Statistical assesment of experimental results: a graphical approach for comparing algorithms

Etor Arza, Josu Ceberio, Ekhiñe Irurozki and Aritz Pérez

Abstract—**Non-deterministic measurements are common in real-world scenarios: the performance of a stochastic optimization algorithm or the total reward of a reinforcement learning agent in a chaotic environment are just two examples in which unpredictable outcomes are common. These measures can be modeled as random variables and compared among each other via their expected values or more sophisticated tools such as null hypothesis statistical tests. In this paper, we propose an alternative framework to compare two random variables according to their cumulative distribution functions. First, we introduce a dominance measure for two random variables that quantifies the proportion in which the cumulative distribution function of one of the random variables is greater than the other. Then, we present a graphical method that allows a visual estimation of the proposed dominance measure, the probability that one of the random variables takes lower values than the other, and a comparison of quantiles of the random variables. With illustrative purposes, we re-evaluate the experimentation of an already published work with the proposed methodology and we show that additional conclusions—missed by the rest of the methods—can be inferred. Additionally, a software package is provided as a convenient way of applying the proposed framework.**

Index Terms—**random variables, cumulative distribution function, first-order stochastic dominance, data visualization**

I. INTRODUCTION

The objective value obtained by an optimization algorithm may be non-deterministic. For example, in stochastic algorithms, the objective value measured depends on the seed used in the random number generator. In these kinds of scenarios, we can think that these non-deterministic measurements are observations of random variables with unknown distributions. Based on these measurements, we sometimes need to choose the random variable that takes the lowest (or largest) values. The expected values of the random variables—usually estimated as an average of several repeated observations—can be used for this purpose. However, many statisticians have claimed that summarizing data with simple statistics such as the average or the standard deviation is misleading, as very different data can still have the same statistics [1], [2].

Motivating example 1): A real-world motivation for this work is as follows. Suppose we need to choose the best option between two stochastic gradient-based methods for optimizing the parameters of a neural network. A neural network classifier trained with a gradient-based method will produce different

Etor Arza (earza@bcamath.org) and Aritz Pérez (aperez@bcamath.org) are with BCAM - Basque Center for Applied Mathematics.
Josu Ceberio (josu.ceberio@ehu.eus) is with the University of the Basque Country UPV/EHU.
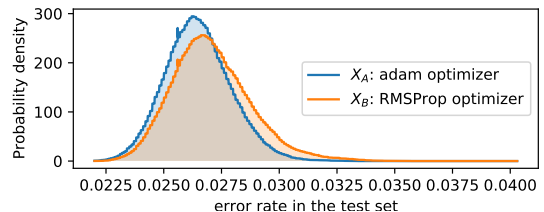Ekhiñe Irurozki (irurozki@telecom-paris.fr) is with Télécom Paris.

Figure 1. Density estimates of the error rates produced by the optimizers *adam* and *RMSProp* in the MNIST dataset. The sci-kit learn[6] package was used in the estimation.

error rates [3] each time it is trained-tested, even if the same train-test dataset is used in each repeated measurement. One of the reasons is that the learned classifier depends on the initialization of its weights (before applying a gradient-based optimizer), which are often initialized randomly [4].

To illustrate the previous scenario, we trained and tested a neural network[1] in the MNIST dataset, and we compared two gradient-based optimizers in this data set: *adam* and *RMSProp* [3]. The error rate in the test set depends on the seed used to train the neural network, and therefore, we can model the error rate of each of the algorithms in this problem as a random variable. An observation of each of the two random variables (the error rate of each gradient-based optimizer is modeled as a random variable) involves training the neural network in the training set and measuring its error rate in the test set: the training and test sets are the same for each trained neural network. Figure 1 shows the kernel density estimations of these random variables using the uniform kernel. As we see in the figure, the error rate is not the same in each measurement and ranges between 0.022 and 0.04. This shows that, in this context, it makes sense to model the error rate as a random variable rather than a constant: a unique value cannot represent the error rate without a significant amount of information loss.

Motivating example 2): In the following, we present another example with synthetic data. Let us consider the two random variables X_A and X_B shown in Figure 2. X_B has a lower expected value than X_A, $\mathbb{E}[X_B] < \mathbb{E}[X_A]$. If we use the expected value as the only criterion, then X_B takes lower values than X_A. However, notice that with a low but nonzero probability, X_B will take very large values that are undesirable in the context of minimization. Without loss of generality, in this paper, we assume that lower values are preferred.

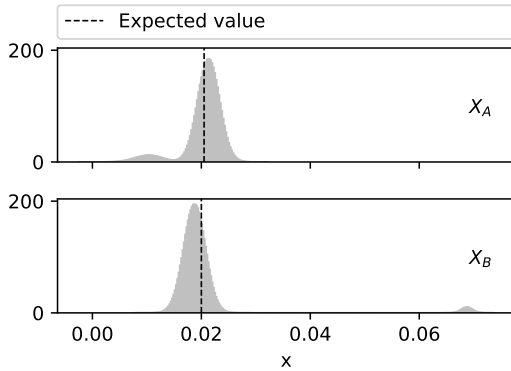In an environment where reliability is key, such as breast can-

Figure 2. The probability density of two random variables X_A and X_B, with probability density functions $g_A = 0.925 \cdot g_{\mathcal{N}(0.210325, 0.002)} + 0.075 \cdot g_{\mathcal{N}(0.010325, 0.025)}$ and $g_B = 0.975 \cdot g_{\mathcal{N}(0.01875, 0.002)} + 0.025 \cdot g_{\mathcal{N}(0.06875, 0.001)}$ where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. Their expected values are $\mathbb{E}[X_A] = 0.0205$ and $\mathbb{E}[X_B] = 0.02$ respectively.

cer detection [7], or certain reinforcement learning tasks [8], [9] like self-driving cars [10], an error with low variance is very important, even if it means a slightly worse expected value.

In other circumstances, obtaining the lowest possible error can be more important than reliability. One could argue that reliability is less important in sentiment analysis [11], or in certain real-world optimization problems [12], where obtaining the best possible solution is key. When obtaining the best possible score is more important than reliability, it may even be worth running an optimization algorithm several times and choosing the best solution out of all the runs. In that case, X_A would also be preferred to X_B, as X_A has a higher probability of taking a value lower than 0.01 (see Figure 2).

a) Related work: In these two examples, we have seen that summarizing and comparing random variables with only the expected value can leave important information out (such as which of the random variables can take lower values), especially when neither random variable clearly takes lower values than the other one. Many works in the literature use null hypothesis tests [13], [14], [15] to analyze observed samples and choose one of the random variables accordingly. Nonetheless, as claimed in Benavoli et al. [16], null hypothesis tests have their limitations too: when the null hypothesis is not rejected—this will happen often when the random variables being compared take similar values—, we get no information. Not only that but even when the null hypothesis is rejected, it does not quantify the amount of evidence in favor of the alternative hypothesis [16].

b) Contribution: In this paper, we propose a graphical framework that compares two random variables using their associated cumulative distribution functions, in the context of choosing the one that takes lower values. The proposed methodology can compare the scores of two stochastic optimization algorithms or the error rates of two classifiers, among other applications. To achieve this, the performances of the

optimization algorithms (or the error rates of the classifiers) are modeled as random variables, and then, we compare them by measuring the *dominance*.

