Aggregation in conformal e-classification

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #48

May 8, 2026

Project web site:
http://alrw.net

Abstract

Aggregating conformal predictors is a standard way of balancing their predictive and computational efficiency while retaining their validity, at least approximately. An important advantage of conformal e-predictors is that they are easier to aggregate without sacrificing their validity. This paper studies experimentally cross-conformal e-prediction, which is an existing method of aggregating conformal e-predictors, and its modifications that are conceptually simpler and more flexible.

Contents

1 Introduction

Full conformal prediction is often computationally inefficient, and a natural way of improving its computational efficiency is to split the training set into two parts, train a base predictor on one part, and calibrate its predictions on the other part. This approach is called inductive conformal prediction (Vovk et al., 2022, Sect. 4.2) or split conformal prediction (Angelopoulos et al., 2026). Inductive conformal prediction typically leads to a loss of predictive efficiency, and a standard way to balance predictive and computational efficiency is to aggregate several inductive conformal predictors (Carlsson et al., 2014).

A simple way of aggregating inductive conformal predictors is cross-conformal prediction (Vovk et al., 2022, Sect. 4.4). A major weakness of cross-conformal prediction and related methods based on p-values such as jackknife+ (Barber et al., 2021) is that their provable property of validity weakens significantly (namely, the guaranteed error probability may increase by a factor of 2, or approximately 2) while their typical empirical property of validity remains the same (see, e.g., Vovk et al. 2022, Sects 4.4.3 and 4.8.4).

An important advantage of the analogue of cross-conformal prediction based on e-values, or *cross-conformal e-prediction*, is that cross-conformal e-predictors inherit the standard property of validity of conformal e-predictors; there is no mismatch between provable validity and typical empirical validity (Vovk, 2025, Sect. 6). This follows from the average of e-values always being an e-value, while nothing of this kind is true for p-values (Vovk and Wang, 2020).

In this paper it will be convenient to use "conformal prediction" and related terms (such as "cross-conformal prediction") in the generic sense, covering both methods based on p-values (*conformal p-prediction*) and methods based on e-values (*conformal e-prediction*). The term "conformal prediction" will also be generic in the sense of covering full, inductive, and cross- conformal prediction. (However, in our acronyms, such as CCP for "cross-conformal p-prediction" or "cross-conformal p-predictor", we will drop an extra "P" in order to be consistent with Vovk et al. 2022.)

The main contribution of this paper is a critical discussion of cross-conformal e-prediction and proposing its more flexible modifications. Our experimental results will only cover simulation studies concerning prediction of observations taking values in a finite set $\mathbf{Y}$ and generated from the standard Bayesian model with Jeffreys's prior. (In the terminology of Vovk et al. 2022, there are no objects and we are dealing with a classification problem.) The Bayesian model will be assumed to be known, in the spirit of the Burnaev–Wasserman programme (Vovk et al., 2022, Sects 2.5 and 2.9.7). The main topic of this paper is conformal e-prediction, but the more standard conformal p-prediction will also be touched upon, especially in early sections.

We start in Sect. 2 from defining suitable ways of measuring predictive efficiency of conformal predictors. This draws on and unifies the discussions in Vovk et al. (2022, Sect. 3.1) and Vovk (2025, Sect. 7). We continue in Sect. 3 by formally introducing our Bayesian model, deriving Bayes predictors for it, and discussing ways of measuring its predictive efficiency. Simulation results for the

predictive efficiency of the Bayes predictors are given in the following section, Sect. 4, where they are compared with results for full conformal predictors. The definition of full conformal predictors is clear-cut only for p-predictors. There are at least two natural definitions of full conformal e-predictors, "ordinary" and "deleted". The deleted version leads to much better results in our simulation studies, and this is what we use in the following sections. While conformal e-predictors are deterministic, there are two versions of conformal p-predictors, deterministic and smoothed, and the difference between them can be very substantial in our simulation studies (since our data sets only involve labels, without objects); we report experimental results for both.

Section 5 discusses inductive conformal prediction, which we continue in Sect. 6 with cross-conformal prediction. Cross-conformal e-prediction (CCEP) has two important advantages over inductive conformal e-prediction, the "Jensen gap" (which is an advantage under our criteria of predictive efficiency) and the cross-conformal e-predictors being nearly deterministic (having low variance); these advantages are discussed later in the paper, in Sect. 7. There are also serious limitations of CCEP: the fraction of the training observations that are used for calibration in each component inductive conformal e-predictor is $1/K$, where $K \geq 2$ (the number of folds) is an integer; therefore, first, we cannot use more than half of the training observations for calibration, and second, we cannot have fine control of the fraction. The limitations are inherited from CCP, but unlike CCP, we can modify CCEP to address them. The first limitation can be overcome by using "inverse cross-conformal e-predictors" (Sect. 6), but a more radical approach overcoming both limitations is to use repeated inductive conformal e-prediction, which involves averaging several independent inductive conformal e-predictors. If we have an idea of a suitable proportion in which to split the training set into proper training and calibration parts, we can replace CCEP by RICEP (repeated inductive conformal e-prediction), in which we can control the Jensen gap and the variance much better; RICEP also allows splitting the training set in any proportion, while for CCEP the proportions are severely limited. This is done in Sect. 7.

A disadvantage of both CCEP and RICEP is that they need the split proportion as their parameter; in the case of CCEP the standard parameter is the number K of folds, which corresponds to the split proportion of $(K-1):1$. In the case of cross-validation, a common belief is that $K = 5$ or $K = 10$ are reasonable choices in wide generality. In Sect. 6 we will see that there are no such universal choices in the case of CCEP (and CCP), and a natural approach is to mix over all possible proportions with uniform weights. This leads to "balanced inductive conformal e-prediction", or BICEP. The predictive performance of RICEP and BICEP is studied in Sect. 7. Section 8 concludes summarizing this paper's findings.

This paper does not contain any non-trivial theoretical results, and its goal is modest: to perform simulation studies of several conformal prediction algorithms and suitable values of their parameters (such as the number K of folds in cross-conformal prediction) and to give some recommendations. All Jupyter notebooks required to reproduce its experimental results can be accessed via

the website https://alrw.net (under Working Paper 48).

In this paper I will follow Greenland (2024) and say "p-surprisal" for $-\ln p$, where p is a p-value. This terminology (unlike "S-value" used by Greenland elsewhere (2019)) allows an easy extension to e-values; namely, $\ln e$, where e is an e-value, will be called "e-surprisal".

2 Evaluating the predictive efficiency of conformal predictors

Starting from the next section we will consider the simple case where there are no objects and each observation is just a label. But the discussion in this section will be more general. Our goal here is to introduce suitable criteria of efficiency for conformal prediction. In the case of conformal e-prediction, it is the main criterion proposed in Vovk (2025); we will also adapt this criterion to conformal p-prediction obtaining a new criterion that is almost (but not quite) a special case of the S criterion of Vovk et al. (2022, Sect. 3.1.1) (as generalized in Vovk et al. 2022, Remark 3.15).

Suppose we have a training set of size l and a test set with labels $y_{l+1}, \dots, y_{l+k}$. In this paper we will evaluate the predictive efficiency of a conformal p-predictor trained on the training set by the value

$$\frac{1}{k(|\mathbf{Y}|-1)} \sum_{i=l+1}^{l+k} \sum_{y \in \mathbf{Y} \setminus \{y_i\}} (-\ln p_i^y), \tag{1}$$

where p_i^y is the conformal p-value for the postulated label y for the ith test observation. Let us call it the *AFS criterion* (average false p-surprisal); it is the arithmetic average of the conformal p-surprisals for all false labels. An attractive property of criteria of efficiency for conformal p-prediction is being *conditionally proper* (Vovk et al., 2022, Sect. 3.1.5). It is reassuring that this property is satisfied by the AFS criterion.

