Proceedings of Machine Learning Research 105:1–23, 2019 Conformal and Probabilistic Prediction and Applications

Conformal Predictor Combination using Neyman-Pearson Lemma

Paolo Toccaceli
PAOLO.TOCCACELI@RHUL.AC.UK
Computer Learning Research Centre
Royal Holloway, Univ. of London
Egham, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo and Evgueni Smirnov

Abstract

The problem of how to combine advantageously Conformal Predictors (CP) has attracted the interest of many researchers in recent years. The challenge is to retain validity, while improving efficiency. In this article a very generic method is proposed which takes advantage of a well-established result in Classical Statistical Hypothesis Testing, the Neyman-Pearson Lemma, to combine CP with maximum efficiency. The merits and the limits of the method are explored on synthetic data sets under different levels of correlation between NonConformity Measures (NCM). CP Combination via Neyman-Pearson Lemma generally outperforms other combination methods when an accurate and robust density ratio estimation method, such as the V-Matrix method, is used.

1. Introduction

Ensembling methods, such as bagging, boosting and their more modern variants, have proved to be very effective in challenging classification problems. While these methods aggregate or refine weak predictors generally of the same type (e.g. short trees or stumps), one can also conceive of combining inherently different ML methods. In such an approach, one might exploit the possibility that the different methods perform differently in different regions of the problem domain. For instance, it may be the case that where method X tends to perform badly, method Y performs well and vice versa. While an ideal combiner that exploits fully these opportunities might be difficult to achieve, there may be still a lot of value in pursuing approximate solutions. One difficulty in combining different methods is that each may output a score which cannot be easily related to those of other methods because each is expressed on a different scale, each with a different functional relationship to the label being predicted.

The framework of Conformal Prediction Vovk et al. (2005) can offer a solution to this problem. Indeed, the notion of p-value that is central to CP provides a natural way to unify the scores produced by almost any arbitrary choice of ML algorithms.

2. Basic terminology

We recall very briefly the main concepts related to Conformal Prediction. Assuming that the training set is made up of ℓ independent identically distributed examples (iid)[1] $(x_i, y_i) \in \mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, if $x_{\ell+1}$ is a test example taken from the same distribution as the training examples, a Conformal Predictor assigns a *p-value* $p_{\bar{y}}$ to a hypothetical assignment of a label $y_{\ell+1} = \bar{y}$ to the object $x_{\ell+1}$. The CP p-value that has the following property: for any chosen $\epsilon \in [0, 1]$, the p-value of test examples (x, y) drawn iid from the same distribution as the training examples are (in the long run) smaller than ϵ with probability at most ϵ. A Conformal Predictor computes a p-value on the basis of Non-Conformity Measures (NCM). The NCM is a real-valued function $A(z; \wr z_1, \ldots, z_k \wr), A : \mathbf{Z} \times \mathbf{Z}^{(k)} \to \mathbb{R}$ that expresses how dissimilar an example appears to be with respect to a bag (or multi-set) of examples, assuming they are all iid. A Non-Conformity Measure can be in principle extracted from any Machine Learning (ML) algorithm. Although there is no universal method to derive it, a default choice is:

$$A((x, y), \wr z_1, \ldots, z_{\ell+1} \wr) := -\Delta(y, f(x))$$

where $f : \mathbf{X} \to \mathbf{Y}'$ is the prediction rule learned on $(z_1, \ldots, z_{\ell+1})$ and $\Delta : \mathbf{Y} \times \mathbf{Y}' \to \mathbb{R}$ is a measure of similarity between a label and a prediction.

Once the p-values for every possible choice of the label for a test object are computed, one can compute a multi-valued prediction (a *prediction set*) that has the *validity* property: given a significance level ϵ, the actual label of test example is not in the set no more than a fraction ϵ of the times. The validity property provides a long term guarantee on the number of errors (where "error" is defined as "actual label not in the prediction set") in the long run. If the prediction set consists of more than one label, the prediction is called *uncertain*, whereas if there are no labels in the prediction set, the prediction is *empty*. A desirable property of a CP is *efficiency*, which is loosely defined as the average size of the region set, when it is not empty.

In this paper we focus only on the *Inductive* form of CP (ICP). In the Inductive form (also referred to as *split* CP) the overall training set is split into a proper training set and a calibration set. The proper training set is used to train the underlying ML method. The function $A()$ is therefore learned once only, on the proper training set. The α_i are computed by evaluating the function $A()$ on the examples of the calibration set and on the hypothetical example. Assuming that the first m examples constitute the proper training set and the remaining $k = \ell - m$ examples the calibration set, the α_i can be formally expressed as:

$$\alpha_i = A((x_i, y_i), \wr z_1, \ldots, z_m \wr) \qquad i = m + 1, \ldots, \ell + 1$$

Once the NCM have been calculated, the p-value for a hypothesis $y_{\ell+1} = \bar{y}$ about the label of test object $x_{\ell+1}$ is defined as follows:

$$p_{\bar{y}} = \frac{|\{i = 1, \ldots, m : \alpha_i \geq \alpha_{\ell+1}\}|}{m + 1}$$

The prediction region Γ_ϵ for a test object x for a chosen significance level $\epsilon \in [0, 1]$ is the set of labels for which the p-value exceeds the significance level:

$$\Gamma_\epsilon(x) := \{y \,|\, p_y > \epsilon\}$$

1. in fact, even a weaker requirement of *exchangeability* is sufficient.

Finally, the validity property as stated above guarantees an error rate over all possible label values, not on per-label value basis. The latter can be achieved with a variant of CP, called *label-conditional CP* (a variant of Mondrian CP). The only change is in the calculation of the p-value: we restrict the α_i only to those that are associated with examples with the same label as the hypothetical label that we are assigning at the test object. So, the p-value for a hypothesis $y_{\ell+1} = \bar{y}$ about the label of test object $x_{\ell+1}$ is defined as follows:

$$p(\bar{y}) = \frac{|\{i = 1, \ldots, (\ell+1) : y_i = \bar{y}, \alpha_i \geq \alpha_{\ell+1}\}|}{|\{i = 1, \ldots, (\ell+1) : y_i = \bar{y}\}|}$$

The property of label-conditional validity is essential in practice when the CP is applied to an "imbalanced" data set, i.e. a data set in which the proportions of labels are significantly different. Empirically, one can observe that with the plain validity property, the overall error rates tend within statistical fluctuation to the chosen significance level, but the minority class(es) are disproportionately affected by errors. This property ensures that, even for the minority class, the long-term error rate will tend to the chosen significance level.

3. Combination of Conformal Predictors

In this paper the objective of the combination of Conformal Predictors is to increase efficiency, while preserving validity. In other words, we aim at reducing the average size the prediction sets, while minimising any deviations of the error rate from the chosen significance level. We will restrict our scope to binary CPs, although the methods can be extended to more than two labels. In the context of binary classification, the maximisation of the efficiency corresponds to the minimisation of the occurrence of uncertain predictions (i.e. prediction sets that contain more than one label). For clarity, the setting for the CP combination is as follows: there are d CPs (which we'll refer to as *base CPs*) and correspondingly d p-values $p^{(1)}, \ldots, p^{(d)}$ for a given label assignment to a test object. We are seeking a function $f(p^{(1)}, \ldots, p^{(d)})$ that computes a p-value that results in a valid and efficient CP. So, in the present approach the problem of combining CP is effectively one of combining p-values. This problem, i.e. obtaining a single test for a common hypothesis, has a long history beginning almost as soon as the modern framework of statistical hypothesis testing was established Fisher (1932) given its interest for many applications, e.g. meta-analysis. A survey of the field can be found in Loughin (2004). It should be observed that the context of CP differs from the more conventional setting of p-value combination. In the latter, often tens of p-values for a common hypothesis (e.g. does drug X reduce the duration of a cold?) are to combined to obtain one overall p-value. The scope of the meta-analysis ends there, when we obtain this single p-value. In the case of CP combination, we have a data set that can contain thousands examples. It is as if we were conducting thousands on meta-analyses on the same population. This gives some opportunities to calibrate the combination methods in an advantageous way.