Specifically, we first propose 8 desirable properties for *dominance measures*: functions that compare two random variables in this context. From the measures in the literature, we find that the *probability that one of the random variables takes a lower value than the other random variable* satisfies most of these properties. In addition, we propose a new dominance measure, the *dominance rate*, that also satisfies most of the properties and is related to the first-order stochastic dominance [17]. Then, we propose a graphical method that involves visually comparing the random variables through these two dominance measures. The graphical method, named *cumulative difference-plot*, can also be used to compare the quantiles of the random variables, and it models the uncertainty associated with the estimate. By re-evaluating the experimentation of a recently published paper with the proposed methodology, we demonstrate that this new plot can be useful to compare two random variables, especially in the case when the random variables take similar values.

Finally, an *R* package named *RVCompare*, available in CRAN, is distributed alongside this paper. With this package, the *cumulative difference-plot* can be conveniently computed. The source code of the package and the supplementary material for the paper are available at github.com/EtorArza[2].

The rest of the paper is organized as follows: in the next section, we propose eight desirable properties for dominance measures. Then, in Section III, we study two dominance measures that satisfy most of these properties. Section IV introduces a graphical method to compare random variables. In Section V, we discuss related methods in the literature and compare them to the proposed approach. Section VI, evaluates the proposed graphical method and other alternatives in an already published experimentation. In Section VII, we state the assumptions and limitations of the proposed *cumulative difference-plot*. Finally, Section VIII concludes the paper.

II. DESIRABLE PROPERTIES FOR DOMINANCE MEASURES

A. Background

When we have two random variables and we need to choose the one that takes the lowest values, we usually take i) the random variable with the lowest expected value or ii) the random variable with the lowest median. The *median* [14] of a continuous random variable X_A, denoted as m_A, is the value that satisfies $\mathcal{P}(X_A < m_A) = \mathcal{P}(X_A > m_A)$. In other words, if m_A is the median of X_A, a sample of X_A is as likely to be lower than m_A as it is to be higher.

Interestingly enough, the median and the expected value have their strengths and weaknesses when it comes to choosing the

---

[2]The source of the package *RVCompare* can be found at github.com/EtorArza/RVCompare. The code to reproduce every figure in the paper is available at github.com/EtorArza/SupplementaryPaperRVCompare.
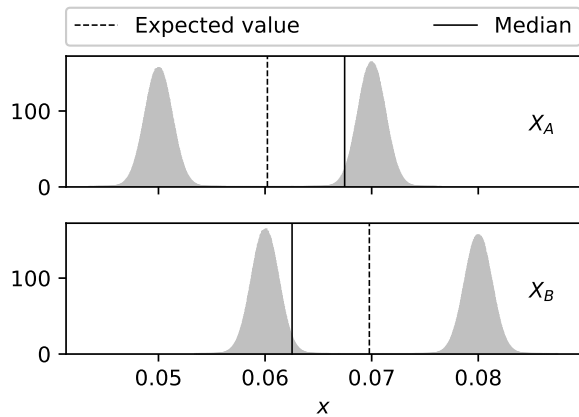
Figure 3. Case 1. The probability density functions of X_A and X_B: $g_A = 0.489 \cdot g_{\mathcal{N}(0.05,0.00125)} + 0.511 \cdot g_{\mathcal{N}(0.07,0.00125)}$ and $g_B = 0.511 \cdot g_{\mathcal{N}(0.06,0.00125)} + 0.489 \cdot g_{\mathcal{N}(0.08,0.00125)}$ where $g_{\mathcal{N}(\mu,\sigma)}$ is the density function of the normal distribution with mean μ and standard deviation σ.
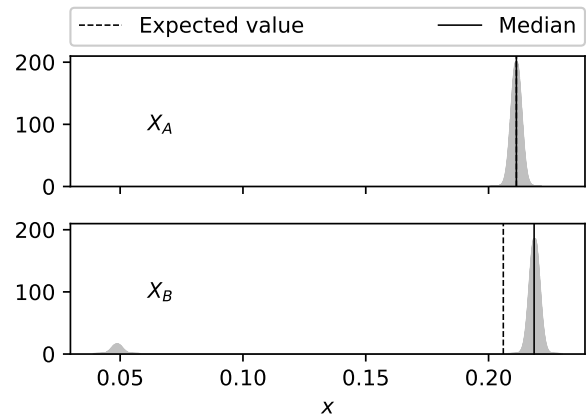

Figure 4. Case 2. The probability density functions of X_A and X_B: $g_A = g_{\mathcal{N}(0.211325,0.002)}$ and $g_B = 0.925 \cdot g_{\mathcal{N}(0.21875,0.002)} + 0.075 \cdot g_{\mathcal{N}(0.04875,0.002)}$ respectively.

random variable that takes the lowest values. In the following, we elaborate on this point with two particular cases of study. The first case is shown in Figure 3, with two random variables X_A and X_B. Each of the random variables is a mixture of two Gaussian distributions with the same shape and similar weight in the mixture. It is clear that X_A tends to take values lower than X_B, as the Gaussian distributions of X_A are centered in 0.05 and 0.06, while the Gaussian distributions of X_B are centered in 0.06 and 0.08. While the expected values of X_A and X_B are aligned with this intuition, the medians are not; as $\mathbb{E}[X_A] < \mathbb{E}[X_B]$ and $m_A > m_B$. However, the expected value does a poor job of summarizing the bimodal shape of X_A or X_B: both of these random variables usually take much higher or much lower values than their expected values.

The second case is shown in Figure 4. With a very high probability, X_A takes lower values than X_B, even though X_B will rarely take really low values, which might prove useful in some particular applications. In this case, $m_A < m_B$ and $\mathbb{E}[X_A] > \mathbb{E}[X_B]$, hence, the comparison of the medians are aligned with the intuition that X_A takes lower values than X_B, while the expected values are not. In the presence of outliers [18], the median is considered more robust than the expected value [19].

Notice that, in the second case, it is not trivial to choose between X_A and X_B, as X_B can take lower values, but X_A is more likely to be lower than X_B. So, when can we claim that one of them clearly takes lower values than the other? When the cumulative distribution of X_A is higher than the cumulative distribution of X_B in the entire domain of definition: in that case, X_A has a higher probability than X_B of taking values lower than x, for all x in the domain of definition. This is known as X_A being stochastically smaller than X_B [13]. Depending on the field of study, it can also be referred to [20], [21], [17] as "X_A stochastically dominates X_B"[3]. The stochastic dominance can be further relaxed, obtaining

what is known as *first-order stochastic dominance* in the literature [20], [17], although, for the sake of brevity, we will call it *dominance* throughout the paper.

Definition 1. *(Dominance) Let X_A and X_B be two continuous random variables defined in a connected subset $N \subseteq \mathbb{R}$. We say that X_A dominates X_B, denoted as $X_A \succ X_B$, when*

i) $G_A(x) \geq G_B(x)$ for all $x \in N$
and
ii) There exists an $x \in N$ such that $G_A(x) > G_B(x)$.

where G_A and G_B are the cumulative distributions of X_A and X_B respectively.