Proposition 1. *The AFS criterion is a conditionally proper criterion of efficiency in the terminology of Vovk et al. (2022, Sect. 3.1.5).*

The proof of Proposition 1, as well as all other proofs in this paper, is given in Appendix A; it is a simple modification of the argument in Vovk et al. (2022, Remark 3.15).

The use of logarithms in (1) has many precedents and is introduced here in analogy with a similar expression for e-values (cf. (2) below). For example, logarithms of p-values are summed in Fisher's method for combining p-values (Fisher, 1925, Sect. 21.1), also known as Fisher's combined probability test. Fisher's argument for this way of combining seems to have been that p-values are akin to probabilities, and independent probabilities combine by multiplication (Fisher, 1925, the end of the first paragraph in Sect. 21.1). Logarithms of p-values were used by Martin-Löf (1966), and now they are ubiquitous in the

algorithmic theory of randomness. The universal p-values are defined to within an additive constant on the logarithmic scale, and adding logarithms of p-values features prominently in many theorems of the algorithmic theory of randomness. In statistics, logarithms of p-values have been widely used by Greenland (2019, 2024) and co-authors. S-values (i.e., p-surprisals) are believed to be additive, at least in the independent case (Greenland, 2019, Sect. 4.4).

The analogue for e-values of the AFS criterion (1) is the *AFES criterion*

$$\frac{1}{k(|\mathbf{Y}|-1)} \sum_{i=l+1}^{l+k} \sum_{y \in \mathbf{Y} \setminus \{y_i\}} \ln e_i^y, \tag{2}$$

where e_i^y is the conformal e-value for the postulated label y for the ith test observation. It was proposed in Vovk (2025) together with its simpler but less natural counterpart

$$\frac{1}{k\,|\mathbf{Y}|} \sum_{i=l+1}^{l+k} \sum_{y \in \mathbf{Y}} \ln e_i^y. \tag{3}$$

It is shown in Vovk (2025, Proposition 15) that in an idealized setting with an infinite training set and $k \to \infty$ the optimal nonconformity e-measure under the AFES criterion (2) is very natural, namely, the nonconformity score of an observation is proportional to the odds against its label (observed for a training observation or postulated for a test observation) conditional on its object. That proposition can be interpreted as the AFES criterion satisfying an analogue for e-values of being conditionally proper. And it is shown in Vovk (2025, Remark 16) that in the same idealized setting the optimal nonconformity e-measure under the modified AFES criterion (3) is even simpler, namely, the nonconformity score of an observation is proportional to the inverse of the conditional probability of its label given its object.

The justification for the use of logarithms in (2) is that logarithms of e-values are additive, as discussed by Shafer (2021, Sect. 2.2.1) with a reference to Kelly. For further discussions, see Ramdas et al. (2023, Sect. 3.1) and Grünwald et al. (2024, Sect. 2.1).

Both criteria, (1) and (2), will be used in this paper in their limiting form as $k \to \infty$.

3 Our Bayesian model and Bayes prediction

As mentioned earlier, in this paper we are interested in the case of classification with no objects; the label space is assumed to be $\mathbf{Y} = \{1, \dots, Y\}$, and therefore, the observation space is also $\{1, \dots, Y\}$. Our postulated model for the data-generating mechanism is Bayesian: the parameter is $\theta \in \Theta$, the parameter space

$$\Theta := \left\{ \theta \in [0,1]^Y : \sum_{y=1}^{Y} \theta_y = 1 \right\}$$

is the standard simplex, P_θ is the probability measure on $\mathbf{Y}$ defined by $P_\theta(\{y\}) := \theta_y$, $y = 1, \dots, Y$, and $\theta \sim \mathrm{Dir}(\alpha, \dots, \alpha)$. Here $\mathrm{Dir}(\alpha, \dots, \alpha)$ is the Dirichlet distribution with all parameters equal to $\alpha \in (0, \infty)$, which we will abbreviate to Dir_α. Important special cases are $\alpha = 1$ (the uniform distribution) and $\alpha = 0.5$ (Jeffreys's prior).

If the training set has size n and there are n_y ys in it for all $y \in \mathbf{Y}$, the probability prediction P for the test observation is

$$\begin{aligned} P(y) &= \frac{\int_\Theta \theta_y^{n_y+1} \prod_{y' \neq y} \theta_{y'}^{n_{y'}} \mathrm{Dir}_\alpha(\mathrm{d}\theta)}{\int_\Theta \prod_{y'} \theta_{y'}^{n_{y'}} \mathrm{Dir}_\alpha(\mathrm{d}\theta)} = \frac{\int_\Theta \theta_y^{n_y+\alpha} \prod_{y' \neq y} \theta_{y'}^{n_{y'}+\alpha-1} \mathrm{d}\theta}{\int_\Theta \prod_{y'} \theta_{y'}^{n_{y'}+\alpha-1} \mathrm{d}\theta} \\ &= \frac{B(n_1 + \alpha, \dots, n_{y-1} + \alpha, n_y + \alpha + 1, n_{y+1} + \alpha, \dots, n_Y + \alpha)}{B(n_1 + \alpha, \dots, n_Y + \alpha)} \\ &= \frac{\Gamma(n_y + \alpha + 1)/\Gamma(n + Y\alpha + 1)}{\Gamma(n_y + \alpha)/\Gamma(n + Y\alpha)} = \frac{n_y + \alpha}{n + Y\alpha}, \end{aligned} \tag{4}$$

where we drop curly braces in expressions such as $P(\{y\})$ and use $\sum_{y'} n_{y'} = n$. This is a generalization of Laplace's rule of succession.

Let us now restate this Bayes prediction rule in terms of p-values and e-values, to facilitate comparison with conformal prediction. Suppose we have a training set of size l. Let n_y, $y \in \mathbf{Y}$, be the number of ys in this training set, as before. The optimal smoothed p-variable, in the sense of the Bayesian analogue of the AFS criterion (1), that is valid under our Bayesian model is

$$p_y = A_y + \tau B_y, \text{ where } A := \sum_{y' : P(y') < P(y)} P(y') \text{ and } B := \sum_{y' : P(y') = P(y)} P(y'), \tag{5}$$

where τ is uniformly distributed on $[0, 1]$; this is the content of the following proposition.

Proposition 2. *The p_y defined by* (5) *comprise a p-variable valid under our Bayesian model attaining*

$$\mathbb{E} \sum_{y' \in \mathbf{Y} \setminus \{y\}} (-\ln p_{y'}) \to \max \tag{6}$$

(cf. (1)*), where the expectation is for $y_1, \dots, y_l, y$ generated from our Bayesian model.*

Let us define the *Bayesian p-predictor* (denoted "p-Bayes" in figures below) as the predictor of the same type as a conformal predictor but outputting the optimal Bayesian p-values (5) in place of conformal p-values. While Vovk et al. (2022, Theorem 3.1 and Remark 3.15) are about an idealized setting with infinite training and test sets and IID observations, Proposition 2 is about a finite training set.