3.1. Merging functions

The first requirement for the combination method is that validity be preserved. A comprehensive analysis of a family of combination methods that ensure validity without requiring

assumptions on the independence of the p-values can be found in Vovk and Wang (2012). Table 1 lists some of methods discussed in the study, namely minimum, maximum, arithmetic average, geometric average. The charts in the left column in Figure 1 illustrate the cumulative distribution of the p-values arising from the combination functions and the merging functions (assuming independence of the base CPs). For the combined CP to preserve the validity property, the distribution of the combined p-values must remain uniform. Consequently, in the charts the traces should follow the dashed diagonal; if the trace is below the diagonal, the predictors are conservative (i.e. leading fewer to incorrect predictions than the significance level) and vice versa. The charts in the left column show that the merging functions would result in conservative CPs when the base CPs are independent. While the absence of independence requirements bestows a wide applicability to the methods, this universal validity guarantee appears to come at the expense of efficiency.

	Combination function	Merging function
Arithmetic average	$p_{arith\_avg} = \frac{1}{d}\sum_{i=1}^{d} p^{(i)}$	$2 \cdot p_{arith\_avg}$
Geometric average	$p_{geom\_avg} = \left(\prod_{i=1}^{d} p^{(i)}\right)^{\frac{1}{d}}$	$e \cdot p_{geom\_avg}$
Min	$p_{min} = \min(p^{(1)}, \ldots, p^{(d)})$	$d \cdot p_{min}$
Max	$p_{max} = \max(p^{(1)}, \ldots, p^{(d)})$	p_{max}

Table 1: Some merging functions. These are special cases of the more general merging functions listed in Table 1 of Vovk and Wang (2012). The merging function for the Minimum is also known as Bonferroni method.

4. Combining independent base CP

After the considerations of the previous section, it seems only natural to turn one's attention to the case of independent p-values. Of course, a degree of correlation is to be expected in any practical scenario, but it may be that the methods result only in a deviation from validity that is acceptable in practical applications.

One approach to recover validity is to rely on a well-known property of univariate cumulative distributions. The methods that exploit this property are often referred to as quantile methods. If we denote by $F_X(x)$ the cumulative distribution of a random variable X, the random variable $F_X(X)$ is uniformly distributed.

$$F_X(x) = \mathbb{P}\{X \leq x\} \Rightarrow F_X(X) \sim U[0,1] \tag{1}$$

We can exploit this fact to obtain a uniformly distributed random variable out of an arbitrary function f() of d p-values if we know the distribution of that function of uniformly

distributed RVs. In fact, the distributions of minimum, maximum, arithmetic average and geometric average of d independent uniformly distributed RVs are known. Their CDFs are either expressed in closed form or are available in popular mathematical software (this is the case for the CDF of the Beta distribution, also known as the regularized beta function). The CDF are presented in Table 2. Assuming that the p-values from the base CPs are uniformly distributed and independent, we can obtain a valid CP combination by combining the p-values and then applying the distribution function. We refer to this class of methods as CDF-calibrated. Figure 1 shows the actual error rate vs. significance level for the four methods. The plots confirm that the p-values combined as prescribed above result in valid CPs (within statistical fluctuation). The effect of dependence between p-values will be discussed in section 10.1

Combination function	CDF	Comment
Arithmetic average (sum)	$\frac{1}{n!}\sum_{k=0}^{\lfloor t \rfloor}(-1)^k\binom{d}{k}(t-k)^d$	Irwin-Hall distribution
Geometric average (product)	$t\sum_{i=0}^{d-1}\frac{(-\log t)^i}{i!}$	Fisher formulaFisher (1948)
Min	`betainc()`	Beta(d, 1)
Max	`betainc()`	Beta(1,d)

Table 2: Some combination functions with known CDFs

5. Adaptive methods

The methods in the previous sections are all *a priori* methods, in the sense that the law with which the p-values are combined does not depend on the observed data. In this section we discuss a class of methods that adapt to the statistics of the observed data, albeit at the cost of having to set aside a fraction of the available observations for this purpose, thereby reducing the size of the training set for the underlying ML algorithms.

5.1. ECDF calibration

The method of ECDF calibration has been described in Toccaceli and Gammerman (2018) and before in Balasubramanian et al. (2015). It is an adaptive version of the idea put forward in section 4. Whereas in that context the distribution $F_X(x)$ was determined on the basis of the known law $f(p_1, \ldots, p_d)$, here the $F_X(x)$ is estimated as Empirical Cumulative Distribution Function on a calibration set. The calibration set on which the ECDF is estimated contains only the examples consistent with the Null Hypothesis, i.e. examples with label 0 when we are combining p_0 and examples with label 1 when we are combining p_1.

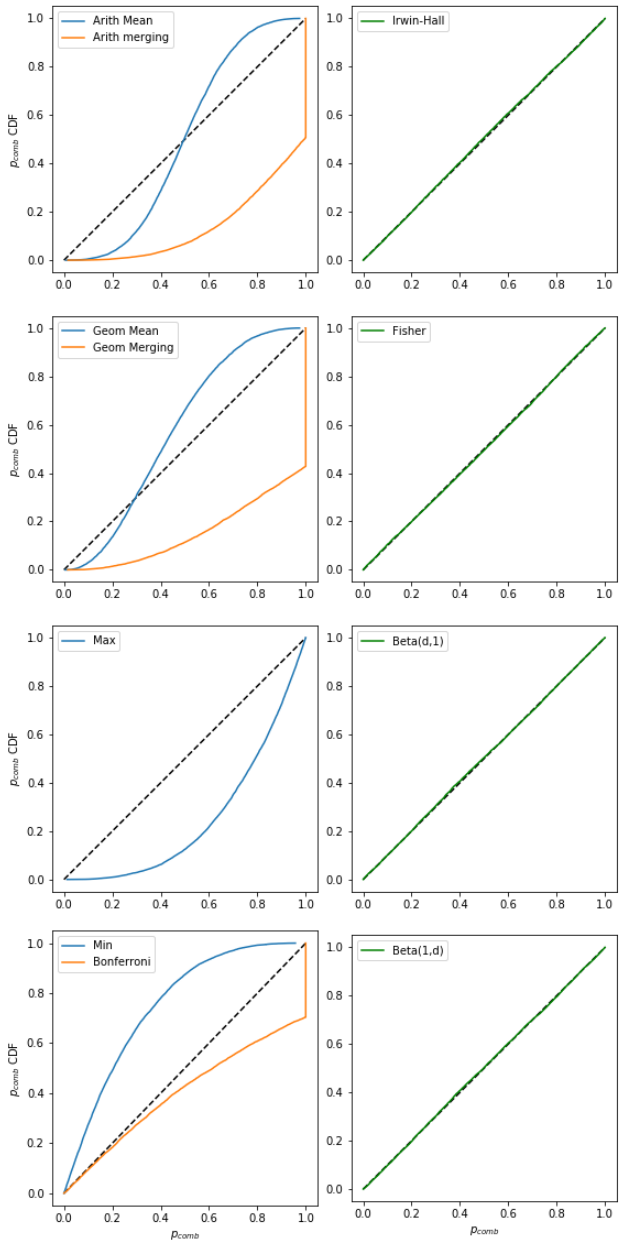


Figure 1: Comparison of validity of combination methods. Each plot shows the CDF of the combined p-value, when the base p-values are independent and uniformly distributed on $[0, 1]$. For the combined CP to be exactly valid, the trace should be the $(0, 0)$-$(1, 1)$ diagonal, indicated here with a dashed line. The left column shows the straightforward methods along with the merging variant that ensures (conservative) validity. The right column shows the CDF-calibrated versions.

Note that this method has two advantages: (a) it allows complete freedom in the choice of the law used to combine p-values, (b) it can account for the dependence in the base p-values.