For X_A not to dominate X_B (denoted as $X_A \not\succ X_B$[4]), either condition i) or ii) must be violated. The special case that $X_A \not\succ X_B$ and $X_B \not\succ X_A$ at the same time is defined, it is said that X_A and X_B *cross* [21], and we denote it as $X_A \lessgtr X_B$. In the non trivial ($X_A \neq X_B$) case that $X_A \lessgtr X_B$, there exists two points $x_1, x_2 \in N$ such that $G_A(x_1) < G_B(x_1)$ and $G_A(x_2) > G_B(x_2)$: we cannot say, for all $x \in N$, that one of the random variables has a higher probability of taking values lower than x.

Let us now see how the cumulative distributions can be used to compare random variables in an example. In Figure 5a, the cumulative distributions of the random variables described in Figure 3 are shown. We can see that $G_A(x) > G_B(x)$ for almost all $x \in N$. But there is at least a point $x \in (0.06, 0.07)$ where $G_A(x) < G_B(x)$, hence, $X_A \lessgtr X_B$. The same happens in the second case (Figure 5b). As in the previous case, $X_A \lessgtr X_B$, because even though $G_A(x) > G_B(x)$ for almost all $x \in N$ (in which $g_A(x) \neq 0$ and $g_B(x) \neq 0$), for all $x \in (0.05, 0.2), G_A(x) < G_B(x)$.

In the following, we will study how to quantify the difference between two random variables, emphasizing the degree to which one of the random variables dominates the other.

---

[3]Without loss of generality, minimization is assumed in this paper.

[4]Note that $X_A \not\succ X_B$ is not equivalent to $X_B \succ X_A$.
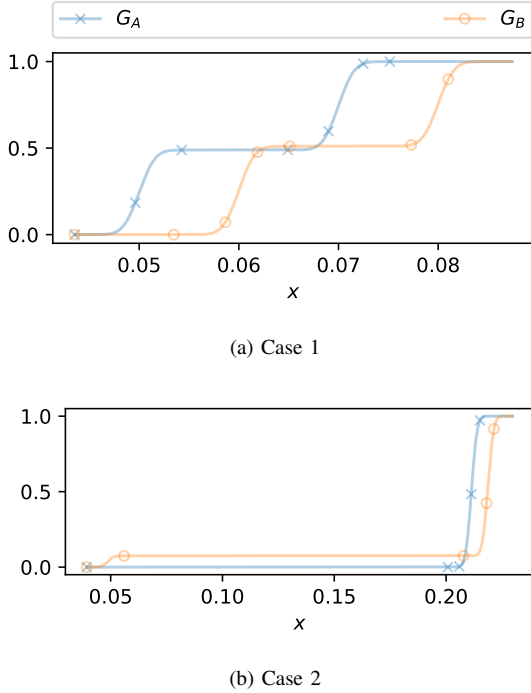
(a) Case 1

(b) Case 2

Figure 5. The cumulative distributions of the two cases shown in Figures 3 and 4.

B. Desirable properties

There are many ways to compare two random variables, each with a different point of view: some aim to find how dissimilar two random variables are (disregarding which of them takes lower values), while other methods try to guess if one of the random variables dominates the other one. In the context of this paper, we are interested in measures that, given two random variables, quantify through the dominance how much one of the random variables tends to take lower values than the other. We use the term *dominance measure* to refer to functions that quantify the difference between two random variables following this intuition. In this section, we define eight desirable properties for these dominance measures, and we study the suitability of several measures from the literature.

Definition 2. *Let X_A and X_B be two continuous random variables. We define a* dominance measure *between two random variables as a function $\mathcal{C}$ that maps two random variables into a real value $\mathcal{C}(X_A, X_B)$.*

It is desirable that $\mathcal{C}(X_A, X_B)$ quantifies the dominance, so $\mathcal{C}(X_A, X_B)$ should be high when X_A takes lower values than X_B and low, otherwise. Formally, this intuitive idea can be represented as:

Property 1. *$\mathcal{C}$ is defined in the $[0,1]$ interval, where:*

i)

$$\mathcal{C}(X_A, X_B) = 1 \iff X_A \succ X_B$$

ii)

$$\mathcal{C}(X_A, X_B) = 0 \iff X_B \succ X_A$$

iii)

$$\mathcal{C}(X_A, X_B) \in (0,1) \iff X_B \lessgtr X_A$$

Proposition 1. *If a dominance measure $\mathcal{C}$ satisfies Property 1 i) and ii), then it also satisfies Property 1 iii).*

Proof. By definition, $X_B \lessgtr X_A$ iff $X_A \not\succ X_B$ and $X_B \not\succ X_A$. Property 1 i) and ii) implies that $X_A \not\succ X_B$ and $X_B \not\succ X_A$ iff $\mathcal{C}(X_A, X_B) \neq 1$ and $\mathcal{C}(X_A, X_B) \neq 0$. From Property 1 i) also $\mathcal{C}(X_A, X_B) \in [0,1]$, thus $X_B \lessgtr X_A$ iff $\mathcal{C}(X_A, X_B) \in (0,1)$. □

Property 2. *(Antisymmetry) $\mathcal{C}(X_A, X_B)$ and $\mathcal{C}(X_B, X_A)$ add up to 1.*

$$\mathcal{C}(X_A, X_B) = 1 - \mathcal{C}(X_B, X_A)$$

It is noteworthy that Property 1 ii) can be inferred from Property 1 i) and Property 2.

Property 3. *The inversion (under the sum) of the operands of $\mathcal{C}$ equals the inversion of $\mathcal{C}$:*

$$\mathcal{C}(-1 \cdot X_A, -1 \cdot X_B) = 1 - \mathcal{C}(X_A, X_B)$$

Property 4. *When X_A and X_B are equal, $\mathcal{C}$ is symmetric.*

$$X_A = X_B \implies \mathcal{C}(X_A, X_B) = \mathcal{C}(X_B, X_A)$$

Assuming Property 2 holds, we can rewrite the previous property as:

$$X_A = X_B \implies \mathcal{C}(X_A, X_B) = 0.5.$$

Property 5. *(Invariance to translation) Moving the domain of definition of X_A and X_B by the same amount does not change $\mathcal{C}$[5].*

$$\text{for all } \lambda \in \mathbb{R}, \quad \mathcal{C}(X_A + \lambda, X_B + \lambda) = \mathcal{C}(X_A, X_B)$$

Property 6. *(Invariance to scaling) Scaling both X_A and X_B by the same positive amount does not change $\mathcal{C}$.*

$$\text{for all } \lambda > 0, \quad \mathcal{C}(\lambda \cdot X_A, \lambda \cdot X_B) = \mathcal{C}(X_A, X_B)$$

There are two more properties based on the intuition explained in the following. In the second case, shown in Figure 4, we saw that for all $x \in (0.075, 0.2) \subseteq N$, $G_A(x) < G_B(x)$. However, intuitively, we would think that $G_A(x) > G_B(x)$ should have a higher weight in $\mathcal{C}(X_A, X_B)$, as most of the probability mass of both X_A and X_B is in the interval $(0.2, 0.23)$. In other words, the *small* mass of X_B centered in 0.05 can only account for a *small* part of $\mathcal{C}(X_A, X_B)$. In what follows, this is formalized as X_B being a mixture of two distributions, where one of the distributions represents this small mass with a small weight in the mixture. Property 7 states that the change in $\mathcal{C}$ produced by the distribution of small weight in the mixture can be, at most, its weight in the mixture.

