^m$, and any $(\alpha_1, \dots, \alpha_m) \in [0,\infty)^m$,

$$(\alpha_1, \dots, \alpha_m) = A(z_1, \dots, z_m)$$
$$\implies (\alpha_{\pi(1)}, \dots, \alpha_{\pi(m)}) = A(z_{\pi(1)}, \dots, z_{\pi(m)}).$$

In other words, each nonconformity score α_i is unaffected by the ordering of z_i or the other elements in the sequence.

The conformal e-predictor f corresponding to such A is defined by

$$f(z_1, \dots, z_n, x, y) := \alpha_{n+1},$$

where

$$(\alpha_1, \dots, \alpha_n, \alpha_{n+1}) := A(z_1, \dots, z_n, (x, y)).$$

A *conformal e-predictor* is a function obtained from a nonconformity e-measure in the manner described above. Given a training set $z_1, \dots, z_n$ and a test object x, the full prediction for x produced by a conformal e-predictor f is the family of conformal e-values

$$\big(f(z_1, \dots, z_n, x, y) \mid y \in \mathcal{Y}\big). \tag{3}$$

This family of e-values, one for each possible label y, can be viewed as a soft set predictor. By applying a threshold to f, we obtain a (hard) set predictor: we include in the prediction set for the label of x all labels y whose conformal e-value is smaller than a chosen level. However, the threshold does not need to be specified in advance. Instead, the conformal e-value $f(z_1, \dots, z_n, x, y)$ can be interpreted as the degree to which the label y is excluded from the soft prediction set.

Our goal is to construct conformal predictors that are both valid and efficient. Validity means that the true label should not be excluded, while efficiency means that incorrect labels should be excluded whenever possible.

The *full prediction* for the label of x can be summarized in several ways. For example, the point prediction can be defined as

$$\hat{y} \in \arg\min_y f(z_1, \dots, z_n, x, y),$$

assuming the minimum is attained at a unique label.

We call a nonconformity e-measure (and the corresponding conformal e-predictor) *admissible* if equality holds in condition (2),

$$\frac{1}{m}\sum_{i=1}^{m} \alpha_i = 1.$$

We typically restrict our attention to admissible nonconformity e-measures and admissible conformal e-predictors.

A nonnegative nonconformity measure is a measurable function

$$A : \mathcal{Z}^+ \to [0, \infty)^+$$

defined in the same way as a nonconformity e-measure but without imposing condition (2). Given such a function A, we can always construct an admissible nonconformity e-measure A' by normalizing A as follows:

$$A'(z_1, \dots, z_m) := \frac{m}{\sum_{i=1}^{m} \alpha_i}(\alpha_1, \dots, \alpha_m), \tag{4}$$

where $(\alpha_1, \dots, \alpha_m) := A(z_1, \dots, z_m)$. If $A(z_1, \dots, z_m) = (0, \dots, 0)$, we define

$$A'(z_1, \dots, z_m) := (1, \dots, 1)$$

to ensure that A' is admissible.

The corresponding conformal e-predictor is said to be based on A.

Let $Z_1, Z_2, \dots$ denote random elements whose observed realizations are the data points $z_1, z_2, \dots$. More generally, we use (X, Y) or Z to represent random elements taking values in the observation space $\mathcal{Z}$.

Proposition 3. *For any conformal e-predictor f and any n, if $Z_1, \dots, Z_n, (X, Y)$ are IID (or exchangeable), then*

$$\mathbb{E}\, f(Z_1, \dots, Z_n, X, Y) \le 1. \tag{5}$$

Running example

We illustrate the conformal e-prediction framework with a simple binary classification example.

Let the object space be $\mathcal{X} = \mathbb{R}$ and the label space be $\mathcal{Y} = \{A, B\}$. We are given the training data:

$$z_1 = (1, A), \quad z_2 = (2, A), \quad z_3 = (4, B), \quad z_4 = (5, B).$$

We aim to predict the label of a new object $x = 3$.

We define the nonconformity score as the

distance from the class mean:

$$\alpha_i = \begin{cases} |x_i - \mu_A| & \text{if } y_i = A, \\ |x_i - \mu_B| & \text{if } y_i = B, \end{cases}$$

where μ_A and μ_B are the sample means of the classes.

Case 1: new label $y = A$

We augment the dataset with $(3, A)$. Then:

$$\mu_A = \frac{1+2+3}{3} = 2, \quad \mu_B = 4.5.$$

The nonconformity scores are:

$$(\alpha_1, \dots, \alpha_5) = (1, 0, 0.5, 0.5, 1).$$

Their average is 0.6.