5.2. Multivariate ECDF

As stated in point (b) in the previous section, the ECDF calibration allows to recover validity after combining p-values with an arbitrary law. We illustrate this point further with an adaptive combination method, i.e. one in which the combination law varies with the observed data. The method we propose here combines p-values by computing the value of multivariate joint distribution of the p-values and then calibrates it to a $U[0,1]$ with the ECDF calibration discussed in section 5.1.
More formally, given RVs $X_1, \ldots, X_d$, the joint CDF is:

$$F_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = \mathbb{P}\{X_1 \leq x_1, \ldots, X_d \leq x_d\}$$

The combination method discussed in this section can be expressed as:

$$p_{mecdf} = F_{P^{(1)},\ldots P^{(d)}}(p^{(1)}, \ldots p^{(d)}) \tag{2}$$

To perform the ECDF calibration, one must fist p_{mecdf} on a calibration set so that an ECDF of the p_{mecdf} can be computed. Then, the combined p-value is

$$p_{comb} = F_{P_{mecdf}}(p_{mecdf}) \tag{3}$$

Note that calibration step above is needed to recover validity because the CDF property stated in eq. 1 for the univariate case does not hold in the multivariate case. That is, if $X^{(1)}, \ldots X^{(d)}$ are independent uniformly distributed RVs, $F_{X^{(1)},\ldots X^{(d)}}(X^{(1)}, \ldots X^{(d)})$ is not distributed according to $U[0,1]$ [2]

Such CDF is unknown, but we can estimate it by computing the Multivariate ECDF on calibration data.

$$F_{\ell_{cal}}(x^{(1)},\ldots,x^{(d)}) = \frac{1}{\ell_{cal}} \sum_{i=1}^{\ell_{cal}} \prod_{k=1}^{d} \theta(x^{(k)} - x_i^{(k)})$$

6. Combination via Neyman-Pearson Lemma

The Neyman-Pearson Lemma is a result in Statistic Hypothesis Testing on which basis it is possible to define a test statistic and a threshold so that the resulting significance test has Uniform Maximum Power (UMP). Here, *power* is defined as the probability to reject correctly the Null Hypothesis H_0.

This can be applied to CP by noticing that when we calculate, say, p_0, we assume as Null Hypothesis that the label is 0 and compute a p-value for the test object under this assumption. The p_0 can be interpreted as the probability of drawing from the same set as the calibration set an example that is as or more contrary to the hypothesis of randomness as the hypothetical test example.

2. The distribution of $F_{X^{(1)},\ldots X^{(d)}}(X^{(1)}, \ldots X^{(d)})$ is referred to as Kendall distribution function Genest and Rivest (2001).

The Neyman-Pearson Lemma is particularly relevant to CP combination because it can optimise efficiency (i.e. results smaller prediction sets). To see this, consider that with higher power one rejects more often H_0 when indeed it should be rejected. Consider also that the prediction set contains all the hypothetical label assignments that could not be rejected at the chosen significance level (as it contains all the labels y for which $p_y > \epsilon$). This means that the higher the power of a test, the less likely it will be that the prediction set will contain incorrect labels. Note also that, in so for as validity is satisfied, the rate at which the correct label is in the prediction set is equal to the significance level.

6.1. Statement of the Neyman-Pearson Lemma

The most powerful test between two simple hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ is the one that uses as test statistic the likelihood ratio:

$$\Lambda(x) := \frac{\mathcal{L}(\theta_0|x)}{\mathcal{L}(\theta_1|x)} \tag{4}$$

and as threshold the value η that satisfies

$$\epsilon = \mathbb{P}\left[\Lambda(X) \leq \eta \mid H_0\right] \tag{5}$$

where ϵ is the significance level.

6.2. Application to Combination of Conformal Predictors

Let's assume that we have k separate CPs, each using some different underlying ML algorithm, producing for a test object the k p-values $p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)}$ for the hypothetical label assignment $y = \bar{y}$. The likelihood ratio $\Lambda(p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)})$ can be computed as:

$$\Lambda_0(p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)}) = \frac{\mathbb{P}\left[p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)} \,\middle|\, y = \bar{y}\right]}{\mathbb{P}\left[p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)} \,\middle|\, y \neq \bar{y}\right]} \tag{6}$$

If we denote by $F_{\Lambda_0}(\lambda)$ the (cumulative) distribution function of $\Lambda(p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)})$ given H_0, the p-value for the combination is then obtained as:

$$p_{\bar{y}}^{(\mathrm{NP})} = F_{\Lambda}(\Lambda(p_{\bar{y}}^{(1)}, \ldots, p_{\bar{y}}^{(k)})) \tag{7}$$

To justify the last equation, consider that eq. 5 can be expressed also as $\epsilon = F_{\Lambda_0}(\eta)$ In principle, there is no need to compute explicitly η. An alternative way of interpreting eq. 5 is saying that the hypothesis should be rejected when the value of cumulative distribution function for the hypothetical example is less than or equal to ϵ. By computing $p_{\bar{y}}^{(\mathrm{NP})}$ according to eq. 7 we achieve precisely that.

7. Implementation of Neyman-Pearson Combination

The method just described promises optimal efficiency, with no assumptions on the absence of correlation or even dependence among CPs. In principle, this method should outperform any other combination method, at least in terms of efficiency. However, the method revolves around the ratio of the likelihoods under the null and under the alternative hypothesis. The estimation of a density, let alone a density ratio, is an ill-posed problem as pointed out as early as Vapnik (1995). The difficulty of this estimation is further compounded by its multivariate nature. It is therefore important to investigate the question of how the method actually performs in practice, especially when only limited amounts of noisy data are available. The performance of the methods chosen for the estimation of the density ratio is critical for this method to realise its full potential.

Three approaches are described: Naïve Neyman-Pearson, Multivariate Histogram, and V-Matrix.

7.1. Naïve Neyman-Pearson

To apply the method described in section 6.2, one needs to compute the likelihood $\mathcal{L}(\theta_1|x)$, that is, the density $\mathbb{P}[X|\theta_1]$ evaluated at x. In particular, we are looking for the likelihood for the joint event of $p_1, p_2, \ldots, p_k$.

To make the estimation more tractable, one approach is to make the *naïve* assumption that the p-values are independent (this is analogous to the independence assumption made in Naïve Bayes). So the density of the joint event can be calculated as the product of the densities of each of the simple events.

Consequently, a method that we refer to here as Naïve Neyman-Pearson obtains first an estimate of the (marginal) density of each of the p-values and then simply calculates the likelihood for the joint event as product of those densities. The likelihood $\mathcal{L}(\theta_0|p)$ for each p-value is 1 by construction. So, the NPL statistic can be expressed as:

$$\Lambda(X) = \frac{1}{\prod_{i=1}^{k} f_1(p_i)} \qquad \text{where} \quad X = (P_1, P_2, \ldots, P_k)$$

To obtain the combined p-value, we start from recalling that the threshold η is chosen so that the significance level ϵ is:

$$\epsilon = \mathbb{P}[\Lambda(X) \leq \eta | H_0]$$

We can therefore transform the statistic value λ into a p-value by applying to it the CDF of the NPL statistic evaluated on the H_0 cluster.

$$p_{\text{comb}} = CDF_{H_0}(\lambda)$$

where

$$CDF_{H_0}(\lambda) = \mathbb{P}[\Lambda(X) \geq \lambda \,|\, H_0]$$

Note that this ensures that the p-value for the Null Hypothesis be uniformly distributed.

One obvious limit of this approach is that it is hardly ever the case that the p-values of the base CPs are independent.

7.2. V-Matrix

To account fully for an arbitrary dependence between p-values one has to attempt to estimate the multivariate joint density ratio. Density estimation is central to statistical inference and the problem has been studied for decades, resulting in a variety of methods. A rigorous approach was proposed by Vapnik first inVapnik (1995), and then in Vapnik et al. (2015) and Vapnik and Izmailov (2015). The method is referred to as V-Matrix method. We'll recap just the key points here and refer the reader to papers just cited for the full derivation and all the attendant details.