---

[5]We define $X_A + \lambda$ as the random variable that is sampled in two steps: first obtain an observation from X_A and then add λ to this observation. We define $\lambda \cdot X_A$ in a similar way.

Property 7. *Let $X_B = \mathcal{M}_{[1-\tau,\tau]}(X_{B1}, X_{B2})$ be the mixture[6] distribution of X_{B1} and X_{B2} with weights $1 - \tau$ and τ respectively and let X_A be another random variable. Then,*

$$|\mathcal{C}(X_A, X_B) - \mathcal{C}(X_A, X_{B1})| \leq \tau$$

In addition, the position of this small mass of X_B centered in 0.05 is irrelevant, as long as it moves without overlapping with the mass of X_A and changing its position with respect to it. For example, the small mass could have been centered in 0.1 or 0.15 instead of 0.05, without any changes to the quantification of which of the random variables takes lower values. In addition to the position, the shape of the mass can also be altered without any changes in $\mathcal{C}(X_A, X_B)$, as long as its weight in the mixture stays the same and does not overlap with X_A. We formalize this intuitive idea in Property 8 by first defining the small mass as a component in the mixture distribution, and then we impose the condition that the dominance measure does not change when this mass is translated or scaled, as long as it does not overlap with the support of the other random variable.

Property 8. *Let $X_B = \mathcal{M}_{[1-\tau,\tau]}(X_{B1}, X_{B2})$ be the mixture distribution of X_{B1} and X_{B2} with weights $1 - \tau$, and τ, respectively and let X_A be another random variable. Suppose that $\mathrm{supp}(X_{B2}) \cap \mathrm{supp}(X_A) = \varnothing$. Let $\lambda_1 \in \mathbb{R}^+, \lambda_2 \in \mathbb{R}$ be two numbers such that for all $\xi_1, \xi_2 \in [0,1]$, $\mathrm{supp}((1 + (\lambda_1 - 1)\xi_1) \cdot X_{B2} + \xi_2\lambda_2) \cap \mathrm{supp}(X_A) = \varnothing$.*

Then,

$$\mathcal{C}(X_A, X_B) = \mathcal{C}(X_A, \mathcal{M}_{[1-\tau,\tau]}(X_{B1}, \lambda_1 \cdot X_{B2} + \lambda_2))$$

Unfortunately, it is impossible that a dominance measure satisfies Properties 1 and 7 at the same time. Intuitively, the problem is that, given the distributions X_A and $X_B = \mathcal{M}_{[1-\tau,\tau]}(X_{B1}, X_{B2})$, it is possible that $X_A \succ X_{B1}$ and at the same time $X_B \succ X_A$ with $\tau < 0.5$[7]. We formalize and prove this claim in the following proposition:

Proposition 2. *Let $\mathcal{C}$ be a dominance measure.*

i) If $\mathcal{C}$ satisfies Property 1, then it fails to satisfy Property 7.

ii) If $\mathcal{C}$ satisfies Property 7, then it fails to satisfy Property 1.

Proof. A dominance measure only satisfies a property when that property is true for every possible random variable. Consequently, to prove this proposition, it is enough to find four random variables X_A, X_B, X_{B1} and X_{B2} where

i) $X_B = \mathcal{M}_{[0.1,0.9]}(X_{B1}, X_{B2})$,
ii) $X_A \succ X_{B1}$,
iii) $X_B \succ X_A$.

If four random variables can be found that satisfy these three statements, then with Property 1 we obtain that

[6]The probability density function of $\mathcal{M}_{[1-\tau,\tau]}(X_{B1}, X_{B2})$ is defined as $(1 - \tau) \cdot g_{B1}(x) + \tau \cdot g_{BS2}(x)$. Note that $\tau \in [0, 1]$.
[7]See https://etorarza.github.io/pages/2021-interactive-comparing-RV.html for an interactive example that illustrates the above point.
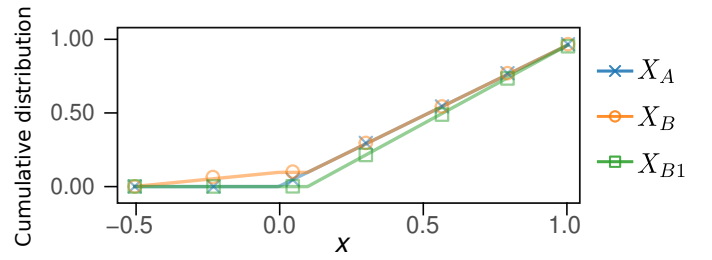

Figure 6. The cumulative distribution functions of X_A, X_B and X_{B1}.

Table I
WHICH OF THE PROPERTIES IN SECTION II-B DOES EACH MEASURE SATISFY?

	1	2	3	4	5	6	7	8
Kullback-Leibler divergence				✓	✓	✓		✓
Jensen-Shannon divergence				✓	✓	✓		✓
Total-Variation				✓	✓	✓	✓	✓
Hellinger distance				✓	✓	✓	✓	✓
Wasserstein distance				✓	✓			
$\mathcal{C}_\mathcal{P}$: Probabitlity of $X_A < X_B$		✓	✓	✓	✓	✓	✓	✓
$\mathcal{C}_\mathcal{D}$: Dominance rate of X_A over X_B	✓	✓	✓	✓	✓	✓		✓

A checkmark ✓ indicates that the measure satisfies the property.

$\mathcal{C}(X_A, X_{B1}) = 1$ and $\mathcal{C}(X_A, X_B) = 0$. This contradicts Property 7, because $|\mathcal{C}(X_A, X_B) - \mathcal{C}(X_A, X_{B1})| \nleq 0.1$. The same is true the other way around, Property 7 states that $|\mathcal{C}(X_A, X_B) - \mathcal{C}(X_A, X_{B1})| \leq 0.1$ and this contradicts Property 1, with $\mathcal{C}(X_A, X_{B1}) < 1$ or $\mathcal{C}(X_A, X_B) > 0$.

A simple example in which this happens is for the random variables

$X_A = \mathcal{U}(0, 1)$,

$X_B = \mathcal{M}_{[0.9, 0.1]}(\mathcal{U}(0.1, 1), \mathcal{U}(-0.5, 0))$,

$X_{B1} = \mathcal{U}(0.1, 1)$,

$X_{B2} = \mathcal{U}(-0.5, 0)$.