The expected p-surprisal for (5) is

$$\mathbb{E}_\tau (-\ln p_y) = F(A_y, B_y), \tag{7}$$

where F is defined as the definite integral

$$F(A,B) := -\int_0^1 \ln(A+B\tau)\,\mathrm{d}\tau = -\left[\left(\frac{A}{B}+\tau\right)\ln(A+B\tau)-\tau\right]_0^1$$
$$= \frac{A}{B}\ln A + 1 - \frac{A+B}{B}\ln(A+B) \tag{8}$$

for $A \geq 0$ and $B > 0$. In computer code the case $A = 0$ should be considered separately, and the last expression should be replaced by $1 - \ln B$. The value of the AFS criterion (1) for a given training set and an infinite test set is, a.s.,

$$\sum_{y\in\mathbf{Y}} \theta_y \frac{1}{Y-1} \sum_{y'\neq y} F(A_{y'}, B_{y'}) = \frac{1}{Y-1} \sum_{y\in\mathbf{Y}} (1-\theta_y) F(A_y, B_y), \tag{9}$$

where $\theta = (\theta_y)$ is the realized parameter value (generated from the Dirichlet distribution). In our computer code, we average it over many simulations of θ and the training set.

In analogy with Vovk (2025, Proposition 15), the optimal Bayesian e-value (valid under our Bayesian assumption) for a test observation y is

$$e_y := \frac{1}{Y-1} \frac{1 - \frac{n_y+\alpha}{l+Y\alpha}}{\frac{n_y+\alpha}{l+Y\alpha}} = \frac{1}{Y-1}\left(\frac{l+Y\alpha}{n_y+\alpha} - 1\right). \tag{10}$$

This is spelled out in the following proposition, which extends Vovk (2025, Proposition 15) from conformal to Bayesian e-prediction.

Proposition 3. *The e-values e_y defined by* (10) *provide the only solution to the optimization problem*

$$\mathbb{E} \sum_{y'\in\mathbf{Y}\setminus\{y\}} \ln e_{y'} \to \max \tag{11}$$

(cf. (2)*), where the expectation is for $y_1, \ldots, y_l, y$ generated from our Bayesian model.*

Similarly to the case of p-values, we define the *Bayesian e-predictor* (denoted "e-Bayes" in figures) as the predictor of the same type as a conformal e-predictor but outputting the optimal Bayesian e-values (10) in place of conformal e-values. Similarly to Proposition 2, Proposition 3 modifies its conformal counterpart, Vovk (2025, Proposition 15), making it more realistic in that the training set becomes finite, but on the negative side we have to make an extra Bayesian assumption.

Now the value of the AFES criterion (2), i.e., the expectation of the average conformal e-surprisal for false labels, for a given parameter (θ_y), given training set, and an infinite test set, is

$$\sum_{y\in\mathbf{Y}} \theta_y \frac{1}{Y-1} \sum_{y'\neq y} \ln e_{y'} = \frac{1}{Y-1} \sum_{y\in\mathbf{Y}} (1-\theta_y) \ln e_y \tag{12}$$

a.s., where e_y is defined by (10).

A simpler modification of the e-values (10) is

$$e_y := \frac{1}{Y} \frac{1}{\left(\frac{n_y+\alpha}{l+Y\alpha}\right)} = \frac{1}{Y} \frac{l+Y\alpha}{n_y+\alpha}. \tag{13}$$

It is also valid under our Bayesian model, but is optimal in a different sense.

Proposition 4. *The e_y defined by* (13) *provide the only solution to the optimization problem*

$$\mathbb{E} \sum_{y' \in \mathbf{Y}} \ln e_{y'} \to \max$$

(cf. (3)*).*

The *suboptimal Bayesian e-predictor* ("suboptimal e-Bayes" in figures) is defined as the Bayesian e-predictor except that it outputs (13) instead of (10). Of course, it is suboptimal only because of our decision to use (2) as our main criterion of efficiency; it would have been optimal had we used (3). For a given parameter (θ_y), a given training set, and an infinite test set the value of the modified AFES criterion (3) is, a.s.,

$$\sum_{y \in \mathbf{Y}} \theta_y \frac{1}{Y} \sum_{y' \in \mathbf{Y}} \ln e_{y'} = \frac{1}{Y} \sum_{y \in \mathbf{Y}} \ln e_y. \tag{14}$$

The suboptimal Bayesian e-predictor consistently achieves better (i.e., larger) values of (14) than the Bayesian e-predictor does (but this is not shown in the figures below).

4 Full conformal prediction

The main disadvantage of the Bayesian p-values and e-values derived in the previous section is that they are only valid under the postulated Bayesian model. In this sense this Bayesian model serves as hard model in the terminology of Vovk et al. (2022, Sect. 2.5.1). In this and following sections we will use the Bayesian model only for defining nonconformity measures for conformal prediction, and the resulting p-values and e-values will be valid under exchangeability. Only their efficiency will depend on the postulated Bayesian model, and so the latter becomes our soft model.

In this section we derive the full conformal predictor (CP; we do not claim any formal properties of predictive efficiency for it, nor for its computationally efficient analogues introduced below). As before, n_y, $y \in \mathbf{Y}$, is the number of ys in the training set of size l. As conformity score of an observation y we use an estimate of the conditional probability of y or, equivalently, the number of ys in the comparison bag (the resulting conformal p-value does not depend on the details of how exactly we define the comparison bag; cf. the discussion of

ordinary, studentized, and deleted e-values below). If the true test observation is y, for each false test observation $y' \neq y$ the conformal p-value is

$$p_{y'} = A_{y'} + \tau B_{y'}, \text{ where}$$
$$A := \frac{\sum_{y'' \in \mathbf{Y} \setminus \{y'\} : n_{y''} < n_{y'}+1} n_{y''}}{l+1} \text{ and } B := \frac{\sum_{y'' : n_{y''} = n_{y'}+1} n_{y''} + n_{y'} + 1}{l+1}, \tag{15}$$

where τ is, as before, uniformly distributed on $[0,1]$. The expected p-surprisal is then given by (7) and (8). The value of the AFS criterion for a given training set and an infinite test set is given by the right-hand side of (9).

The situation with e-values is more complicated; the e-values can be ordinary, studentized, and deleted (see, e.g., Vovk et al. 2022, Sect. 7.3.1), and these versions are all different (unlike in the case of p-values). Let $\sigma \in [0,1]$ be a parameter indicating how ordinary our conformal e-predictor is; its precise meaning will be explained after introducing nonconformity scores (16) and (17). Each of the n_y observations y in the training set has nonconformity score

$$\frac{1 - \frac{n_y - 1 + \sigma + \alpha}{l + \sigma + Y\alpha}}{\frac{n_y - 1 + \sigma + \alpha}{l + \sigma + Y\alpha}} = \frac{l + \sigma + Y\alpha}{n_y - 1 + \sigma + \alpha} - 1 \tag{16}$$

in the augmented training set (the training set augmented by the postulated test observation) unless y coincides with the postulated test observation; if it does, the nonconformity score is

$$\frac{1 - \frac{n_y + \sigma + \alpha}{l + \sigma + Y\alpha}}{\frac{n_y + \sigma + \alpha}{l + \sigma + Y\alpha}} = \frac{l + \sigma + Y\alpha}{n_y + \sigma + \alpha} - 1; \tag{17}$$

the nonconformity score of a postulated test observation y is also given by (17).

The nonconformity scores (16) and (17) both use the Bayesian probabilities (4). The case $\sigma = 0$ is the main one used in this paper; it is the *deleted* version of the definition. The Bayesian training set (serving as our comparison bag) is the augmented training set minus the observation for which we are computing its nonconformity score. The size of the Bayesian training set is l, as in both (16) and (17). Both (16) and (17) are the estimated odds against y. In (16), we use $n_y - 1$ as the number of ys since we are excluding one of the observations y. And (17) covers two cases: if we are computing the nonconformity score of the test observation, there is no need to decrease n_y since n_y only includes training observations; and if we are computing the nonconformity score of a training observation, we first exclude it but then we include the test observation y in our count of ys.