7.2.1. Direct Constructive Setting

Let's consider first the problem of density estimation. Let's assume that we are given ℓ d-dimensional samples $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})$ from a (cumulative) distribution $F(x)$. We are seeking a density $f(x)$ such that:

$$\int_{-\infty}^{x} f(t)dt = F(x)$$

The distribution F(x) is unknown, but from the samples we can compute the empirical cumulative distribution

$$F_\ell(x^{(1)}, \ldots, x^{(d)}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{k=1}^{d} \theta(x^{(k)} - x_i^{(k)})$$

where $\theta()$ is the step function defined as:

$$\theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

A key result in Vapnik-Chervonenkis theory guarantees that the uniform convergence of $F_\ell(x)$ to $F(x)$ as $\ell \to \infty$ is fast:

$$\mathbb{P}\left[\sup_x |F_\ell(x) - F(x)| > \epsilon\right] \leq 2\exp(-c^*\epsilon^2\ell)]$$

where $c^* = 1 - \frac{(d-1)\log \ell}{\epsilon^2 \ell}$.
In other words, the cumulative distribution function can be estimated from a limited amount of samples with a relatively small error. The direct constructive setting consists in estimating the density $f()$ as solution of the integral equation using the approximation given by empirical distribution function $F_\ell(x)$ in place of the actual but unknown $F(x)$.

7.2.2. Density Ratio

In the case of the density ratio estimation, we are given ℓ_{num} d-dimensional samples $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})$ from a (cumulative) distribution $F_{num}(x)$ and ℓ_{den} d-dimensional samples $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})$ from a (cumulative) distribution $F_{den}(x)$. We are seeking a density $r(x)$ such that:

$$\int_{-\infty}^{x} r(t)dF_{den}(t) = F_{num}(x) \tag{8}$$

Analogously to the density estimation case above, we estimate $r(x)$ by solving the integral equation after replacing $F_{num}(x)$ and $F_{den}(x)$ with their empirical counterparts, $F_{\ell_{num}}(x)$ and $F_{\ell_{den}}(x)$

7.2.3. Solution via regularization method

The integral equations arising from the direct constructive setting are *ill-posed*, in the sense that their solutions are not stable: informally stated, small changes to the right-hand side can result in significant changes to the solution. In the case of the density ratio problem, the difficulty is compounded by the fact that not only the right-side, but the left side are approximately defined. Problems of this nature are called stochastic ill-posed problems.

The method proposed in Vapnik et al. (2015) is to seek the function $r(x)$ that minimizes the sum of the L_2 distance (in a chosen metric space E) between $F_{\ell_{num}}()$ and the left-hand side of eq. 8 and a regularization term. The solution is sought in a Reproducing Kernel Hilbert Space of kernel $K(\cdot,\cdot)$ and has the form:

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(X_i, x) = A^T \mathcal{K}(x) \tag{9}$$

where $A = (\alpha_1, \ldots, \alpha_{\ell_{den}})^T$ and $\mathcal{K}(x)$ is a vector of $K(X_i, x), i = 1, \ldots, \ell_{den}$.
The functional to minimize is expressed as:

$$A^T K V K A - 2\left(\frac{\ell_{den}}{\ell_{num}}\right) A^T K V^* \mathbf{1}_{\ell_{num}} + \gamma A^T K A \tag{10}$$

where K is the $(\ell_{den} \times \ell_{den})$ matrix with elements $K(X_i, X_j)$, $i, j = 1, \ldots, \ell_{den}$ and V and V^* are matrices that reflect the geometry of the observed data. In addition, the solution should take non-negative values and should integrate to 1. These two constraints are expressed in terms of the observed data as:

$$KA \geq \mathbf{0}_{\ell_{den}} \tag{11}$$

$$\frac{1}{\ell_{den}} A^T K V^* \mathbf{1}_{\ell_{num}} = 1 \tag{12}$$

7.2.4. V-Matrix

The $(\ell_{den} \times \ell_{den})$ V matrix and $(\ell_{num} \times \ell_{den})$ V^* matrix mentioned in eq. 10 have elements

$$V_{i,j} = \int \theta(x - X_i)\theta(x - X_j)\sigma(x)d\mu(x). \tag{13}$$

where $\sigma(x)$ and $\mu(x)$ are respectively a weighting function and a measure that arise in the definition of distance in the metric space E. $\sigma(x)$ and $\mu(x)$ allow to craft the definition of distance to suit the specific statistical inference problem. With the choice of $\sigma(x) = 1$ and μ the uniform measure, assuming that data belongs to the upper-bounded interval $[-\infty, u]$,

$$V_{i,j} = \prod_{k=1}^{d} \left(u - \max\left\{X_i^{(k)}, X_j^{(k)}\right\}\right) \tag{14}$$

8. Experiments with synthetic data

In Heard and Rubin-Delanchy (2018), the Neyman-Pearson Lemma is used in combination with the common assumption that the distribution of p-value under the alternative hypothesis is of the form Beta(a, b) with $a \in (0, 1]$ and $b \in [1, +\infty)$. In particular, the paper claims that Fisher's method is the most powerful when the alternative hypothesis $p \sim \text{Beta}(0.5, 1)$. One wonders how warranted this common Beta(a, b) assumption is (see also Sellke et al. (2001)), in particular in the specific context of Conformal Predictors. On a purely intuitive basis, it is not outside the realm of possibility that there may be some deeper relationship between the distribution of CP p-values, which can be seen as rank transformed scores, and the order statistics of the uniform distribution which indeed happen to be Beta-distributed random variables. However, in the present study it was felt that it would be more realistic to generate p-values from appropriate distributions of NCMs, rather than directly.

8.1. A realistic model of NCMs

The NCMs can in principle be obtained from a very wide variety of ML algorithms. One can model the distribution of NCMs as a mixture of two distributions, one for NCMs for examples of one class and the other for the NCMs of the other class. Figure 2 shows an example of the histogram of the distribution that arise in a real-life case. Of course, markedly different distributions can arise from different methods, but the example suggests that it might be relevant to study the case in which the scores for the two classes are distributed as two Gaussians.

Throughout the rest of the paper, we assume that the NCMs are derived from the scores simply by a monotone transformation, e.g. changing the sign, as needed.

In Figure 3 four main cases are identified. In all four cases, the Gaussian distributions have mean -1 and +1. What differs is variance, which reflects the relative uncertainty of the prediction for each class. The four cases allow us to study the effect of larger and asymmetric overlaps.

9. The distribution of p-values under the Alternative Hypothesis

The distribution of p-values under the Null Hypothesis is uniform by construction. The distribution of p-values under the Alternative Hypothesis is determined by the distribution of the Nonconformity Measure. If we denote as $P_0(\alpha)$ the CDF of alphas under H_0 and $p_1(\alpha)$ the PDF for the NCM under H_1, the p-values can be viewed as Random Variables obtained as:

$$\mathbb{P}\left[\alpha_0 \geq \alpha_1\right] = 1 - P_0(A_1)$$

where A_1 is a random variable whose realisations are the NCM α_1 under the Alternative Hypothesis H_1.

For the four cases shown in Figure 3 it is possible to express in closed form the PDF of the p-values under the Alternative Hypothesis. The equations are given in the table in Figure 4[3].

3. The symbolic expressions were computed using Mathematica®.

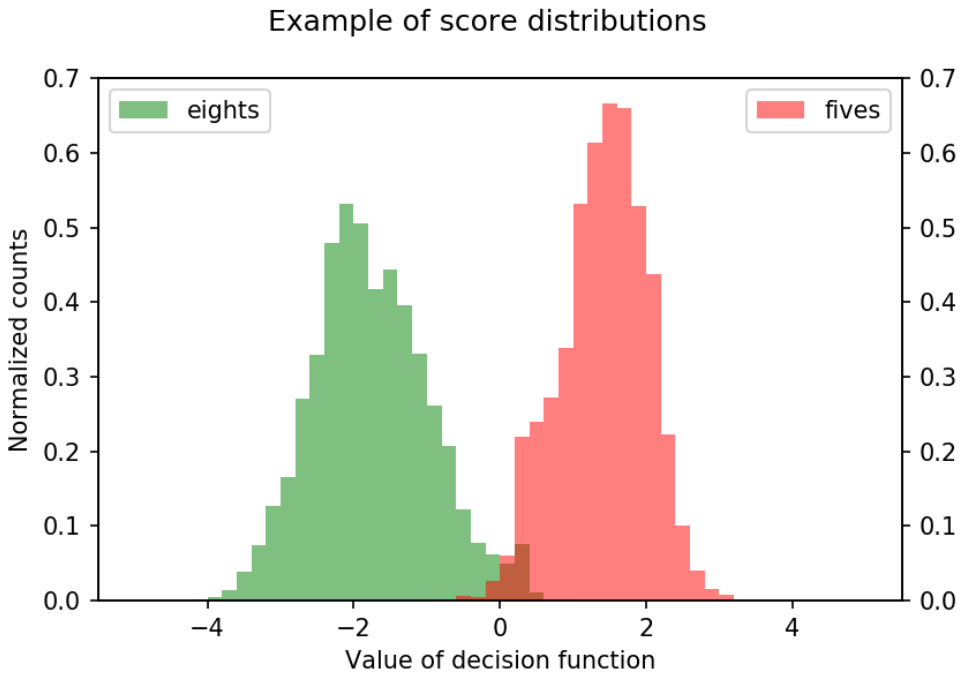


Figure 2: Example of score distribution from a real-life dataset. These scores were obtained as SVC decision function values. The SVC was trained to classify a dataset containing 28x28 images of handwritten "5" and "8" digits from the well-known MNIST dataset.