To get an admissible e-measure, we normalize:

$$\begin{aligned}(\alpha'_1, \dots, \alpha'_5) &= \frac{5}{3}(\alpha_1, \dots, \alpha_5) \\ &= (1.67, 0, 0.83, 0.83, 1.67).\end{aligned}$$

Thus, the conformal e-value is:

$$f(z_1, \dots, z_4, x, A) = \alpha'_5 = 1.67.$$

Case 2: new label $y = B$

We continue with the same procedure. Then:

$$\mu_A = 1.5, \quad \mu_B = \frac{4+5+3}{3} = 4.$$

The nonconformity scores are:

$$(\alpha_1, \dots, \alpha_5) = (0.5, 0.5, 0, 1, 1).$$

Their average is 0.6.

After normalization:

$$\begin{aligned}(\alpha'_1, \dots, \alpha'_5) &= \frac{5}{3}(\alpha_1, \dots, \alpha_5) \\ &= (0.83, 0.83, 0, 1.67, 1.67).\end{aligned}$$

Thus:

$$f(z_1, \dots, z_4, x, B) = \alpha'_5 = 1.67.$$

The conformal e-values are:

$$f(x, A) = 1.67, \quad f(x, B) = 1.67.$$

Hence, both labels are equally plausible. The point prediction is not determined uniquely:

$$\arg\min_{y \in \mathcal{Y}} f(x, y)$$

contains both A and B.

5. Split conformal e-predictors

As we discussed split conformal prediction the previous semester, we will also define the setup for e-predictors. Let Σ be a measurable space (referred to as a *summary space*). A Σ-valued split nonconformity measure is defined as a measurable function

$$A : \mathcal{Z}^+ \to \Sigma.$$

We can say that the value $A(z_1, \dots, z_m, z)$ quantifies how well the observation z conforms to the sample $z_1, \dots, z_m$.

A *normalizing transformation* is a measurable, equivariant mapping

$$N : \Sigma^+ \to [0, \infty)^+,$$

which assigns to any nonempty finite sequence $(\sigma_1, \dots, \sigma_m)$ in Σ a sequence $(\alpha_1, \dots, \alpha_m)$ of nonnegative real numbers of the same length. These numbers must satisfy the condition that their average does not exceed 1.

To construct a split conformal e-predictor using a training sample $z_1, \dots, z_n$, the data are partitioned into two subsets: a *proper training set* $z_1, \dots, z_{n-c}$ and a *calibration set* $z_{n-c+1}, \dots, z_n$, exactly as we did it with p-predictors.

For a new input object x and a candidate label y, define

$$f(z_1, \dots, z_n, x, y) := \alpha_y,$$

where α_y is obtained through the following procedure. First, compute

$$\sigma_i = A(z_1, \dots, z_{n-c}, z_{n-c+i}), \quad i = 1, \dots, c,$$

and

$$\sigma_y = A(z_1, \dots, z_{n-c}, (x, y)).$$

Next, apply the normalizing transformation

$$(\alpha_1^y, \dots, \alpha_c^y, \alpha_y) = N(\sigma_1, \dots, \sigma_c, \sigma_y).$$

For many choices of the functions A and N, the split conformal e-predictor defined above can be implemented efficiently. The following conditions are satisfied in order to achieve computational efficiency:

1. After processing the proper training set

once, there exists a simple rule allowing $A(z_1, \dots, z_{n-c}, z)$ to be computed for any new observation z.
2. The normalizing transformation N can be evaluated with low computational cost.

Proposition 4. *For any split conformal e-predictor f and any n, if $Z_1, \dots, Z_n, (X, Y)$ are exchangeable, we have* (5)*.*

6. Cross conformal e-predictors

Split conformal e-predictors are typically computationally efficient. However, their predictive efficiency may be lower than that of the full conformal e-predictors. This difference arises because full conformal methods effectively utilize the entire training sample for both training and calibration.

Cross-conformal e-prediction is designed to improve predictive efficiency by combining several split conformal predictors.

A Σ-valued split nonconformity measure A is called a Σ-valued *cross-nonconformity measure* if the value $A(z_1, \dots, z_m, z)$ is invariant to permutations of its first m arguments.

Given such a function A together with a normalizing transformation N, the cross-conformal e-predictor is constructed as the following way:

The training sample $z_1, \dots, z_n$ is randomly partitioned into K non-empty multisets (called *folds*) denoted by z_{S_k} for $k = 1, \dots, K$. These folds have equal (or as equal as possible) sizes. Here $K \in \{2, 3, ...\}$ is a parameter of the procedure. $(S_1, \dots, S_K)$ forms a partition of the index set $\{1, \dots, n\}$, and z_{S_k} contains all observations z_i with indices $i \in S_k$.