For $\sigma = 1$ we will have the *ordinary* version, in which an observation y is included in the Bayesian training set (our comparison bag) when computing its nonconformity score. Its inclusion is reflected in adding $\sigma = 1$ to the numerators and denominators of all fractions in (16) and (17). The case $\sigma = 0.5$ resembles

the studentized version in that it is intermediate between deleted and ordinary (although it is between deleted and ordinary in a different sense from the usual studentized residuals Montgomery et al., 2012, Sect. 4.2.2).

The nonconformity scores (16) and (17) give us this formula for the conformal e-value for a postulated test observation y':

$$e_{y'} = \frac{\frac{l+\sigma+Y\alpha}{n_{y'}+\sigma+\alpha} - 1}{\frac{1}{l+1}\left(\sum_{y \neq y'} n_y \left(\frac{l+\sigma+Y\alpha}{n_y - 1+\sigma+\alpha} - 1\right) + (n_{y'}+1)\left(\frac{l+\sigma+Y\alpha}{n_{y'}+\sigma+\alpha} - 1\right)\right)}; \tag{18}$$

the e-value is obtained by normalizing the nonconformity score of the test observation by dividing it by the average nonconformity score in the augmented training set. This defines the (full) *conformal e-predictor* (CEP).

In computer code, we should be careful with the possibility of division by zero in (18), since $n_y - 1 + \sigma + \alpha = 0$ is possible (but only when $n_y = 0$). If $n_y = 0$, the fraction with $n_y - 1 + \sigma + \alpha = 0$ in the denominator should be set to, say, 0 (the exact value does not matter since the fraction is multiplied by $n_y = 0$).

In our code we first compute

$$S = \sum_{y \in \mathbf{Y}} n_y \left(\frac{l+\sigma+Y\alpha}{n_y - 1+\sigma+\alpha} - 1\right), \tag{19}$$

after which we can replace (18) by the quicker expression

$$e_{y'} = \frac{(l+1)\left(\frac{l+\sigma+Y\alpha}{n_{y'}+\sigma+\alpha} - 1\right)}{S - n_{y'}\left(\frac{l+\sigma+Y\alpha}{n_{y'} - 1+\sigma+\alpha} - 1\right) + (n_{y'}+1)\left(\frac{l+\sigma+Y\alpha}{n_{y'}+\sigma+\alpha} - 1\right)}. \tag{20}$$

The value of the AFES criterion is given by (12), and we evaluate its average over 10,000 independent draws of θ from Jeffreys's prior and then of the training set for various σ in Figure 1 (the solid lines). The results for $\sigma = 0$ are much better than those for $\sigma = 1$, and the results for the quasi-studentized case $\sigma = 0.5$ are intermediate. In our experiments we choose $l = 12{,}000$ as the size of our training set for two reasons. First, this choice leads to e-values comparable with Jeffreys's $\sqrt{10}$ and 10 (on the log scale, $\ln\sqrt{10} \approx 1.15$ and $\ln 10 \approx 2.30$) and to p-values comparable with Fisher's 5% and 1% (on the log scale, $\ln(1/0.05) \approx 3.00$ and $\ln(1/0.01) \approx 4.61$); see Vovk and Wang (2023, Sect. 2) for a discussion of these conventional thresholds. Second, 12,000 is convenient in the context of CCP and CCEP since it has many small divisors K (which can be conveniently used as the numbers of folds).

The nonconformity scores (16)–(17) are the counterpart of (10) in full conformal e-prediction (motivated by Vovk 2025, Proposition 15). Now let discuss the counterpart of (13) (motivated by Vovk 2025, Remark 16), where we replace the likelihood ratio against the observation by its inverse probability. It involves a slight modification, which we call *suboptimal CEP*: now we should drop all entries of "-1" outside the expressions "$n_y - 1$" and "$n_{y'} - 1$" in (16), (17), (18),

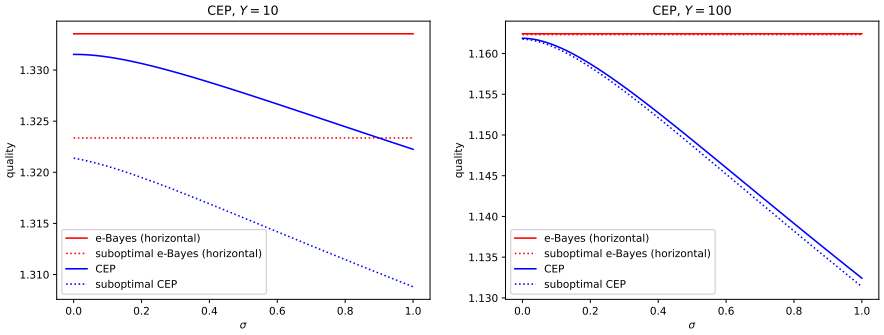

Figure 1: Results for CEP with different degrees of ordinariness $\sigma \in [0, 1]$; $l = 12{,}000$, $Y = 10$ (left), and $Y = 100$ (right). The plots show the average value of the AFES criterion over 10,000 iterations as function of σ. This is for $\alpha = 0.5$ (i.e., Jeffreys's prior).

(19), and (20). The quality of prediction as measured by the AFES criterion (12) then suffers, since this nonconformity measure is optimized for (14).

The pictures in Figure 1 cover both (10) and (13) and their full conformal analogues, with the suboptimal versions shown as dotted lines. The vertical axis is labelled "quality", which refers to the AFES criterion (or the AFS criterion in the case of p-values in some of the figures given below). The case $\sigma = 0$ (deleted CEP) remains the best for the suboptimal versions. While Figure 1 only covers the cases $Y = 10$ and $Y = 100$, the blue lines also look monotonically decreasing (albeit almost horizontal) in the case of binary classification $Y = 2$, which we often include in our experimental studies later in this paper. Therefore, from now on we always set $\sigma := 0$ in CEP and suboptimal CEP.

5 Inductive conformal prediction

In inductive conformal prediction, we randomly split the training set of size l into two parts, the *proper training set* of size $m \in \{1, \dots, l\}$ and the *calibration set* of size $m' := l - m$. Let n_y be the number of ys, $y \in \mathbf{Y}$, in the proper training set and n'_y be the number of ys in the calibration set.

5.1 ICP

Let us derive the inductive conformal p-predictor (ICP) corresponding to the AFS criterion of efficiency (1). In the calibration set, n'_y elements have conformity score $\frac{n_y+\alpha}{m+Y\alpha}$ or, equivalently, n_y, for each $y \in \mathbf{Y}$. Suppose the true test