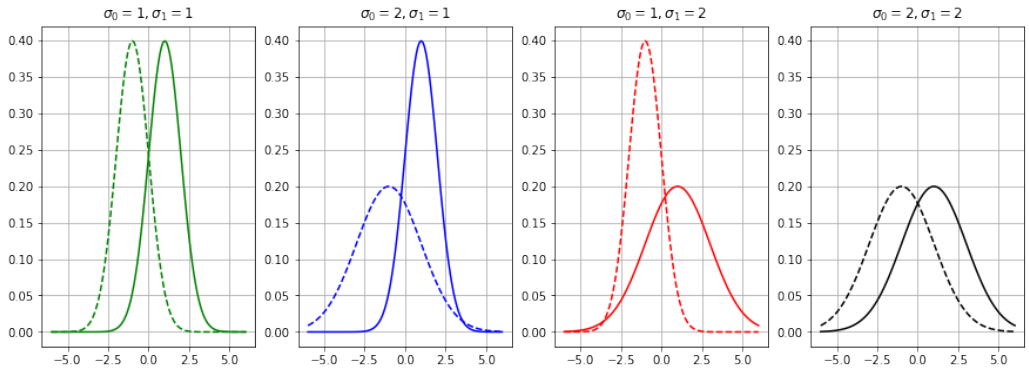
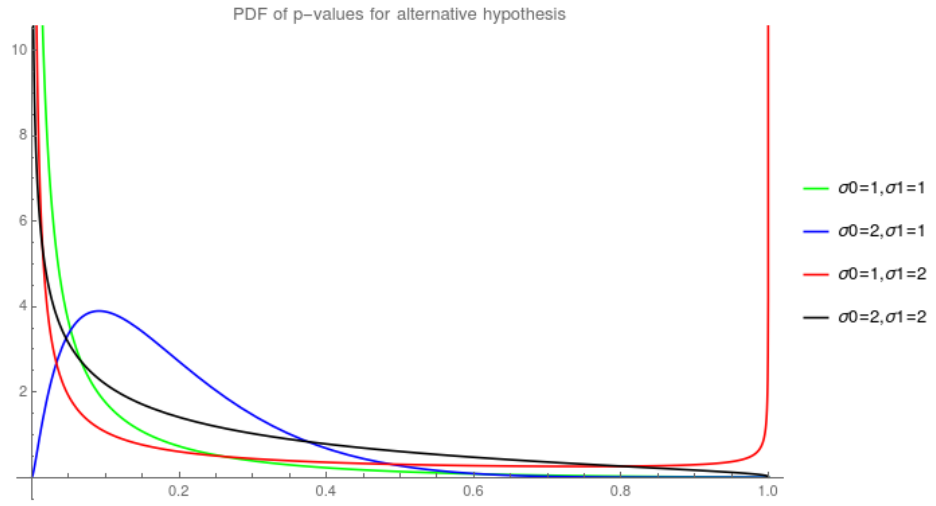


Figure 3: Cases of NCM distributions. The dashed lines correspond to H_0 and the solid lines to H_1. The cases at the extreme left and extreme right assume that the underlying method had the same uncertainty in classifying test examples of either label. The cases differ in the amount of "overlap". The plots in the middle refer to cases in which the classifier had more uncertainty for the Null Hypothesis label (blue) and less uncertainty for the Null Hypothesis label (red)



case	σ_0	σ_1	**PDF of p-values under** H_1
green	1	1	$\exp\left(-2\sqrt{2}\,\text{InvErfc}(2-2x)-2\right)$
blue	2	1	$2\exp\left(-3\,\text{InvErfc}^2(2-2x)-4\sqrt{2}\,\text{InvErfc}(2-2x)-2\right)$
red	1	2	$\frac{1}{2}\exp\left(\frac{1}{4}\left(3\,\text{InvErfc}^2(2-2x)-2\sqrt{2}\,\text{InvErfc}(2-2x)-2\right)\right)$
black	2	2	$\exp\left(\sqrt{2}\,(-\text{InvErfc}(2-2x))-\frac{1}{2}\right)$

Figure 4: The PDF of the p-values under H_1.

It is interesting to observe that while the black and green traces could be in qualitative agreement with the common assumption mentioned earlier in section that the Alternative Hypothesis p-values follow some form of Beta distribution, the blue and the red traces show a different behaviour. But it could be argued that the vertical asymptote at $p = 1$ for the red trace and the behaviour near $p = 0$ for the blue line have to do with the possibly unrealistic long tails of the wider Gaussian. Both occurrences can be explained by the fact that for sufficiently small and sufficiently large the PDF of the Gaussian of larger variance has larger values than that of the Gaussian of lower variance.

10. Experimental results

The CP combination methods discussed in the previous sections were applied to **two** base CPs, denoted here with CP_a and CP_b. Calibration sets and test sets had both 5,000 examples, with the two classes being represented in equal proportions. (Obviously, there is no proper training set as the NCMs are "simulated").

The code was entirely written in Python with the help of Jupyter Notebooks, using `numpy`, `scipy`, `numba` and `scikit-learn`. The V-Matrix implementation used the `cvxopt` package for the solution of the Quadratic Programmming problem.

We assumed that in each CP the NCMs for examples of the two labels could be distributed in the one of four possible cases discussed in the previous section, namely:

- $\sigma_0^2 = 1, \sigma_1^2 = 1$
- $\sigma_0^2 = 1, \sigma_1^2 = 4$
- $\sigma_0^2 = 4, \sigma_1^2 = 1$
- $\sigma_0^2 = 4, \sigma_1^2 = 4$

The total number of pairings of cases, discounting symmetries, is $\frac{(n+1)n}{2} = \frac{5 \cdot 4}{2} = 10$. For each of these pairings, we then used 3 different settings of correlation between the NCMs of CP_a and CP_b. We generated NCM sets with covariance 0 (in fact, they were not only uncorrelated, but independent), covariance 0.8, and covariance -0.8. Figure 5 illustrates the NCMs and the resulting p-values for the 3 different covariance values in the case with $\sigma_0 = 2$, $\sigma_1 = 2$. From the NCMs, p-values for the test objects were computed according to the MICP framework. The p-values were then used to compute the prediction sets and the results, in turn, were summarised into confusion matrices, which provide counts of correct, incorrect, empty, and uncertain predictions. To assess validity, the confusion matrices were computed for different significance levels, namely 0.01, 0.05, 0.1, 0.2.

As stated in section 3, the objective considered in this paper is to improve efficiency, while preserving validity. So, the analysis that follows will focus on these two properties.