The cumulative distribution functions of X_A, X_B and X_{B1} are shown in Figure 6, where it is clear that $X_B \succ X_A$ and $X_A \succ X_{B1}$. $\square$

In the following, we will briefly review several measures in the literature and, specifically, which of the proposed properties they satisfy. Many measures describe the difference between X_A and X_B, disregarding whether the difference in cumulative density is positive or negative. Consequently, they cannot satisfy Property 1 (see Appendix A for details). This is the case for *f-divergences*—including Kullback-Leibler, Jensen-Shannon, the Hellinger distance and the total variation—and for the Wasserstein distance. These measures also fail to satisfy several other properties (see a summary in Table I).

III. DOMINANCE MEASURES

Most of the measures in the literature fail to satisfy the eight properties introduced in Section II-B. However, there is a dominance measure in the literature that overcomes this limitation: the probability that $X_A < X_B$ [14].

A. $\mathcal{C}_\mathcal{P}$: the probability of $X_A < X_B$

We can compare X_A and X_B with the probability that a value sampled from X_A is smaller than a value sampled from X_B. When the random variables are exactly the same, this probability is 0.5. Formally, given two continuous random variables X_A and X_B defined in a connected set $N \subseteq \mathbb{R}$, the probability that $X_A < X_B$ is defined as:

$$\mathcal{P}(X_A < X_B) = \int_N g_B(x) G_A(x) dx. \tag{1}$$

When we consider $\mathcal{P}(X_A < X_B)$ as a dominance measure, we will denote it as $\mathcal{C}_\mathcal{P}(X_A, X_B)$.

One of the advantages of $\mathcal{C}_\mathcal{P}$ is its easy interpretation. In addition, $\mathcal{C}_\mathcal{P}$ is a well behaved dominance measure, as it satisfies Properties 2, 3, 4, 5, 6, 7 and 8. It also satisfies a weak version of Property 1:

$$\mathcal{C}_\mathcal{P}(X_A, X_B) = 1 \implies X_A \succ X_B \implies$$

$$\mathcal{C}_\mathcal{P}(X_A, X_B) \in (0.5, 1]$$

and

$$\mathcal{C}_\mathcal{P}(X_A, X_B) = 0 \implies X_B \succ X_A \implies$$

$$\mathcal{C}_\mathcal{P}(X_A, X_B) \in [0, 0.5).$$

Note that, when $X_A \succ X_B$, $\mathcal{C}_\mathcal{P}(X_A, X_B) \neq 1$ is still possible, and this is why it does not satisfy Property 1 entirely. For instance, when the probability densities of X_A and X_B are two Gaussian distributions with the same variance and the mean of X_A is lower, then $X_A \succ X_B$ but $\mathcal{C}_\mathcal{P}(X_A, X_B) < 1$.

So far, we have seen that $\mathcal{C}_\mathcal{P}$ satisfies most of the properties. Unfortunately, since it does not satisfy Property 1, not all cases of $X_A \succ X_B$ can be identified by $\mathcal{C}_\mathcal{P}$. We now propose a dominance measure that satisfies Property 1 and, thus, can be used to identify cases in which $X_A \succ X_B$.

B. $\mathcal{C}_\mathcal{D}$: dominance rate

Intuitively, the *dominance rate* is a dominance measure that quantifies the extent to which X_A has a lower cumulative distribution function than X_B, normalized by the portion of the probability densities with different cumulative distributions.

Definition 3. *(Dominance density function) Let X_A and X_B be two continuous random variables defined in a connected set $N \subseteq \mathbb{R}$. We define the dominance density function as follows:*

$$\mathcal{D}_{X_A, X_B}(x) = \begin{cases} g_A(x) \cdot k_A & \text{if} \quad G_A(x) > G_B(x) \\ -g_B(x) \cdot k_B & \text{if} \quad G_A(x) < G_B(x) \\ 0 & \text{otherwise.} \end{cases}$$

where $k_A = \left(\int_{\{x \in N \ | \ G_A(x) \neq G_B(x)\}} g_A(t) dt \right)^{-1}$ is the normalization constant and k_B is defined likewise.

Note that the dominance density function is not correctly defined when $\int_N |g_A(x) - g_B(x)| dx = 0$.

Definition 4. *(Dominance rate) Let X_A and X_B be two continuous random variables defined in a connected set $N \subseteq \mathbb{R}$. The dominance rate of X_A over X_B is defined as*

$$\mathcal{C}_\mathcal{D}(X_A, X_B) = \begin{cases} 0.5, & \text{if} \int_N |g_A(x) - g_B(x)| dx = 0 \\ 0.5 \int_N \mathcal{D}_{X_A, X_B}(t) dt + 0.5, & \text{otherwise.} \end{cases}$$

Basically, we are measuring the amount of mass of X_A in which $G_A(x) > G_B(x)$ minus the amount of mass of X_B in which $G_A(x) < G_B(x)$. This value is then normalized so that all sections in which $G_A(x) = G_B(x)$ are ignored, i.e.

$\int_N \mathcal{D}_{X_A, X_B}(t) dt =$

$$\frac{\mathbb{E}_A[\mathcal{I}[G_A(x) > G_B(x)]]}{\mathbb{E}_A[\mathcal{I}[G_A(x) \neq G_B(x)]]} - \frac{\mathbb{E}_B[\mathcal{I}[G_A(x) < G_B(x)]]}{\mathbb{E}_B[\mathcal{I}[G_A(x) \neq G_B(x)]]}$$

Finally, we apply the linear transformation $l(x) = 0.5x - 0.5$ ensuring the dominance rate is defined in the interval $[0, 1]$ (instead of $[-1, 1]$), required to comply with Property 1.

From

i) $\mathcal{C}_\mathcal{D}(X_A, X_B) = 1 \iff X_A \succ X_B$ and

ii) $\mathcal{C}_\mathcal{D}(X_A, X_B) = 0 \iff X_B \succ X_A$,

we deduce that the dominance rate satisfies Property 1. Note that the previous deduction is only possible when g_A and g_B are bounded, as this implies that G_A and G_B are continuous. Specifically, it is enough to find a point in N in which $G_A(x) > G_B(x)$ to satisfy that $\int_{x \in \{t \in N \ | \ G_A(t) > G_B(t)\}} g_A(x) dx > 0$, and this point is guaranteed to exist when $X_A \succ X_B$ because of the definition of the dominance. The dominance rate is also a well behaved dominance measure, as it satisfies Properties 1, 2, 3, 4, 5, 6 and 8.

We have seen that the dominance measures $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ satisfy most of the properties listed in Section II-B. As we will see in the next section, their values are related.