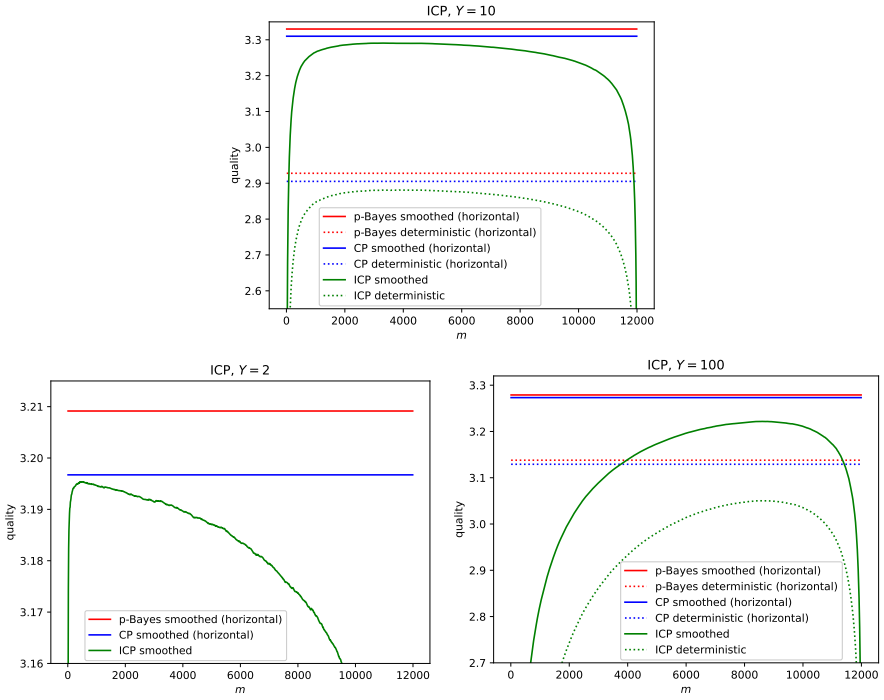

Figure 2: Results for ICP; $l = 12{,}000$, $Y = 2$ (left), $Y = 10$ (top), and $Y = 100$ (right). The average value of the AFS criterion over 10,000 iterations is shown as function of the size $m = 1, \dots, l-1$ of the proper training set. This is for Jeffreys's prior, $\alpha = 0.5$.

observation is y. For each false test observation y', the corresponding p-value is

$$p_{y'} = A_{y'} + \tau B_{y'} := \frac{\sum_{y'': n_{y''} < n_{y'}} n'_{y''} + \tau \left(\sum_{y'': n_{y''} = n_{y'}} n'_{y''} + 1 \right)}{m' + 1} \tag{21}$$

(cf. (15)). The expected p-surprisal is then (7) with y replaced by y' and with $A_{y'}$ and $B_{y'}$ defined by (21). The value of the AFS criterion for given proper training set, calibration set, and θ is still (9).

Some experimental results are given in Figure 2, with the label "quality" for the vertical axis referring to the AFS criterion with $k \to \infty$, as mentioned earlier. For deterministic p-values (corresponding to $\tau := 1$) we replace the function (8) by $F(A, B) := -\ln(A + B)$. The lines for smoothed p-values are significantly higher than those for deterministic p-values. The difference is especially large for $Y = 2$; in this case we do not show the lines for deterministic p-values since they are so far below the lines that are shown.

Figure 2 demonstrates that there is no universally applicable recommendation for a good size m' of the calibration set. For $Y = 100$ the optimal

percentage of calibration observations is small, whereas for $Y = 2$ it is very close to 1; the situation for $Y = 10$ is intermediate. This is natural as learning is much quicker for $Y = 2$: in this case we have much fewer parameters to learn than for $Y = 100$.

5.2 ICEP

In the case inductive conformal e-predictors (ICEP), we can again aim at optimizing either (2) or (3). Let us first find the ICEP corresponding to our main criterion, (2). The n'_y ys in the calibration set have

$$\frac{1 - \frac{n_y+\alpha}{m+Y\alpha}}{\frac{n_y+\alpha}{m+Y\alpha}} = \frac{m + Y\alpha}{n_y + \alpha} - 1 \tag{22}$$

as their nonconformity score (cf. (17)). Therefore, the conformal e-value for a postulated test observation y' is

$$e_{y'} = \frac{\frac{m+Y\alpha}{n_{y'}+\alpha} - 1}{\frac{1}{m'+1}\left(\sum_{y'' \in \mathbf{Y}} n'_{y''}\left(\frac{m+Y\alpha}{n_{y''}+\alpha} - 1\right) + \left(\frac{m+Y\alpha}{n_{y'}+\alpha} - 1\right)\right)}. \tag{23}$$

Overall, the value of the AFES criterion is given by (12), as in the case of full conformal e-prediction.

If we aim to optimize (3), we should drop the two subtrahends in (22) and drop the three entries of "-1" in (23). We refer to this predictor as *suboptimal ICEP* since we continue to use the criterion (2) (with $k \to \infty$) when evaluating it in Figure 3.

The analogue of Figure 2 for e-values is given as Figure 3; in the case $Y = 2$ we drop the plots for the suboptimal versions, which are very far below the area shown. The dotted lines are for the suboptimal versions; the quality of prediction as measured by (12) (and shown in the plots) suffers, but as measured by (14), it improves (which is not shown in our plots). The difference between the solid and dotted lines almost disappears for $Y = 100$.

6 Cross-conformal prediction

In this section we will see that the two varieties of cross-conformal prediction, those based on p-values and e-values, behave very differently in our experiments. Whereas for cross-conformal p-predictors (CCP) the greater the number K of folds the better, as far as predictive efficiency is concerned, for cross-conformal e-prediction (CCEP) there is a "sweet spot", and moving beyond it leads to loss of predictive efficiency. The price to pay is that the standard property of validity for full conformal p-prediction becomes violated for CCP.

We use K to denote the number of folds. Let us assume that the size l of the training set is divisible by K. Then $m' := l/K$ is the "size of the calibration set" and $m := l - m'$ is the "size of the proper training set". For each fold

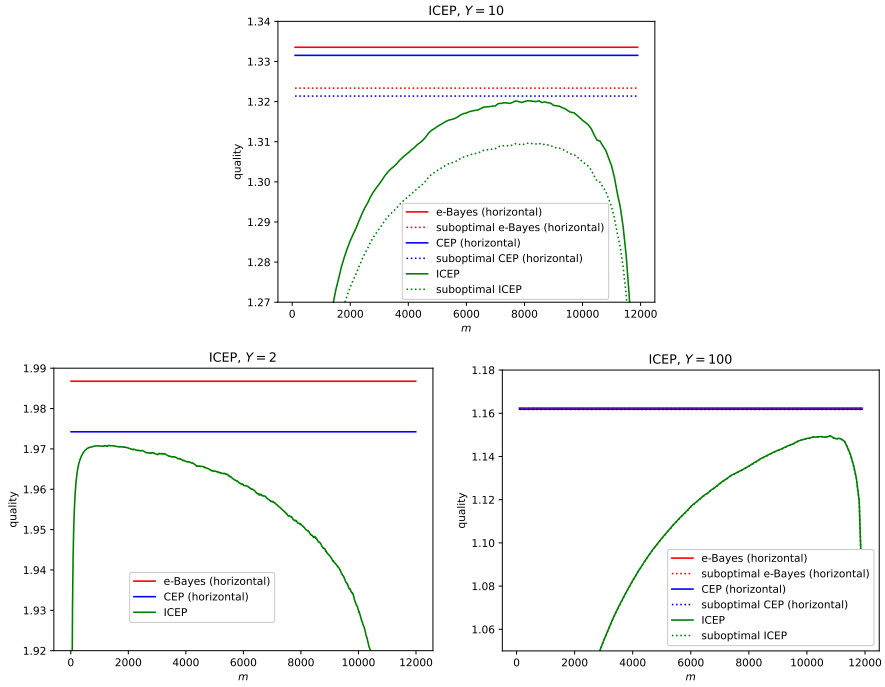

Figure 3: Results for ICEP, averaged over 10,000 iterations; $l = 12{,}000$, $Y = 2$ (left), $Y = 10$ (top), $Y = 100$ (right), and $\alpha = 0.5$.

$k = 1, \ldots, K$, let $n'_{k,y}$ be the number of ys in the calibration set (i.e., fold k) and $n_{k,y}$ be the number of ys in the proper training set (i.e., the remaining $K-1$ folds).

The shapes of the green lines in Figures 2 and 3 suggest that there is no universally suitable good value for K: for $Y = 100$ we want to include the bulk of training observations in the proper training set (and so make K large), while for $Y = 2$ we want to use as many training observations as possible for calibration in each component inductive conformal predictor (and so set $K := 2$). In fact this expectation is confirmed only for CCEP.