For each $k \in \{1, \dots, K\}$ and for every potential label $y \in \mathcal{Y}$ associated with a new object x, compute the value α_k produced by the split conformal e-predictor (based on A and N). The set $z_{S_{-k}}$ is used as the proper training sample and z_{S_k} as the calibration sample, where

$$S_{-k} := \bigcup_{j \neq k} S_j = \{1, \dots, n\} \setminus S_k$$

denotes the complement of S_k.

Finally, the cross-conformal e-predictor is defined as

$$f(z_1, \dots, z_n, x, y) := \frac{1}{K} \sum_{k=1}^{K} \alpha_k.$$

Proposition 5. *For any cross-conformal e-predictor f and any n, if $Z_1, \dots, Z_n, (X, Y)$ are exchangeable, we have* (5)*.*

7. Experimental Setup

We consider a synthetic regression problem designed to compare the two prediction methods stated above. Let $X \sim \text{Uniform}(0, 10)$ and define the response variable as

$$Y = \sin(X) + 0.5X + \varepsilon,$$

where the noise term ε is normally distributed with input-dependent variance:

$$\varepsilon \sim \mathcal{N}\left(0,\ \sigma^2(X)\right), \quad \text{with} \quad \sigma(X) = 0.5 + 0.3X.$$

A dataset of $n = 400$ observations is generated and randomly split into a training set (300 points) and a test set (100 points).

I used a **Random Forest** regression model as the base predictor in the conformal methods.

Then I implemented the two conformal prediction methods: Regarding the split conformal prediction, the training data were further divided into a proper training set and a calibration set. The model was trained on the proper training set, and prediction intervals are constructed using quantiles of absolute residuals computed on the calibration set. The leave-one-out procedure was used for cross conformal prediction, where a model is trained n times, each time excluding one observation.

I also calculated the average interval length, which is important for measuring efficiency.

7.1. Visualization

The following plots help us visualize the comparison of these two methods:

- **Figure 1.:** A bar plot showing the runtime in seconds, showing which method is more efficient computationally. Cross conformal prediction deals with multiple folds, resulting in a longer runtime.

- **Figure 2.:** A bar plot showing the average interval length for each method, illustrating predictive efficiency. The narrower the prediction band, the more effective the method is, because the 90% coverage is guaranteed. Cross conformal prediction seems a better method in this regard.

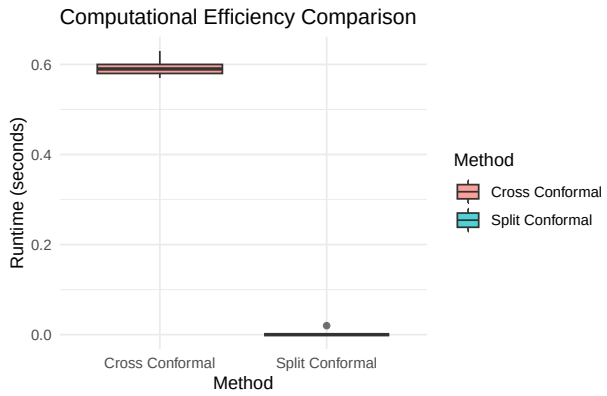

Figure 1. Computational efficiency

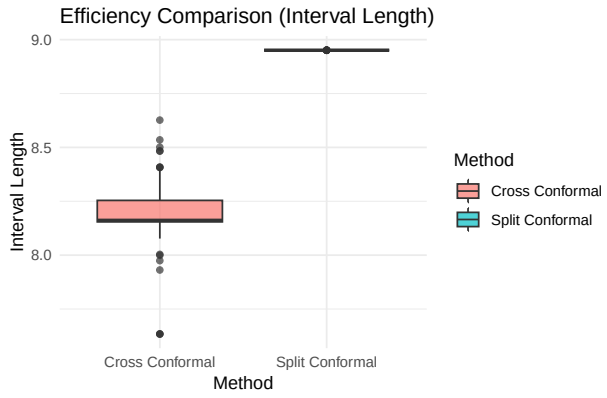

Figure 2. Predictive efficiency

In the following parts of this report, we consider three settings that arise in real-world applications where standard conformal prediction fails. However, when using p-value-based methods, key limitations emerge, we show that these challenges can be addressed by constructing conformal sets using e-values, an approach known as conformal e-prediction. I rely primarily on source [3] for the next sections.

8. Batch anytime valid conformal prediction

We consider the following setup: Data arrive sequentially in batches b_t, $t = 1, 2, \dots$, where each batch contains $n_t \geq 0$ calibration data points $S_{t,1}, \dots, S_{t,n_t}$ and one test data point S_{t,n_t+1}. Batches arrive sequentially, and their total number may be unknown in advance. Moreover, the data distribution may shift between batches; the observations in different batches are not necessarily identically distributed. We only assume that, within each batch, the data (both calibration and test points) are exchangeable conditional on the previous batches.