C. The relationship between $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$

In Section II-A we stated that $\mathcal{C}_\mathcal{P} = 1$ is a stronger condition than $\mathcal{C}_\mathcal{D} = 1$, because $\mathcal{C}_\mathcal{P}(X_A, X_B) = 1$ implies that for all x in N that $G_A(x) < 1$, $G_B(x) = 0$. On the other hand, $\mathcal{C}_\mathcal{D} = 1$ implies that $X_A \succ X_B$ (the two conditions in Definition 1), which is weaker. In the diagram below, we show the values

of $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ that imply other values of $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$. Each arrow can be interpreted as an implication. The implications are transitive: i.g. $\mathcal{C}_\mathcal{D}(X_A, X_B) = 1$ implies $\mathcal{C}_\mathcal{P}(X_A, X_B) > 0.5$.
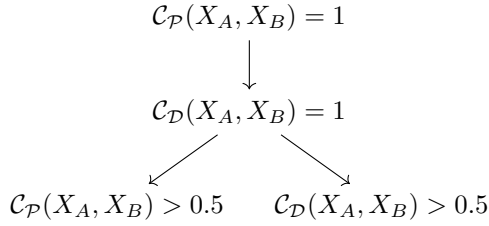
$$\mathcal{C}_\mathcal{P}(X_A, X_B) = 1$$
$$\downarrow$$
$$\mathcal{C}_\mathcal{D}(X_A, X_B) = 1$$

$$\mathcal{C}_\mathcal{P}(X_A, X_B) > 0.5 \qquad \mathcal{C}_\mathcal{D}(X_A, X_B) > 0.5$$

Figure 7. Implications between the values of $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$.

D. Estimating $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$

In the previous sections, we have assumed that the random variables X_A and X_B are known, but usually, we only have a few observed values from each random variable. Therefore, it may be interesting to estimate $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ from the observed samples. With this purpose, we propose the following empirical estimates of $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$.

Definition 5. (estimation of $\mathcal{C}_\mathcal{P}$)
Let X_A and X_B be two continuous random variables and $A_n = \{a_1, ..., a_n\}$ and $B_n = \{b_1, ..., b_n\}$ their n observations respectively. We define the estimation of the probability that $X_A < X_B$ as

$$\widetilde{\mathcal{C}_\mathcal{P}}(A_n, B_n) = \sum_{i,k=1...n} \frac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2}.$$

This estimator is well known in the literature because it is the U statistic of the Mann-Whitney test [13].

Definition 6. (estimation of $\mathcal{C}_\mathcal{D}$)
Let X_A and X_B be two continuous random variables and $A_n = \{a_1, ..., a_n\}$ and $B_n = \{b_1, ..., b_n\}$ their n observations respectively. Let $C_{2n} = \{c_j\}_{j=1}^{2n}$ be the list of all the sorted observations of A_n and B_n where c_1 is the smallest observation and c_{2n} the largest. Suppose that $a_i \neq b_j$ for all $i, j = 1, ..., n$. We define the estimation of the dominance rate as

$$\widetilde{\mathcal{C}_\mathcal{D}}(A_n, B_n) = \sum_{i=1}^{2n} \frac{\mathcal{I}(\hat{G}_A(c_j) > \hat{G}_B(c_j) \wedge c_j \in A_n)}{2n} -$$
$$\sum_{i=1}^{2n} \frac{\mathcal{I}(\hat{G}_A(c_j) < \hat{G}_B(c_j) \wedge c_j \in B_n)}{2n} + \frac{1}{2}.$$

where $\mathcal{I}$ is the indicator function and $\hat{G}_A(x)$ and $\hat{G}_B(x)$ are the empirical distributions estimated from A_n and B_n respectively.

For simplicity, this estimator of the dominance rate assumes there are no repeated samples. However, it can be extended to take into account repeated values (see Appendix C).
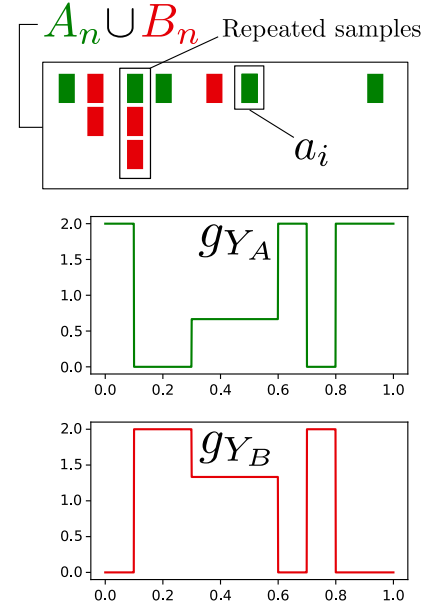

Figure 8. An example of the probability density functions of Y_A and Y_B given the observed samples $\boldsymbol{a} \cup \boldsymbol{b}$.

IV. CUMULATIVE DIFFERENCE-PLOT

In this section, we propose a graphical method called *cumulative difference-plot* that shows the estimations of $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ decomposed by quantiles: $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ can be visually estimated from the difference plot. In addition, the proposed plot allows a comparison of quantiles of the two random variables. The proposed approach also models the uncertainty associated with the estimation of the *cumulative difference-plot* from the data.

A. Quantile random variables

From a practical point of view, it is unlikely that the probability densities of the compared random variables X_A and X_B are known. Usually, we only have n observations $A_n = \{a_1, ..., a_n\}$ and $B_n = \{b_1, ..., b_n\}$ from each random variable. The proposed *cumulative difference-plot* is based on two random variables Y_A and Y_B that are defined with these observations. Specifically, we define the densities of the two *quantile random variables* Y_A and Y_B as a mixture of several uniform distributions in the interval $[0, 1]$.

The uniform distributions in the quantile random variables are placed according to their rank in $A_n \cup B_n$. Assuming no repetitions, for each value k in $A_n \cup B_n$, its corresponding kernel is centered in $\frac{\text{rank}(k)+0.5}{2n}$ where $\text{rank}(k)$ is the rank of k in $A_n \cup B_n$. The kernels have a bandwidth of $1/4n$, ensuring that the sum of the densities of Y_A and Y_B is constant. If there are repeated values in $A_n \cup B_n$, their corresponding kernel is placed at the middle of the previous and the next rank, and the width of the kernel is increased proportionally with respect to the number of repetitions. See Figure 8 for an example. In Appendix B-A we show how to compute the probability densities of Y_A and Y_B step by step.
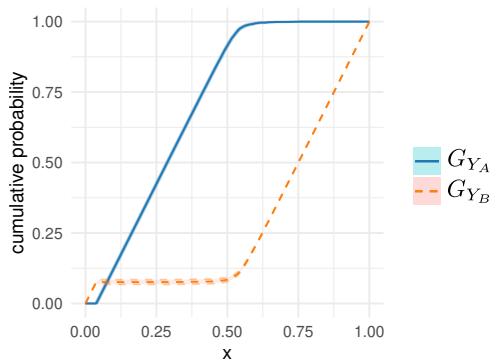
Figure 9. The confidence bands of the cumulative distributions of the quantile random variables Y_A and Y_B corresponding to the distributions X_A and X_B in Case 2 (shown in Figure 4).

As we will later see, Y_A and Y_B have interesting properties essential to define the *cumulative difference-plot*. The quantile random variables have the same $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ as the kernel density estimations of X_A and X_B (shown in Appendix B-B). In addition, $g_{Y_A}(x) + g_{Y_B}(x) = 2$ for all $x \in [0, 1]$.

B. Confidence bands

The *cumulative difference-plot* is based on the cumulative distribution functions of Y_A and Y_B, which are estimated from the observed samples. This means that we need to model the uncertainty associated with the estimations. Confidence bands are a suitable choice in this scenario: a confidence band is a region in which the cumulative distribution is expected to be with a certain confidence. The size of the band is determined by the number of samples and the desired level of confidence: a high number of samples or a low level of confidence are associated with a small band size. There is an extensive literature [22], [23], [24], [25], [26], [27], [28] on how to estimate the confidence bands of cumulative distributions, and, in this work, we use a simple bootstrap approach[8].