The results for the $10 \times 3 \times 4 = 120$ cases (each repeated 25 times) are summarized in Tables 3, 4, 5, 6. In Figure 6 we show one representative case out of the 120.
In the charts, the entries are grouped as follows:

base predictors: The base CPs, identified as "a" and "b"

reference: the theoretical optimal methods under the assumption of independence, listed as "Naive Neyman-Pearson Ideal"

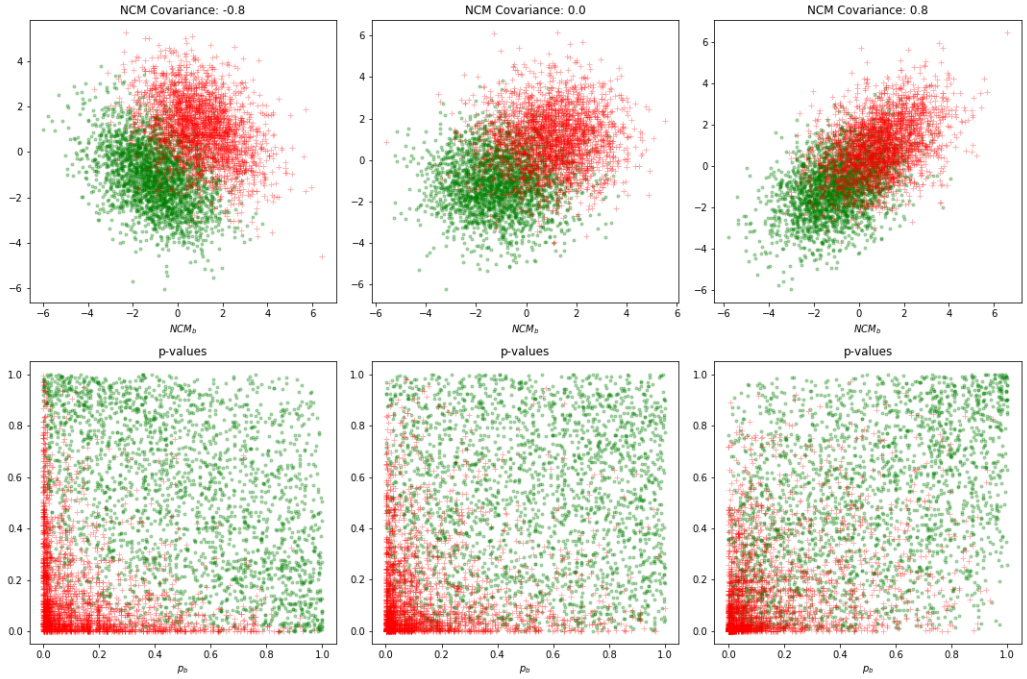


Figure 5: Example of NCM sets with same variance, but different covariance. The red crosses correspond to the data points with label 1 and the green dots to the data points with label 0. Note the variance and covariance referred to here are those of each component of the Gaussian mixture, i.e. between the NCMs for set a and set b for label 0 and the NCMs for set a and set b for label 1. The top row shows the NCMs, the bottom row the resulting MICP p-values.

basic methods: arithmetic average, geometric average, maximum, minimum

merging functions: methods discussed in sec. 3.1 which guarantee (conservative) validity

CDF calibrated methods: arithmetic average (CDF), geometric average (CDF), maximum (CDF), minimum (CDF)

ECDF calibrated methods: arithmetic average (ECDF), geometric average (ECDF), maximum (ECDF), minimum (ECDF)

adaptive methods: Multivariate ECDF, Naive Neyman-Pearson (histogram), V-Matrix

An orange background is applied to the groups as a visual reminder of their significant deviations from validity. Table 3 reports the average fraction of uncertain predictions (inversely related to efficiency) for one value of significance level, namely $\epsilon = 0.05$. In Tables 4,5,6, we present the rankings in terms of efficiency, averaged over the 25 repetitions, and disaggregated by significance level, correlation, and variances, respectively. Also, in these tables we removed the methods that deviate significantly from validity so that the ranking is fairer.

10.1. Findings

The analysis of the results confirms the observations made earlier while describing the methods. More specifically, taking Figure 6 as a representative case, we can in fact observe that:

1. all the basic methods (arithmetic average, geometric average, maximum, minimum) exhibit deviation from validity
2. the merging functions are extremely conservative, perhaps with the exception of Bonferroni for low value of significance level.
3. The CDF calibrated methods are indeed valid when the base predictors are independent, but exhibit different forms of deviation in the presence of correlation.
4. ECDF calibrated methods exhibit small deviation form validity also in the presence of correlation.

The basic methods and the merging functions will not be discussed further as their deviation from exact validity defeats the purpose of CP combination considered in this study.

Turning now our attention to efficiency, the results shown in Tables 3,4,5,6 support the following findings:

1. In the case of positive correlation, there is not much efficiency improvement in combining (refer to Table 3) This may be intuitively justified by observing that if the p-value are strongly correlated, they convey the same information. Bringing this to an extreme, we would not expect to see any improvement by combining a CP with itself. Conversely, negative correlation offers the best opportunities for efficiency gains.

2. The accuracy and robustness of density ratio estimation is critical to the success of the application of the Neyman-Pearson method. When a simple method such as histogram is used, the N-P method often fails to improve CP efficiency. The improvements require the use of a more accurate and robust method such V-Matrix.

3. The superiority of V-Matrix method fails to manifest itself fully for very low values of the significance level (refer to Figure 4). This is indicative of inaccuracy in the low end of the prediction range (i.e. for values close to 0). This may be overcome with a better choice of kernel. In this study, the Gaussian RBF kernel was chosen after some experiments with Polynomial and INK-Spline Kernel failed to provide encouraging results. It is possible that a kernel on a [0,1] support and with a better suited functional form might perform better.

4. The Multivariate ECDF method performs well and it is competitive with respect to V-Matrix. This is particularly interesting given the simplicity of the method and the absence of any parameters that need optimisation (the V-Matrix method has a regularisation parameter and, possibly, also a kernel parameter).

11. Future directions

This study focused on the combination of just 2 CPs. It would be worthwhile to investigate how the performance varies when more than 2 CPs are combined. The curse of dimensionality might affect density ratio estimation to an extent that would limit the advantages of the N-P method. Also, imbalance, i.e. the different proportion of examples of the two classes, might affect negatively the adaptive methods. A natural application to study is in Cross-Conformal Predictors Vovk (2015). More in general, a comparison should carried out on real-world data sets and a variety of underlying ML methods to gain a better understanding of their merits and limitations.

12. Conclusions

When the objective of CP combination is efficiency improvement while preserving exact validity, the Neyman-Pearson Lemma can be used to obtain a combination method that offers the best efficiency at the cost of using part of the training set for calibration purposes. The critical component of the method is density ratio estimation and we showed on a realistic synthetic data set that an accurate and robust method such V-Matrix can be used successfully. We also showed that other approximate methods exist that provide, with much less complexity, only slightly inferior results.

13. Acknowledgements

The author gratefully acknowledges AstraZeneca for the financial support of this research through grant R10911-10 “Automated Chemical Synthesis”. The author also thanks Prof. Alexander Gammerman and Prof. Zhiyuan Luo for helpful discussions and the anonymous reviewers for their constructive comments.