Given $\alpha \in (0, 1)$, the goal is to construct a sequence of batch anytime-valid conformal sets $\widehat{C}_t$ for all $t \geq 1$, based on all previous batches and the calibration data from batch t, such that

$$\mathbb{P}\Big(\forall t \geq 1,\ S_{t,n_t+1} \in \widehat{C}_t\Big) \geq 1 - \alpha, \qquad (6)$$

where the probability is taken over all data points.

If (6) holds, we call $\{\widehat{C}_t\}_{t\geq 1}$ a sequence of batch anytime-valid conformal sets. There are many real-world applications, where this setting can be used. An example of this problem arises in pharmaceutical drug deployment across hospitals. When a new drug is introduced, we aim to seek statistical guarantees on its efficacy. The drug's effect on patient i in hospital b_t is denoted by $S_i^{(t)}$, where each hospital corresponds to a data batch.

In hospital b_t, where the drug is tested on n_t volunteer patients, the goal is to construct a conformal prediction set for $S_{n_t+1}^{(t)}$, representing the treatment outcome of the next patient in that hospital.

We will also discuss why the standard method, defined with p-values, in conformal prediction cannot be used for batch anytime valid conformal prediction. With this method a set $\widehat{C}_t$ is produced for a batch of exchangeable data

$$S_1^{(t)}, \dots, S_{n_t}^{(t)}, S_{n_t+1}^{(t)}$$

such that

$$1 - \alpha \ \leq\ \mathbb{P}\Big(S_{n_t+1}^{(t)} \in \widehat{C}_t\Big) \ <\ 1 - \alpha + \frac{1}{n_t + 1},$$

where the rightmost inequality holds if the data

are almost surely distinct.

In dynamic scenarios such as batch anytime-valid conformal prediction, simply applying the standard conformal prediction method to each batch does not yield batch anytime-valid guarantees. We formulate this fact in the following lemma:

Lemma 1. *Let $\widehat{C}_t$ be the conformal set obtained using standard conformal prediction on batch b_t for any $t \geq 1$. Suppose that:*

(i) almost surely, there are no ties within any batch;
(ii) batches are independent; and
(iii) $n_t \geq 2\alpha^{-1}$ for all $t \geq 1$.

Then $\{\widehat{C}_t\}_{t\geq 1}$ is not a sequence of batch anytime-valid conformal sets.

Proof. Let

$$T := \frac{\log(1-\alpha)}{\log(1-\alpha/2)}.$$

We have

$$\mathbb{P}\Big(\forall t \geq 1,\ S^{(t)}_{n_t+1} \in \widehat{C}_t\Big) \leq \mathbb{P}\Big(\forall t = 1,\dots,T,\ S^{(t)}_{n_t+1} \in \widehat{C}_t\Big).$$

By independence of the batches,

$$= \prod_{t=1}^{T} \mathbb{P}\Big(S^{(t)}_{n_t+1} \in \widehat{C}_t\Big) \ < \ \prod_{t=1}^{T}\left(1-\alpha+\frac{1}{n_t+1}\right).$$

Under the assumption $n_t \geq 2\alpha^{-1}$ for all $t \geq 1$, we have

$$1-\alpha+\frac{1}{n_t+1} \leq 1-\frac{\alpha}{2},$$

so that

$$\prod_{t=1}^{T}\left(1-\alpha+\frac{1}{n_t+1}\right) < (1-\alpha/2)^T.$$

By the definition of T,

$$(1-\alpha/2)^T \leq 1-\alpha.$$

The equality above follows from the fact that each set $\widehat{C}_t$ depends only on data from batch b_t, and the batches are independent. The last two inequalities follow from the assumption on n_t and the definition of T, respectively. □

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P}$ is a probability measure. We will work with real-valued random variables.

Theorem 1 (Ville's Inequality)**.** *Let $\{M_t\}_{t\geq 0}$ be a nonnegative supermartingale. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\left(\forall t \geq 0,\ M_t < \frac{1}{\alpha}\right) \geq 1-\alpha.$$

In the following, we focus on supermartingales, as they are essential for e-prediction. We begin with the definition, and then state the above theorem for these processes in a more specialized setting.