6.1 CCP

For each false test observation y', compute

$$A_k := \sum_{y'' : n_{k,y''} < n_{k,y'}} n'_{k,y''} \qquad \text{and} \qquad B_k := \sum_{y'' : n_{k,y''} = n_{k,y'}} n'_{k,y''}.$$

Setting

$$A := \frac{\sum_{k=1}^{K} A_k}{l+1} \qquad \text{and} \qquad B := \frac{\sum_{k=1}^{K} B_k + 1}{l+1}, \tag{24}$$

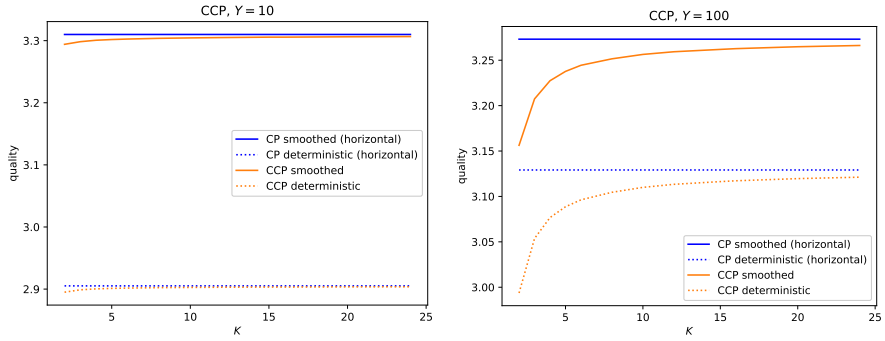

Figure 4: Results for CCP; $l = 12{,}000$, $Y = 10$ (left), $Y = 100$ (right), and $\alpha = 0.5$. The value of the AFS criterion is averaged over 10,000 iterations and shown as function of the number of folds.

we obtain the expression $p_{y'} := A + \tau B$ for the cross-conformal p-value; the expression for the deterministic cross-conformal p-value is obtained by setting $\tau := 1$. The expected p-surprisal is still (7), and the value of the AFS criterion for a given training set, its split into folds, and θ is still given by (9).

Experimental results for CCP are shown in Figure 4. The AFS quality graphs for CCP are monotonic in K in these experiments, and on the right (at $K = 12000$, which is not shown) they merge with the value for the full conformal predictor (for $Y = 10$ exactly and for $Y = 100$ with extremely high accuracy). The numbers of folds in the figure are taken from the set $K \in \{2, 3, 4, 5, 6, 8, 10, 12, 15, 16, 20, 24\}$ (so that all K are divisors of $l = 12{,}000$). The price to pay for such nice behaviour is that CCP lack provable validity, which may even show in experiments (such as those involving excessive randomization). Let us check formally (in the appendix) the coincidence of full conformal p-prediction and CCP with the maximal number of folds (*leave-one-out CCP*).

Proposition 5. *On the same data, a leave-one-out CCP produces the same p-values as the corresponding full conformal p-predictor.*

6.2 CCEP

For cross-conformal e-prediction (CCEP), we compute the e-value (23) for each fold k and each false test observation y', where all $n_y := n_{k,y}$ and $n'_y := n'_{k,y}$ now also depend on k. For each y', we then average these e-values over $k = 1, \dots, K$ getting one e-value $e_{y'}$. Finally this e-value is fed into (12).

Experimental results for CCEP are shown in Figure 5. In the left panel, K ranges over $\{2, 3, 4, 5, 6, 8, 10\}$ (with the best quality attained at $K = 3$), and in the right panel, it ranges over $\{2, 3, 4, 5, 6, 8, 10, 12, 15, 16, 20, 24\}$ (with the best quality attained at $K = 10$). For a smaller number Y of classes, we need fewer observations to learn a good prediction rule. For comparison, we also show in

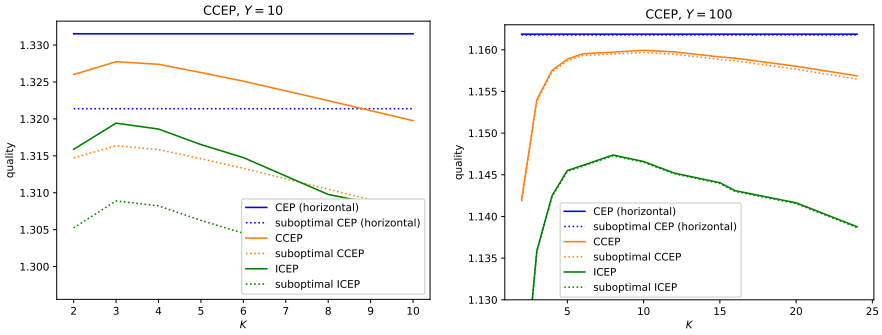

Figure 5: Results for CCEP; $l = 12{,}000$, $Y = 10$ (left), $Y = 100$ (right), and $\alpha = 0.5$. The average value of the AFES criterion over 10,000 iterations is shown as function of the number of folds (smallest divisors of l).

green the AFES quality of ICEP with the sizes $\frac{K-1}{K}l$ and $\frac{1}{K}l$ of the proper training and calibration sets, respectively.

6.3 Inverse CCEP

For $Y = 2$ (and very small Y in general) the best value of K is 2, and it appears that a smaller K, if it were possible, would be preferable. So in this subsection we consider the *inverse CCEP*. Each fold in turn is used as proper training set, and the remaining folds are used for calibration; the resulting e-values are then averaged. See Figure 6 for experimental results; now we use $\frac{1}{K}l$ and $\frac{K-1}{K}l$ as the sizes of proper training and calibration sets, respectively, for ICEP and drop "inverse" in the legend. The optimal number of folds is $K = 20$, and it produces visibly better results than $K = 2$, for which CCEP and inverse CCEP coincide.

While inverse CCP are hardly ever useful, as we have seen above, the situation with CCEP is very different. Both CCEP and inverse CCEP are subsumed, to some degree, by RICEP, which will be introduced in the next section.

7 RICEP and BICEP

It is clear that CCEP, even if complemented by inverse CCEP, is too inflexible since the number K of folds can only be an integer starting from 2. In this section we will discuss much more flexible methods.

In *repeated inductive conformal e-prediction* (RICEP), we choose the size $m \in \{1, \dots, l-1\}$ of the proper training set arbitrarily (this is a parameter of the algorithm). We then split the training set N times randomly into disjoint proper training set of size m and calibration set of size $l - m$, for each split compute the corresponding e-value (23), and finally average the resulting e-values over the splits. Therefore, there are two new parameters, m and N. By

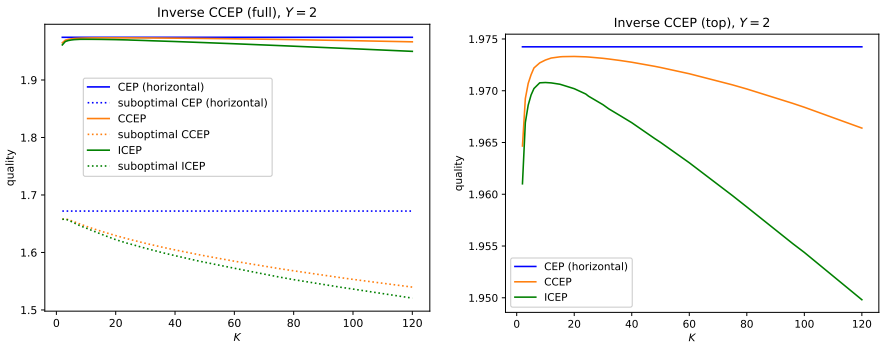

Figure 6: Results for the inverse CCEP: the full picture is in the left panel and its top part is in the right one; $l = 12{,}000$, $Y = 2$ (binary classification), and $\alpha = 0.5$. The average value of the AFES criterion over 10,000 iterations is shown as function of the number of folds (the 25 smallest divisors of l).

the law of large numbers (under mild regularity conditions), there will be not much randomness in RICEP when N is large (unlike in ICEP); RICEP (repeated ICEP) are "pseudo-deterministic". In Figures 7–9, we mention N after "RICEP" in the legend, so that, e.g., "RICEP 100" stands for RICEP with $N = 100$.