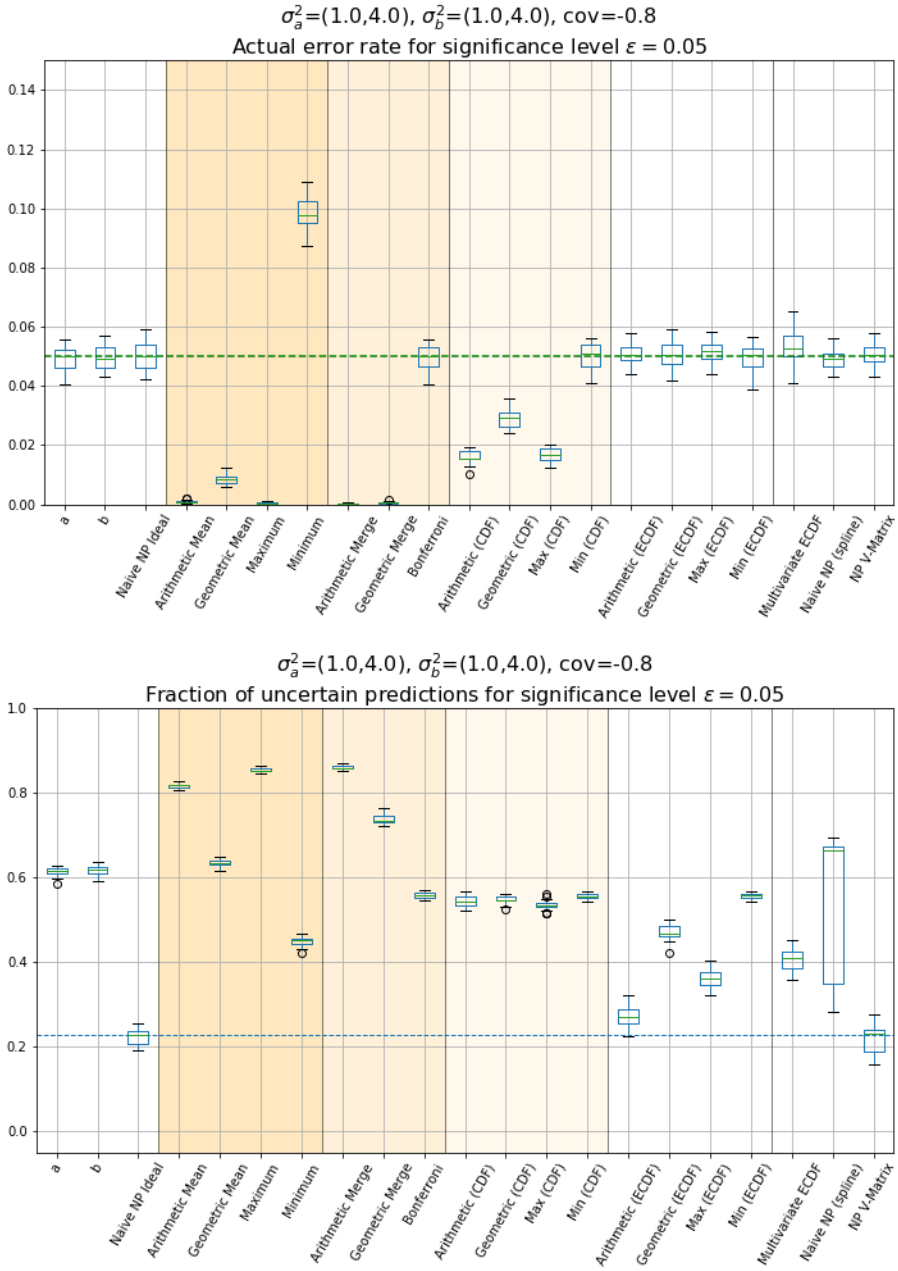


Figure 6: Boxplots for the error rate and for the fraction of uncertain predictions for one "representative" scenario. In top chart, which refers to error rate, the methods in the shaded areas show significant validity deviations (compare with dashed green line, which corresponds to the significance level). In the bottom chart, which refers to the fraction of uncertain predictions, we can see that NP V-Matrix outperform all the other methods. (The green line is the median rate for "Naïve NP" which we take here as reference.)

	σ_a^2=(1.0,1.0), σ_b^2=(1.0,1.0)			σ_a^2=(1.0,1.0), σ_b^2=(1.0,4.0)			σ_a^2=(1.0,1.0), σ_b^2=(4.0,1.0)			σ_a^2=(1.0,1.0), σ_b^2=(4.0,4.0)			σ_a^2=(1.0,4.0), σ_b^2=(1.0,4.0)		
	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800
a	0.313	0.310	0.313	0.313	0.309	0.312	0.310	0.312	0.309	0.303	0.315	0.316	0.617	0.617	0.611
b	0.312	0.315	0.313	0.615	0.616	0.613	0.616	0.617	0.611	0.690	0.692	0.692	0.618	0.615	0.617
Naive NP Ideal	0.000	0.067	0.271	0.026	0.148	0.264	0.025	0.149	0.269	0.091	0.230	0.332	0.227	0.301	0.339
Arithmetic Mean	0.521	0.463	0.351	0.779	0.685	0.599	0.776	0.688	0.599	0.805	0.744	0.687	0.818	0.787	0.725
Geometric Mean	0.080	0.248	0.303	0.405	0.407	0.418	0.407	0.406	0.415	0.470	0.482	0.490	0.633	0.617	0.593
Maximum	0.704	0.591	0.433	0.846	0.783	0.720	0.845	0.784	0.719	0.886	0.831	0.780	0.853	0.830	0.779
Minimum	0.018	0.094	0.192	0.168	0.193	0.224	0.168	0.194	0.219	0.188	0.216	0.244	0.451	0.436	0.471
Arithmetic Merge	0.777	0.638	0.507	0.861	0.809	0.745	0.861	0.810	0.743	0.918	0.865	0.812	0.860	0.841	0.809
Geometric Merge	0.502	0.506	0.522	0.680	0.641	0.623	0.680	0.640	0.623	0.786	0.746	0.721	0.735	0.735	0.728
Bonferroni	0.092	0.214	0.344	0.306	0.328	0.363	0.303	0.328	0.357	0.347	0.371	0.409	0.558	0.560	0.583
Arithmetic (CDF)	0.131	0.112	0.024	0.277	0.257	0.175	0.280	0.257	0.176	0.407	0.362	0.303	0.544	0.445	0.354
Geometric (CDF)	0.007	0.095	0.147	0.184	0.226	0.241	0.183	0.226	0.242	0.244	0.276	0.297	0.547	0.467	0.435
Max (CDF)	0.215	0.170	0.038	0.385	0.320	0.199	0.390	0.320	0.206	0.479	0.424	0.351	0.535	0.451	0.313
Min (CDF)	0.090	0.210	0.338	0.301	0.323	0.359	0.300	0.324	0.355	0.338	0.368	0.404	0.557	0.559	0.580
Arithmetic (ECDF)	0.014	0.110	0.269	0.087	0.255	0.391	0.086	0.252	0.401	0.205	0.357	0.464	0.271	0.443	0.525
Geometric (ECDF)	0.000	0.095	0.270	0.082	0.222	0.334	0.081	0.225	0.345	0.158	0.273	0.374	0.469	0.465	0.504
Max (ECDF)	0.065	0.169	0.282	0.202	0.314	0.451	0.201	0.317	0.460	0.316	0.422	0.504	0.361	0.443	0.494
Min (ECDF)	0.090	0.208	0.288	0.305	0.322	0.322	0.303	0.324	0.327	0.344	0.369	0.390	0.558	0.555	0.559
Multivariate ECDF	0.000	0.090	0.267	0.024	0.216	0.314	0.025	0.221	0.325	0.109	0.266	0.368	0.409	0.456	0.493
Naive NP (histo)	0.479	0.533	0.630	0.533	0.576	0.675	0.528	0.574	0.684	0.133	0.453	0.361	0.665	0.613	0.499
NP V-Matrix	0.004	0.078	0.267	0.014	0.170	0.281	0.012	0.174	0.283	0.109	0.239	0.312	0.230	0.380	0.341