Definition 6. *A process $(M_t)_{t\in\mathcal{T}}$ is called a test supermartingale for $\mathcal{P}$ if for every $P \in \mathcal{P}$ it satisfies:*

1. *$M_t \geq 0$ P-almost surely for all $t \in \mathcal{T}$,*
2. *$(M_t)_{t\in\mathcal{T}}$ is a supermartingale under P,*
3. *$\mathbb{E}_P[M_0] \leq 1$.*

A family of processes $(M^P)_{P\in\mathcal{P}}$ is called a test supermartingale family if each M^P is a test supermartingale for P.

Remark 2 (Ville's Inequality with Stopping Times on Test Supermartingals)**.** *An alternative formulation of Ville's inequality involves stopping times.*

Given a nonnegative test supermartingale $\{M_t\}_{t\geq 0}$, Ville's inequality extends to stopping times as follows:

$$\mathbb{P}\left(M_\tau < \frac{1}{\alpha}\right) \geq 1-\alpha.$$

This ensures that the coverage guarantee remains valid regardless of when the process is stopped based on observed data.

To apply Ville's inequality, we first need to choose a nonnegative test supermartingale, which involves defining an appropriate filtration. Throughout this section, we consider the filtration of σ-algebras generated by all random variables from the data batches obtained so far.

$$\mathcal{F}_1 = \sigma(S_1^1,\dots,S_1^{n_1},S_1^{n_1+1}),$$
$$\mathcal{F}_2 = \sigma(S_1^1,\dots,S_1^{n_1+1},S_2^1,\dots,S_2^{n_2+1}),$$

and more generally,

$$\mathcal{F}_t = \sigma(S_1^1, \dots, S_1^{n_1+1}, \dots, S_t^1, \dots, S_t^{n_t+1}).$$

Specifically, $\mathcal{F}_0$ is the trivial σ-algebra $\{\emptyset, \Omega\}$.

Now we need to define a sequence of random variables $\{M_t\}_{t\geq 0}$, which will be essential for the anytime-valid conformal sets.

Theorem 2. *For all $t \geq 0$, the sequence of random variables $\{M_t\}_{t\geq 0}$ defined by*

$$M_t = \prod_{s=1}^{t} E_s, \quad \textit{where}$$
$$E_s = \frac{S_s^{n_s+1}}{\frac{1}{n_s+1}\sum_{j=1}^{n_s+1} S_s^j} \quad \textit{for all } s \geq 1,$$

is a nonnegative test supermartingale.

Proof. [3] [5] Firstly, $M_0 = 1$, because it is an empty product. Since we assume that all scores are positive, it is clear that $M_t \geq 0$ for all $t \geq 0$.

It remains to show that $\{M_t\}_{t\geq 0}$ is a supermartingale. Let $t \geq 1$. Then,

$$\begin{aligned}\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\Big[\prod_{s=1}^{t} E_s \,\Big|\, \mathcal{F}_{t-1}\Big] \\ &= \prod_{s=1}^{t-1} E_s \cdot \mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \\ &= M_{t-1} \cdot \mathbb{E}[E_t \mid \mathcal{F}_{t-1}].\end{aligned}$$

We are going to show that $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] = 1$. Introduce the random variables

$$E_s^i = \frac{S_s^i}{\frac{1}{n_s+1}\sum_{j=1}^{n_s+1} S_s^j}, \qquad i = 1, 2, \dots, n_s + 1.$$

Due to the exchangeability of $S_s^1, \dots, S_s^{n_s+1}$, the random variables $E_s^1, \dots, E_s^{n_s+1}$ are identically distributed, and hence they all have the same expectation $\mathbb{E}(E_s^1)$.

Observe that

$$E_s^1 + E_s^2 + \cdots + E_s^{n_s+1} = \frac{\sum_{i=1}^{n+1} S_s^i}{\frac{1}{n+1}\sum_{j=1}^{n+1} S_s^j} = n + 1.$$

Taking expectations yields

$$\begin{aligned}\mathbb{E}(E_s^1 + E_s^2 + \cdots + E_s^{n_s+1}) &= (n+1)\mathbb{E}(E_s^{n_s+1}) \\ &= (n+1)\mathbb{E}(E_s).\end{aligned}$$

Since the sum equals $n + 1$, we obtain

$$(n+1)\mathbb{E}(E_s) = n + 1,$$

and therefore $\mathbb{E}(E_s) = 1$. We assumed that, within a given batch, the data are exchangeable conditional on the previous batches, so we have $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] = 1$.