To illustrate how to interpret the confidence bands of the cumulative distributions of Y_A and Y_B, we will assume that we have observed $n = 400$ samples from each random variable X_A and X_B from Case 2 (see Figure 4 in Section II-A). We show the 95% confidence bands of the cumulative distribution functions of Y_A and Y_B in Figure 9. The estimated cumulative distribution functions of Y_A and Y_B resemble the cumulative distribution functions of X_A and X_B from Figure 5b. However, there are several relevant differences. In Figure 9, we observe that Y_A and Y_B are defined in the interval $[0, 1]$,

while the cumulative distribution functions of X_A and X_B are defined in the sample space. Each of the values in this interval can be used to deduce the distribution with the lowest quantile: at $x = 0.5$, the cumulative distribution function of Y_A is larger than the cumulative distribution function of Y_B, hence, the median of X_A is lower than the median of X_B. In addition, the sum of the density function of Y_A and Y_B is constant. As a result, unlike X_A and X_B, the probability density functions of Y_A and Y_B do not have large areas where the probability density is zero.

C. The cumulative difference-plot

In this section, we introduce a new graphical method designed to visually analyze the dominance of X_A and X_B. Without loss of generality, a minimization[9] setting is assumed: lower values in X_A and X_B are preferred to higher values. It builds upon the difference function defined as

$$\begin{aligned} \text{diff} : [0,1] &\longrightarrow [-1,1] \\ x &\longmapsto G_{Y_A}(x) - G_{Y_B}(x), \end{aligned} \tag{2}$$

where $G_{Y_A}(x)$ and $G_{Y_B}(x)$ are the cumulative distribution functions of Y_A and Y_B, respectively.

The *cumulative difference-plot* is the plot of the difference function (the difference between the cumulative distributions of Y_A and Y_B), including a confidence band. A positive value in the *cumulative difference-plot* can be interpreted as a quantile in which the cumulative distribution function of X_A is larger than the cumulative distribution function of X_B. Hence, if the difference is positive at 0.5, the median of X_A is lower than the median of X_B (assuming minimization). In this sense, the best values obtained by both random variables are compared on the left side, and the worst values are compared on the right side.

1) $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ in the cumulative difference-plot: $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ can be directly obtained from the proposed plot. The integral of the difference between Y_A and Y_B is $\mathcal{C}_\mathcal{P} - 0.5$ (we prove this in Appendix C). Formally, $\mathcal{C}_\mathcal{P} = 0.5 + \int_0^1 \text{diff}(x)dx$. However, in practice, $\mathcal{C}_\mathcal{P}$ can be visually estimated by adding 0.5 to the difference in the areas over and under 0. For the example shown in Figure 10, $\mathcal{C}_\mathcal{P} = 0.5 - \text{Area1} + \text{Area2}$. The difference can only be in the area highlighted in blue in the cumulative difference-plot. When the probability that $X_A < X_B$ is 1, the

---

[8]The bootstrapping [29] method involves considering the observed values as a population from which random samples with replacement are drawn. These samples are then used to estimate the upper and lower pointwise confidence intervals of the cumulative distribution of Y_A and Y_B. Since a pointwise estimation of the confidence interval is used, we can expect that a portion proportional to α will fall outside the confidence band.
Note that we are interested in having an overall confidence of $1 - \alpha$, thus, we want that the cumulative distributions of Y_A and Y_B are inside their confidence bands at the same time with this level of confidence [30], [31]. This means that we have to use a higher confidence level for each band: $\sqrt{(1 - \alpha)}$.

[9]Note that if the random variables being compared take values in a maximization setting (higher values are preferred), then the random variables need to be redefined as the inverse with respect to the sum (this simply means the sampled values are multiplied by -1) before generating the cumulative difference-plot. With this change, the interpretation of the cumulative difference-plot is consistent and intuitive: for either minimization or maximization, on the left side of the cumulative difference-plot, the most desirable values that the random variables take are compared. If the difference is positive on the left side of the cumulative difference-plot, then the best values that X_A takes are better than the best values that X_B takes. Similarly, the worst values are compared on the right side of the cumulative difference-plot: if the difference is positive on this side, then the worst values of X_A are better than the worst values of X_B.
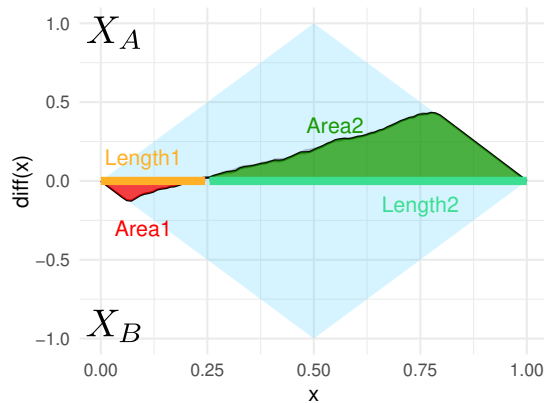
Figure 10. The areas and lengths in the cumulative difference-plot that can be used to deduce $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$.
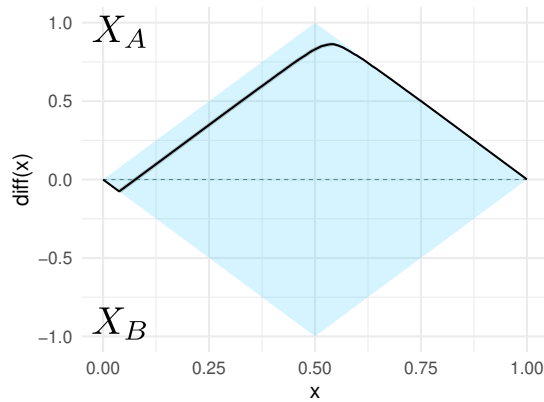

Figure 11. The cumulative difference-plot for Case 2: the difference between the cumulative distribution functions of Y_A minus Y_B corresponding to the distributions X_A and X_B in Case 2 (shown in Figure 4).

difference is at its maximum: in the cumulative difference-plot we see a line from $(x, f(x)) = (0, 0)$ to $(0.5, 1)$ and from $(0.5, 1)$ to $(1, 0)$. Similarly, when the probability that $X_A < X_B$ is 0, the difference between Y_A and Y_B is equal to the lowest possible values inside the light blue area.