	σ_a^2=(1.0,4.0), σ_b^2=(4.0,1.0)			σ_a^2=(1.0,4.0), σ_b^2=(4.0,4.0)			σ_a^2=(4.0,1.0), σ_b^2=(4.0,1.0)			σ_a^2=(4.0,1.0), σ_b^2=(4.0,4.0)			σ_a^2=(4.0,4.0), σ_b^2=(4.0,4.0)		
	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800	-0.800	0.000	0.800
a	0.612	0.616	0.615	0.612	0.616	0.616	0.615	0.615	0.616	0.617	0.617	0.615	0.691	0.692	0.689
b	0.618	0.617	0.616	0.694	0.694	0.689	0.618	0.613	0.615	0.692	0.690	0.691	0.690	0.693	0.691
Naive NP Ideal	0.196	0.283	0.343	0.333	0.404	0.456	0.227	0.296	0.340	0.329	0.405	0.455	0.471	0.544	0.587
Arithmetic Mean	0.905	0.846	0.779	0.892	0.854	0.809	0.814	0.789	0.725	0.890	0.853	0.811	0.915	0.887	0.853
Geometric Mean	0.518	0.502	0.493	0.673	0.650	0.632	0.634	0.619	0.592	0.671	0.651	0.634	0.780	0.759	0.739
Maximum	0.971	0.941	0.907	0.935	0.911	0.880	0.850	0.831	0.781	0.934	0.911	0.881	0.952	0.931	0.906
Minimum	0.318	0.324	0.334	0.420	0.429	0.441	0.452	0.436	0.472	0.426	0.427	0.443	0.470	0.480	0.487
Arithmetic Merge	0.981	0.954	0.922	0.946	0.927	0.902	0.858	0.842	0.809	0.946	0.927	0.903	0.969	0.952	0.930
Geometric Merge	0.681	0.664	0.649	0.817	0.801	0.790	0.736	0.735	0.726	0.816	0.801	0.786	0.941	0.919	0.902
Bonferroni	0.446	0.445	0.446	0.571	0.574	0.583	0.559	0.560	0.580	0.573	0.574	0.586	0.650	0.655	0.662
Arithmetic (CDF)	0.427	0.373	0.305	0.554	0.495	0.438	0.554	0.449	0.357	0.551	0.495	0.438	0.612	0.570	0.524
Geometric (CDF)	0.376	0.365	0.353	0.503	0.482	0.470	0.553	0.471	0.434	0.508	0.484	0.472	0.577	0.565	0.554
Max (CDF)	0.530	0.452	0.362	0.596	0.533	0.460	0.539	0.452	0.313	0.591	0.534	0.457	0.642	0.601	0.548
Min (CDF)	0.444	0.442	0.445	0.566	0.570	0.579	0.557	0.555	0.579	0.566	0.571	0.583	0.647	0.651	0.659
Arithmetic (ECDF)	0.218	0.369	0.484	0.392	0.492	0.557	0.272	0.445	0.522	0.392	0.490	0.556	0.508	0.568	0.602
Geometric (ECDF)	0.299	0.361	0.410	0.437	0.484	0.524	0.477	0.469	0.510	0.437	0.480	0.517	0.517	0.562	0.598
Max (ECDF)	0.323	0.449	0.562	0.452	0.526	0.580	0.358	0.452	0.495	0.457	0.529	0.575	0.560	0.596	0.622
Min (ECDF)	0.442	0.442	0.433	0.569	0.565	0.574	0.560	0.556	0.560	0.570	0.572	0.572	0.647	0.649	0.655
Multivariate ECDF	0.257	0.358	0.414	0.400	0.473	0.522	0.409	0.467	0.487	0.400	0.481	0.517	0.493	0.556	0.598
Naive NP (histo)	0.252	0.644	0.671	0.640	0.452	0.494	0.667	0.612	0.724	0.373	0.438	0.493	0.488	0.552	0.595
NP V-Matrix	0.249	0.320	0.346	0.396	0.442	0.453	0.203	0.376	0.330	0.399	0.434	0.451	0.484	0.547	0.592

Table 3: Fraction of uncertain predictions for significance level $\epsilon = 0.05$. There are 10 scenarios in terms of the variances of the NCMs for the two labels and the 2 CPs. In the headings the two number for σ^2 are the variances for the NCMs for examples with label "0" and for examples with label "1". In each such scenario, different levels of correlations (-0.8, 0, +0.8) were injected between corresponding NCMs. The reported values are averages over 25 runs. The lower the fraction, the higher the CP efficiency.

	0.010	0.050	0.100	0.150	0.200
Naive NP Ideal	1.433	1.400	1.400	1.167	1.000
Arithmetic (ECDF)	5.467	4.900	4.600	4.567	5.167
Geometric (ECDF)	4.100	4.733	4.767	4.500	3.700
Max (ECDF)	6.933	6.233	6.400	6.833	7.200
Min (ECDF)	5.800	6.933	7.200	7.300	7.233
Multivariate ECDF	2.100	3.733	4.000	4.267	4.000
Naive NP (histo)	6.733	6.100	5.600	5.033	4.500
NP V-Matrix	3.433	1.967	2.033	2.333	3.200

Table 4: Average rank of the method when sorted by efficiency, as a function of significance level. Apart from the $\epsilon = 0.01$ case at the left, NP V-Matrix is consistently the best after the Naïve NP Ideal.

	-0.800	0.000	0.800
Naive NP Ideal	1.280	1.020	1.540
Arithmetic (ECDF)	4.660	4.880	5.280
Geometric (ECDF)	4.440	4.340	4.300
Max (ECDF)	6.740	6.760	6.660
Min (ECDF)	7.320	7.000	6.360
Multivariate ECDF	3.160	3.520	4.180
Naive NP (histo)	5.320	5.780	5.680
NP V-Matrix	3.080	2.700	2.000

Table 5: Average rank of the method when sorted by efficiency, as a function of correlation. NP V-Matrix is consistently the best after the Naïve NP Ideal.

	(1.0,1.0), (1.0,1.0)	(1.0,1.0), (1.0,4.0)	(1.0,1.0), (4.0,1.0)	(1.0,1.0), (4.0,4.0)	(1.0,4.0), (1.0,4.0)	(1.0,4.0), (4.0,1.0)	(1.0,4.0), (4.0,4.0)	(4.0,1.0), (4.0,1.0)	(4.0,1.0), (4.0,4.0)	(4.0,4.0), (4.0,4.0)
Naive NP Ideal	1.467	1.333	1.200	1.333	1.400	1.000	1.200	1.533	1.200	1.133
Arithmetic (ECDF)	5.067	5.467	5.400	6.000	4.533	4.400	4.667	4.400	5.000	4.467
Geometric (ECDF)	2.667	3.533	3.533	3.933	5.267	4.533	4.933	4.867	5.067	5.267
Max (ECDF)	6.867	7.000	7.000	7.600	5.400	6.800	7.133	5.200	7.200	7.000
Min (ECDF)	7.000	6.067	6.067	6.667	6.867	6.467	7.400	6.800	7.600	8.000
Multivariate ECDF	2.867	2.667	2.800	3.133	4.000	3.933	4.200	3.800	4.600	4.200
Naive NP (histo)	6.333	7.000	7.133	4.267	6.800	6.133	4.400	7.800	3.067	3.000
NP V-Matrix	3.733	2.933	2.867	3.067	1.733	2.733	2.067	1.600	2.267	2.933

Table 6: Average rank of the method when sorted by efficiency, for the various scenarios of σ_a and σ_b. With the exception of the two case at the left, NP V-Matrix is consistently the best after the Naïve NP Ideal.

References

Vineeth N. Balasubramanian, Shayok Chakraborty, and Sethuraman Panchanathan. Conformal predictions for information fusion. *Annals of Mathematics and Artificial Intelligence*, 74(1):45–65, Jun 2015. ISSN 1573-7470. doi: 10.1007/s10472-013-9392-4. URL https://doi.org/10.1007/s10472-013-9392-4.

R. A. Fisher. *Statistical methods for research workers, 4th. ed.* Edinburgh Oliver & Boyd, 1932.

R. A. Fisher. Question 14: Combining independent tests of significance. *The American Statistician*, 2(5):30–30, 1948.

Christian Genest and Louis-Paul Rivest. On the multivariate probability integral transformation. *Statistics & Probability Letters*, 53(4):391–399, 2001. URL https://EconPapers.repec.org/RePEc:eee:stapro:v:53:y:2001:i:4:p:391-399.

N. A. Heard and P. Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018. doi: 10.1093/biomet/asx076. URL http://dx.doi.org/10.1093/biomet/asx076.

Thomas M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004.

Thomas Sellke, M. J Bayarri, and James O Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001. doi: 10.1198/000313001300339950. URL https://doi.org/10.1198/000313001300339950.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, Aug 2018. ISSN 1573-0565. doi: 10.1007/s10994-018-5754-9. URL https://doi.org/10.1007/s10994-018-5754-9.

Vladimir Vapnik and Rauf Izmailov. Statistical inference problems and their rigorous solutions. In Alexander Gammerman, Vladimir Vovk, and Harris Papadopoulos, editors, *Statistical Learning and Data Sciences*, pages 33–71, Cham, 2015. Springer International Publishing. ISBN 978-3-319-17091-6.

Vladimir Vapnik, Igor Braga, and Rauf Izmailov. Constructive setting for problems of density ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(3):137–146, 2015. doi: 10.1002/sam.11263. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11263.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995. ISBN 0-387-94559-8.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, Jun 2015. ISSN 1573-7470. doi: 10.1007/s10472-013-9368-4. URL https://doi.org/10.1007/s10472-013-9368-4.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *arXiv e-prints*, art. arXiv:1212.4966, Dec 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.