Therefore, $\{M_t\}_{t\geq 0}$ is a martingale, and in particular a supermartingale. □

By applying Ville's inequality to the test supermartingale $\{M_t\}_{t\geq 0}$, we deduce the following corollary:

Corollary 1. *For $t \geq 1$, let $\widehat{C}_t$ be the subset of $\mathbb{R}$ defined by*

$$\widehat{C}_t := \left\{ \upsilon \in \mathbb{R}_+ : \prod_{s=1}^{t-1} E_s \frac{\upsilon}{\frac{1}{n_t+1}\sum_{j=1}^{n_t} S_t^j + \upsilon} < \frac{1}{\alpha} \right\}. \tag{7}$$

Then $\{\widehat{C}_t\}_{t\geq 0}$ is a sequence of batch anytime-valid conformal sets.

In the definition of $\widehat{C}_t$, the variable υ serves as a placeholder for the random variable $S_{n_t+1}^{(t)}$. The set $\widehat{C}_t$ is constructed to ensure that $S_{n_t+1}^{(t)}$ falls within it with high probability.

9. Experiment

I conducted a simulation study in which data arrived sequentially in batches. Each batch consists of (n_t) calibration observations and one test observation. My goal with the experiment was to compare standard and anytime-valid conformal prediction if the data distribution changes over time.

The observations in the first part of the sequence were generated from a normal distribution with variance 1, while after batch 30 the variance increased to 5.

For each batch, two conformal prediction intervals were constructed and compared. The first was the standard conformal prediction interval, obtained by computing the empirical $((1-\alpha))$-quantile of the conformity scores within

the batch. The conformity score was defined as the absolute residual from the batch mean. This construction guarantees marginal coverage of approximately $(1-\alpha)$ within each batch.

The second method was the batch anytime-valid conformal procedure based on e-prediction based on a test martingale we have constructed earlier. The e-value for the batches and the corresponding test martingale was defined the same as in **Theorem 2**.

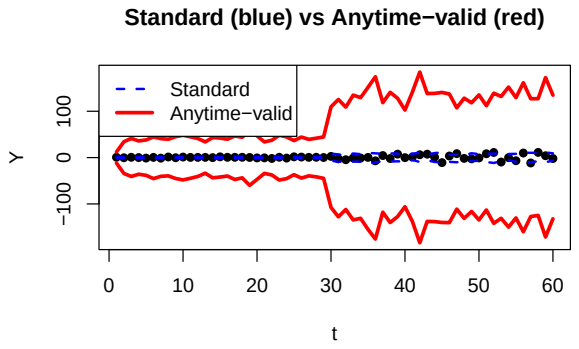

Figure 3. Conformal band

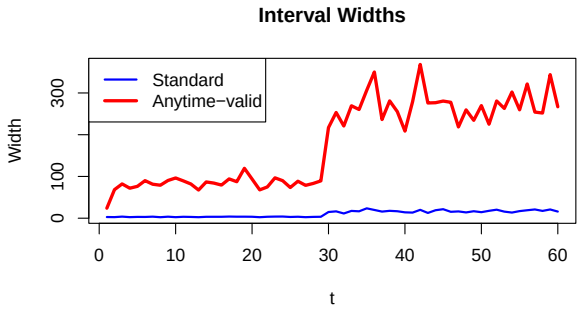

Figure 4. Interval Width

Figure 3. and Figure 4. show the difference between the behaviour of the two conformal prediction methods. For the first part of the batches, when the data-generating distribution remained stable, the standard conformal and the anytime-valid intervals were similar in width. Both methods had a relatively narrow prediction band. However, after the variance increased, the standard conformal intervals remained relatively narrow because they depended only on the current batch, providing coverage for one batch. The problem with the standard method is, that it does not guarantee validity over the full sequence. In contrast, the anytime-valid conformal intervals widened substantially after the shift. This expansion was due to the fact that the prediction used the martingale and over time it was needed to maintain the global coverage guarantee. These results highlight the main trade-off between the two approaches.

Standard conformal prediction is efficient under stable conditions but does not provide sequential validity, whereas batch anytime-valid conformal prediction is suitable in non-stationary environments.

10. Data-dependent coverage

Conformal prediction methods I used last semester operate by fixing a coverage level α in advance, ensuring that the true label falls within the conformal set with probability at least $1-\alpha$. This method does not regulate the size of the conformal prediction sets, which can differ depending on the data distribution. Instead of fixing α a priori, we allow the coverage level to adjust according to the observed data, enabling more flexible conformal sets.

Definition 7. *We say that a nonnegative random variable P is a post-hoc p-variable if*

$$\sup_{\tilde{\alpha}} \frac{\mathbb{P}(P \le \tilde{\alpha} \mid \tilde{\alpha})}{\tilde{\alpha}} \le 1,$$

where the supremum is over every random variable $\tilde{\alpha} > 0$ that may not be independent from P.

Remarkably, post-hoc p-variables are exactly the inverses of e-variables.