In *balanced ICEP*, or *BICEP*, the size of the calibration set is chosen randomly from $\{1, \dots, l-1\}$, and then the calibration set of that size is also chosen randomly; the proper training set is then determined uniquely as the complement of the calibration set. In *semi-BICEP*, the size of the calibration set is chosen randomly from $\{1, \dots, \lfloor l/2 \rfloor\}$ (so that the proper training set is at least as large as the calibration set) and the calibration set of that size is also chosen randomly.

Figures 7, 8, and 9 correspond to $Y = 10$, $Y = 100$, and $Y = 2$, respectively. In RICEP we use $\frac{K-1}{K}l$ and $\frac{1}{K}l$ as the sizes of proper training and calibration sets, except for Figure 9, where, naturally, the sizes are swapped. Figure 9 demonstrates that the full flexibility of BICEP (as compared with semi-BICEP) is sometimes needed, even though the performance of BICEP suffers in Figures 7 and 8 because of its flexibility.

In Figures 7–9, the plots for RICEP N, BICEP N, and semi-BICEP N move up as N increases, so that the best quality is attained when N is large. This can be understood in terms of Jensen's inequality, which implies

$$\ln \frac{e_1 + \dots + e_N}{N} \geq \frac{\ln e_1 + \dots + \ln e_N}{N}. \tag{25}$$

The difference between the left-hand and the right-hand sides is known as the *Jensen gap*, and it is especially substantial when $e_1, \dots, e_N$ are very different; in the first approximation, the Jensen gap is proportional to the variance of $e_1, \dots, e_N$.

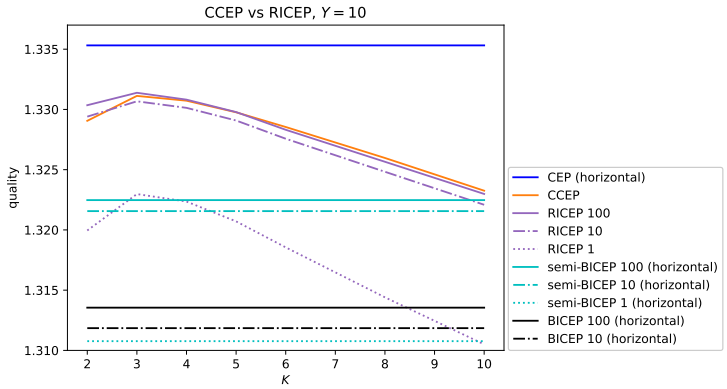

Figure 7: Results for CCEP vs RICEP; $l = 12{,}000$, $Y = 10$, and $\alpha = 0.5$. The average value of the AFES criterion over 10,000 iterations is shown as function of the number of folds (the 7 smallest divisors of l).

Inequality (25) explains why RICEP 10 and RICEP 100 are higher than RICEP 1. We can also apply (25) to 10 in place of N and to averages of 10 e-values in place of $e_1, \dots, e_{10}$, which explains why RICEP 100 is higher than RICEP 10. The Jensen gap between RICEP 100 and RICEP 10 is not as big as the Jensen gap between RICEP 10 and RICEP 1 since the averages of 10 e-values are less variable than the original e-values. Of course, this argument is also applicable to BICEP and semi-BICEP.

The Jensen gap also works in favour of CCEP as compared with ICEP, since CCEP involves averaging of the K e-values coming from different folds. (And this explains why the difference between the CCEP and ICEP lines tends to increase as K grows in Figures 5 and 6.) Therefore, RICEP can be expected to be comparable with CCEP for $N \approx K$; however, CCEP look slightly more efficient than that, especially in Figure 7. In Figures 8 and 9 the lines for CCEP and RICEP 100 almost coincide.

Remark 6. The Jensen gap is an artefact of our way of measuring predictive efficiency that involves logarithms (see (2)). In principle, it does not mean that an improvement from, say, RICEP 1 to RICEP 100 in Figure 7 is in any sense "real". (How real it is is a direction of further research.) What is undoubtedly real is that the averaging in RICEP 100 produces much more stable results. An example of instability of e-values is shown in Figure 10, which results from an experiment in which a random dataset of 12,000 observations with $Y := 10$ is generated from the multinomial distribution whose parameter $\theta \in \Theta$ is generated from $\mathrm{Dir}_{0.5}$ (i.e., Jeffreys's prior). The parameter vector θ is generated only once, and in our experiment $\theta_3 \approx 7.93 \cdot 10^{-5}$ is its smallest component (and so $y = 3$ is the least likely observation); the dataset is also generated only once. Then the dataset is randomly split 1000 times into proper training and

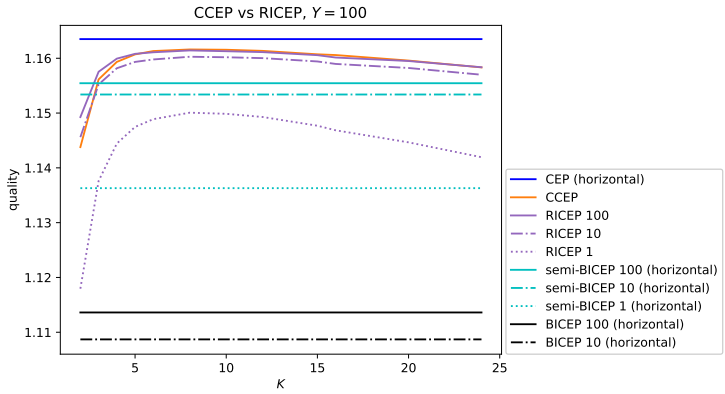

Figure 8: The analogue of Figure 7 for $Y = 100$ and the 12 smallest divisors of l as K.

calibration sets and the corresponding e-values (23) are computed for the least likely observation $y = 3$ (leading to high e-values). We take $m := 8000$ as the size of proper training sets, which is a reasonable value in view of the top panel of Figure 3. The histogram (in blue) shows an anomalous behaviour of the e-values; they are clearly separated into two clusters, above and below 800. The red vertical line shows the average of the 1000 e-values, and it is between the two clusters; in general, the average is still a random variable, but it is much more concentrated and less anomalous.

8 Conclusion

This paper proposes new ways of using e-values in conformal prediction. An important advantage of e-values over p-values is that they do not require smoothing (represented by the random number τ in (5), (15), etc.). Deterministic p-values perform particularly poorly in the context of this paper, where we have no objects in a classification problem, which leads to a large number of ties (especially for a small Y). Another advantage is that CCEP have the same property of validity as CEP, while for CCP their (provable) property of validity becomes weaker. The advantage first described in this paper is that CCEP can be modified to RICEP and BICEP to address some limitations that are shared between CCEP and CCP.