Theorem 3. *P is a post-hoc p-variable if and only if*

$$\mathbb{E}\left[\frac{1}{P}\right] \le 1.$$

Therefore, conformal e-prediction methods yield data-dependent coverage guarantees, in the following sense.

Proposition 6. *Consider a calibration set $\{(X_i, Y_i)\}_{i=1}^n$ and a test data point (X_{n+1}, Y_{n+1}) such that*

$$(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$$

are exchangeable. Let $\tilde{\alpha}$ be any coverage level that may depend on this data. Then we have that

$$\mathbb{E}\left(\frac{\mathbb{P}\left(Y_{n+1} \notin \hat{C}_n^{\tilde{\alpha}}(X_{n+1}) \mid \tilde{\alpha}\right)}{\tilde{\alpha}}\right) \leq 1.$$

where define $\hat{C}_n^{\tilde{\alpha}}(x)$ as

$$\left\{y : \frac{S(x,y)}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S(X_i, Y_i) + S(x,y)\right)} < \frac{1}{\tilde{\alpha}}\right\}.$$

Proof. We have seen earlier that the random variable

$$E = \frac{S(X_{n+1}, Y_{n+1})}{\frac{1}{n+1}\sum_{i=1}^{n+1} S(X_i, Y_i)} \tag{8}$$

is an e-variable. Therefore, by the previous theorem, $P = 1/E$ is a post-hoc p-variable. Therefore, we only need to see that

$$P \leq \tilde{\alpha} \iff E \geq \frac{1}{\tilde{\alpha}}$$

However,

$$E \geq \frac{1}{\tilde{\alpha}} \iff \frac{S(X_{n+1}, Y_{n+1})}{\frac{1}{n+1}\sum_{i=1}^{n+1} S(X_i, Y_i)} \geq \frac{1}{\tilde{\alpha}}$$
$$\iff Y_{n+1} \notin \hat{C}_n^{\tilde{\alpha}}(X_{n+1})$$

by definition of $\hat{C}_n^{\tilde{\alpha}}$. So we proved the equivalence. □

In applications, one can define a data-dependent random variable $\tilde{\alpha}$ that depends only on the observed data $\{(X_i, Y_i)\}_{i=1}^n$ and the test input X_{n+1}, but not on the unknown response Y_{n+1}.

When $\tilde{\alpha} = \alpha$ is a fixed constant independent of the data, the guarantee in *Proposition 5* reduces to

$$\mathbb{P}\left(Y_{n+1} \notin \hat{C}_\alpha^n(X_{n+1})\right) \leq \alpha,$$

which coincides with the core assumption of conformal prediction.

E-variables enable conformal prediction guarantees that adapt to the observed data. In particular, they allow the construction of prediction sets with controlled size.

This approach allows practitioners to adaptively choose $\tilde{\alpha}$ based on the data. Such post hoc guarantees are only possible with e-variables, in contrast to classical conformal prediction methods that rely on a fixed miscoverage level α. We can also construct conformal sets with a specific size. Suppose our aim is to get conformal sets of size at most C. Then we can define:

$$\tilde{\alpha} := \inf\left\{\alpha \in (0,1) : \left|\hat{C}_n^{\alpha}(x)\right| \leq C\right\} \tag{9}$$

Clearly, it is not achievable with standard conformal prediction.

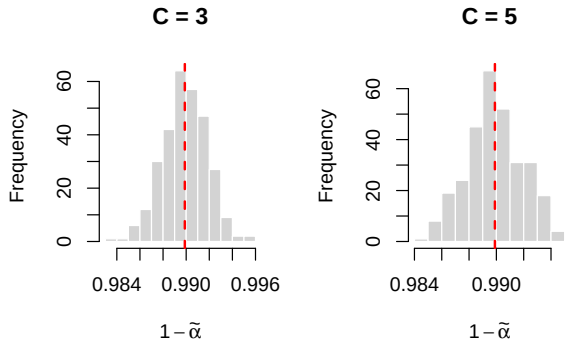

Figure 5. Histogram of $1 - \tilde{\alpha}$ for $C = 3$ and $C = 5$, computed across 300 iterations. Conformal sets were constructed using (9) on synthetic data. In each repetition, we generate synthetic calibration and test data based on random score matrices. The calibration data consist of a matrix in $\mathbb{R}^{200\times10}$ where each entry is independently sampled from a uniform distribution on $[0,1]$. The calibration sample size is $n_{\text{cal}} = 200$ and $K = 10$ is the number of scores per observation. The test observation is represented by a vector whose components are also independently sampled from a uniform distribution. For each observation, the score is defined as the largest value among the K candidate scores in that row.