These are some advantages and disadvantages of RICEP as compared with CCEP, with "+" indicating advantages and "−" disadvantages:

+ RICEP with a large number N of repetitions is pseudo-deterministic, while CCEP may be more volatile;

+ RICEP is much more flexible; we do not need to choose the size of cali-

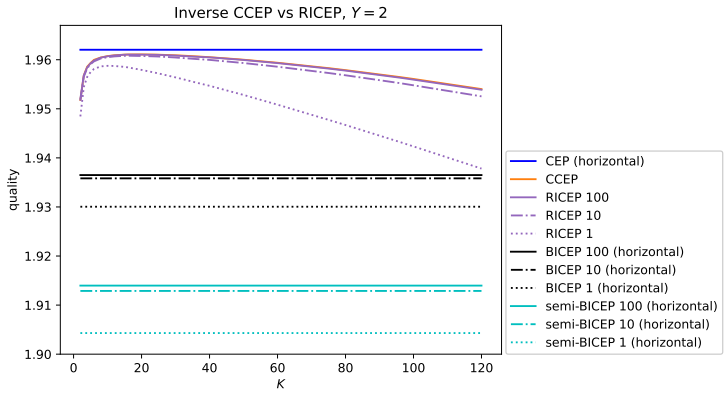

Figure 9: The analogue of Figure 7 for $Y = 2$, inverse CCEP, and the 25 smallest divisors of l as K.

bration sets from divisors (or at least approximate divisors) of the size of the training set; and we can easily combine different sizes, as in BICEP;

\+ in RICEP, the Jensen gap is under our control and is independent of the split proportion, unlike CCEP;

− judging by Figures 7–9, CCEP may produce slightly better results for the same numbers of repetitions (K in the case of CCEP and N in the case of BICEP).

In summary, this paper's experimental results show that RICEP has important advantages over CCEP and should be used when it is clear *a priori* what a reasonable percentage of observations used for calibration may be. Otherwise, BICEP is safer. When some prior knowledge is available about a reasonable percentage of observations used for calibration, we should use instead a *partial BICEP*, choosing the size of the calibration set from some prior distribution on $\{1, \dots, l-1\}$ rather than from the uniform distribution.

References

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. Technical Report arXiv:2411.11824 [math.ST], arXiv.org e-Print archive, March 2026. Pre-publication version of a book to be published by Cambridge University Press.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *Annals of Statistics*, 49: 486–507, 2021.

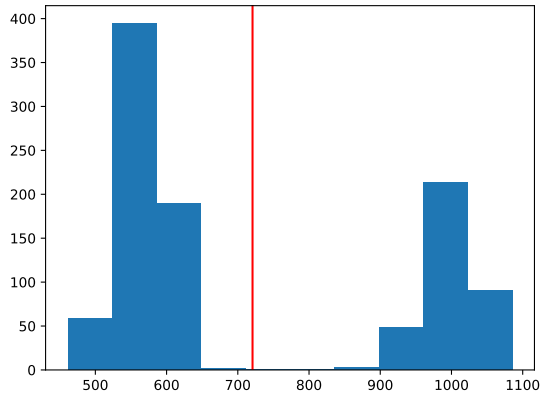

Figure 10: A histogram of RICEP e-values, as described in text.

Lars Carlsson, Martin Eklund, and Ulf Norinder. Aggregated conformal prediction. In Lazaros Iliadis, Ilias Maglogiannis, Harris Papadopoulos, Spyros Sioutas, and Christos Makris, editors, *AIAI Workshops, COPA 2014*, volume 437 of *IFIP Advances in Information and Communication Technology*, pages 231–240, Berlin, 2014. Springer.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925. Section 21.1 first appeared in the 4th edition (1932).

Sander Greenland. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *American Statistician*, 73(S1):106–114, 2019.

Sander Greenland. Contribution to the discussion of "Safe testing" by Grünwald, de Heide, and Koolen. *Journal of the Royal Statistical Society B*, 86:1148–1149, 2024.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society B*, 86:1091–1128, 2024.

Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, Hoboken, NJ, fifth edition, 2012.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38: 576–601, 2023.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society A*, 184:407–478, 2021.

Vladimir Vovk. Conformal e-prediction. *Pattern Recognition*, 166:article 111674, 2025. Special Issue on Conformal Prediction and Distribution-Free Uncertainty Quantification.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107:791–808, 2020.

Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38:329–354, 2023. Revised version: arXiv:1912.13292v5.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

A Some proofs

A.1 Proof of Proposition 1

The argument of Vovk et al. (2022, Sect. 3.4.1, Remark 3.15) can be modified to cover this case as well (it is not covered by the original argument since the function $p \in [0,1] \to -\ln p$ takes value ∞ at $p := 0$, as mentioned at the end of the remark). Namely, we can replace the second displayed equation in Vovk et al. (2022, Remark 3.15) by

$$\sum_{y\in\mathbf{Y}}(-\ln p(x,y)) = \sum_{y\in\mathbf{Y}}\int_0^\infty 1_{\{u\le -\ln p(x,y)\}}\,\mathrm{d}u = \int_0^\infty \sum_{y\in\mathbf{Y}} 1_{\{p(x,y)\le \mathrm{e}^{-u}\}}\,\mathrm{d}u$$
$$= \int_0^\infty \left(|\mathbf{Y}| - \left|\Gamma^{\mathrm{e}^{-u}}(x)\right|\right)\,\mathrm{d}u = \int_0^1 \left(|\mathbf{Y}| - |\Gamma^\epsilon(x)|\right)\frac{\mathrm{d}\epsilon}{\epsilon}. \quad (26)$$

The last integral converges since $|\mathbf{Y}| - |\Gamma^\epsilon(x)| = 0$ for sufficiently small $\epsilon > 0$.

A.2 Proof of Proposition 2

Proposition 2 can be easily deduced from Vovk et al. (2022, Theorem 3.1) and the modification (26) of the argument in Vovk et al. (2022, Remark 3.15). Under our Bayesian model and conditionally on the training set, we are in the idealized setting of Vovk et al. (2022, Sect. 3.1.4) with the data-generating distribution equal to the predictive distribution for the test observation. According to Vovk et al. (2022, Theorem 3.1), the probability measure (4) is an N-optimal idealized conformity measure conditionally on the training set. This implies that it is N-optimal unconditionally, which in combination with (26) implies that the p-values (5) solve (6).

In general, a p-variable solving (6) is not unique since, according to Vovk et al. (2022, Theorem 3.1), we can replace the probability measure (4) used in (5) in the role of idealized conformity measure by its refinement; this will lead to the same value of (6).

A.3 Proofs of Proposition 3 and 4

To prove Proposition 3, apply Vovk (2025, Proposition 15) conditionally on the training set. Then (11) holds conditionally on the training set and, therefore, unconditionally.

To show that there is a unique solution to (11), suppose there are two different solutions. Then their arithmetic mean will provide a better value for the objective function in (11) by Jensen's inequality. The value will be strictly better since every finite sequence of observations has a positive probability under our assumptions.

The proof of Proposition 4 is a simple modification; cf. Vovk (2025, Remark 16).

A.4 Proof of Proposition 5

We need to check that the A and B defined by (15) coincide with those defined by (24) when $K = l$. Let us first do it for the two As. Since $n'_{k,y''}$ is the indicator function of y'' being the observation in the kth fold, A_k is the indicator function of $n_{k,y_k} < n_{k,y'}$. Since the last inequality is impossible for $y_k = y'$, A_k is the indicator function of the conjunction of $y_k \neq y'$ and $n_{y_k} - 1 < n_{y'}$. Therefore, the sum of A_k over k coincides with the numerator of A in (24).

Now let us do it for the two Bs. Notice that B_k is the indicator function of $n_{k,y_k} = n_{k,y'}$, this equality is true for $y_k = y'$, and this equality is equivalent to $n_{y_k} - 1 = n_{y'}$ when $y_k \neq y'$. Therefore, the sum of B_k over k coincides with the numerator of B in (15) apart from the "$+1$".