11. Monte Carlo conformal e-prediction

The third method emerges in machine learning, where labels are uncertain, particularly in scenarios where multiple experts predict the outcome. Instead of standard feature-label pairs (X_i, Y_i), we may encounter $(X_i, Y_i^{(j)})$ for $j = 1, \ldots, m$, where m is the number of experts providing predictions.

This framework arises, for instance, in medicine, where a patient X_i is assessed by multiple experts, each predicting a diagnosis $Y_i^{(j)}$. As a result, the data are no longer exchangeable, which means the application of

traditional conformal prediction methods based on rank statistics cannot be used here.

We describe an approach, called *Monte Carlo conformal prediction* to this setting. The calibration set consists of features $X_i \sim P_X$, and for each X_i, m labels $Y_i^{(j)} \sim P_{Y|X_i}$ for $j = 1, \dots, m$.

Consequently, the calibration set

$$\left\{(X_i, Y_i^{(j)})\right\}_{i=1,\dots,n}^{j=1,\dots,m}$$

no longer consists of exchangeable data, making the standard conformal prediction technique inapplicable.

The conformal e-prediction framework also provides valid guarantees in Monte Carlo conformal prediction.

Instead of choosing a single label to enforce exchangeability, using all m labels helps reduce variability. A method based on averaging arbitrary p-variables was initially introduced, which yields a valid p-variable. However, this comes at the cost of achieving coverage $1 - 2\alpha$ instead of $1 - \alpha$.

We now show that using e-variables instead yields conformal sets with coverage $1 - \alpha$. Suppose $E_1, \dots, E_m$ are e-variables, then their arithmetic mean

$$\bar{E} := \frac{1}{m} \sum_{j=1}^{m} E_j \tag{10}$$

is also an e-variable.

For each $j \in \{1, \dots, m\}$, we denote $S_i^{(j)} := S(X_i, Y_i^{(j)})$ for $i = 1, \dots, n$ as the nonconformity score, and $S_{n+1} := S(X_{n+1}, Y_{n+1})$. They are all exchangeable; therefore,

$$E_j := \frac{S_{n+1}}{\frac{1}{n+1}\left(\sum_{i=1}^{n} S_i^{(j)} + S_{n+1}\right)}$$

is an e-variable.

Applying Markov's inequality to $\bar{E}$ yields a new construction of conformal sets for Monte Carlo conformal prediction.

Theorem 4.

$$\hat{C}_n(x) := \left\{y : \bar{E} < \frac{1}{\alpha}\right\} \tag{11}$$

satisfies that

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}_n(X_{n+1})\right) \geq 1 - \alpha$$

The theorem can be interpreted as showing that incidental errors are reduced by averaging across the contributions of all m experts.

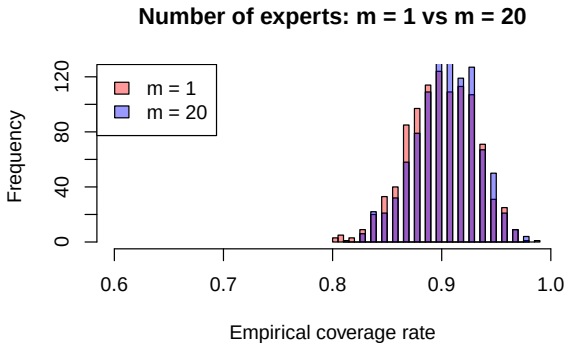

Figure 6. Comparison of coverage using e-variables in Monte Carlo conformal prediction with $m = 1$ or $m = 20$ experts, with $\alpha = 0.15$ in **Theorem 4.** The experiment was repeated 1000 times. In each repetition, a training and testing dataset of size $n = 100$ were generated independently from a normal distribution with additive Gaussian noise.

Disclaimer. I used generative AI for code generation and for grammar and language proofreading.

References

[1] Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Boston, MA: Springer US, 2005.

[2] Ramdas, Aaditya, and Ruodu Wang. "Hypothesis testing with e-values." Foundations and Trends® in Statistics 1.1-2 (2025): 1-390.

[3] Gauthier, Etienne, Francis Bach, and Michael I. Jordan. "E-values expand the scope of conformal prediction." arXiv preprint arXiv:2503.13050 (2025).

[4] Vovk, Vladimir. "Conformal e-prediction." Pattern Recognition 166 (2025): 111674.

[5] Balinsky, A. A., and Balinsky, A. D. (2024). Enhancing conformal prediction using e-test statistics. In Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications (Proceedings of Machine Learning Research, Vol. 230, pp. 65–72).

[6] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS Biology, 13(3), e1002106.

[7] Csillag, Daniel, et al. "Extending Prediction-Powered Inference through Conformal Prediction." arXiv preprint arXiv:2510.16166 (2025).