By contrast, $\mathcal{C}_\mathcal{D}$ is represented in the plot as the total length in which the difference is positive minus the total length in which the difference is negative. Specifically,

$$\mathcal{C}_\mathcal{D} = \frac{\dfrac{\int_0^1 \mathcal{I}[\mathrm{diff}(x) > 0] - \mathcal{I}[\mathrm{diff}(x) < 0]dx}{2} + \frac{1}{2}}{\int_0^1 \mathcal{I}[\mathrm{diff}(x) \neq 0]dx}, \quad (3)$$

where $\mathcal{I}$ is the indicator function (we prove this in Appendix C). As an example, $\mathcal{C}_\mathcal{D}$ is proportional to Length2 $-$ Length1 in Figure 10: it is higher than 0.5, because Length2 $>$ Length1. In this example, there is no need to divide by the total length in which the difference is nonzero because the difference is zero in only a limited number of points. In such cases, $\mathcal{C}_\mathcal{D}$ can also be estimated as the total length in which $\mathrm{diff}(x) > 0$. In the example in Figure 10, the estimation is $\mathcal{C}_\mathcal{D} = $ Length2 ≈ 0.75. Note that Equation (3) is not correctly defined when Y_A and Y_B are equal, but this is an easy case to identify, as the difference is constantly 0.

2) Illustrative example: Figure 11 shows the cumulative difference-plot for the random variables X_A and X_B from Case 2 (their densities were shown in Figure 4). First, we see that the difference is both negative and positive, hence, neither random variable dominates the other. The difference is negative when $x = 0.05$ or lower. This can be interpreted as X_B having a smaller 5% quantile than X_A. The difference is positive otherwise, thus we deduce from the cumulative difference-plot that the 25%, the 50% (the median), the 75% and 95% quantiles are smaller in X_A than in X_B. In other words, the random variable X_B can take really low values with a small probability, but apart from these really low values, X_A takes lower values than X_B. This is also reflected by $\mathcal{C}_\mathcal{D}(X_A, X_B) > 0.75$, as deduced from $\mathrm{diff}(x) > 0$ for all $x \in (0.25, 1)$. The difference is also near its maximum value, implying that its integral is high and thus $\mathcal{C}_\mathcal{P}$ (the probability that $X_A < X_B$) is also near 1.

V. RELATED WORK

Statistical assessment of experimental results is a very studied research topic. In this section, we locate our proposal in the field and focus on similarities and differences with respect to other random variable comparison methods.

1) Visualizing densities: As mentioned in the introduction, it makes sense to model the performance of stochastic optimization algorithms as random variables. Therefore, statistical tools that compare random variables have become an increasingly important part of the analysis of experimental data. Among these tools, visualization techniques such as histograms or box-plots are usually applied before the rest of the methods. The advantage of these methods is their simplicity. If one of the random variables clearly takes lower values than the other, then these two methods effectively convey this message simply and naturally. Unfortunately, when both random variables have similar probability densities, these two methods might fail to represent the random variables in a way that makes it easy to compare them (example shown in Section VI).

The simplicity of these methods is also a drawback: for example, they give no information about the uncertainty associated with the estimates. The histogram suffers from the bin positioning bias [32], [33]. A kernel density estimation with the uniform kernel—considered to be the moving window equivalent of the histogram—overcomes this limitation [32], at the cost of using a more complex model. Similarly, the box-plot has a "non-injectivity" problem: very different data can still have the same box-plot [1], [2]. The violin-plot [34] is an extension of the box-plot that overcomes the above limitation by combining the kernel density estimate of the random variables with the traditional box-plot. The proposed cumulative difference-plot improves on these methods because it represents the data clearly, even when the two random variables being compared are similar.

2) Null hypothesis statistical testing: Null hypothesis tests can be used to compare random variables without having to visually represent them. In a very general way, carrying

out a null hypothesis test involves the following: first, a null hypothesis is proposed. Under certain assumptions, the null hypothesis implies that a given statistic obtained from the data follows a known distribution. Then, assuming the null hypothesis is true, the probability of obtaining data with a more extreme statistic value[10] than the observed [14] is computed. When the probability under the null hypothesis of the observed statistic is lower than a predefined threshold, the null hypothesis is rejected and the alternative hypothesis is accepted [35]. Usually, this threshold is set at an arbitrary but well established [36] $p = 0.05$, although recently [37], [38], further reducing the threshold to $p = 0.005$ has been proposed.

In the context of comparing two random variables X_A and X_B, in general, we cannot assume that a statistic obtained from the data follows a known distribution under the null hypothesis. In this case, a non-parametric test [14] is a suitable choice. Specifically, the Mann-Whitney [13] test is a good choice, as the samples observed from the random variables are i.i.d for each random variable [11]. With this test, the null hypothesis is that $\mathcal{P}(X_A > X_B) = \mathcal{P}(X_B > X_A)$, and a possible alternative hypothesis is that $X_A \succ X_B$ [13].

Null hypothesis tests have some limitations: for example, the p-value does not separate between the effect size and the sample size [16], [39]. In addition, rejecting the null hypothesis does not always mean that there is evidence in favor of the alternative hypothesis: it just means that the observed statistic (or a more extreme statistic) is very unlikely when the null hypothesis is true.

To show this, we generate 400 samples of the distributions X_A and X_B from Case 2 (density functions shown in Figure 4) and we apply the Mann-Whitney test, rejecting the null hypothesis when $p < 0.005$. If we repeat this experiment 10^4 times (with different samples each time), the null hypothesis is rejected every time[12]. However, $X_A \not\succ X_B$ and $X_B \not\succ X_A$, implying that the alternative hypothesis is not true. Note that the proposed cumulative difference-plot (shown in Figure 11) avoids this problem because it correctly points out that neither random variable dominates the other one, for the same case and with the same number of samples.

3) Bayesian analysis: As an alternative to the limitations of null hypothesis test, Bayesian analysis has been proposed [16], [39]. Bayesian analysis[40], [41] estimates the probability that a hypothesis is true, conditioned to the observed data. This estimation requires the prior probabilities of the hypotheses and the data, but usually, they are assumed to follow a distribution that gives equal probability to all hypotheses or data. Recently a Bayesian version of the Wilcoxon signed-rank test [42], [16] has been proposed. In this paper, we will consider the simplex-plot of its posterior distribution. For

---

[10]The definition of what *data with a more extreme statistic value* is not the same for every null hypothesis test, and it depends on the test being used.

[11]For paired data, the Wilcoxon signed-rank test [15] or the sign test [14] should be used. However, in the context of this paper, the samples observed from the random variables are not paired.

[12]The source code to replicate this experiment is available in the file mann_whitney_counter_example.R in our Github repository.
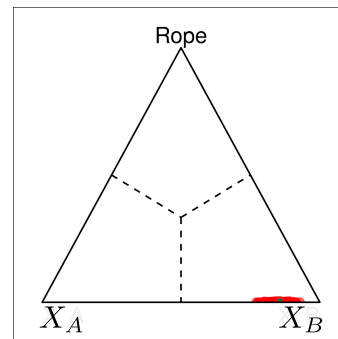

Figure 12. The simplex-plot computed with the package *scmamp* [43] of the posterior distribution for Case 2.

convenience, in the rest of the paper, we will call it *simplex-plot*.

Once the posterior distribution is known, the probability that the difference between a sample from X_A and a sample from X_B is in the interval $(-\infty, -r)$, $[-r, r]$ or $(r, +\infty)$ can be computed. These probabilities can be interpreted as the probability that $X_A > X_B$, $X_A = X_B$ and $X_B > X_A$, where two samples x_a and x_b are considered equal when $|x_a - x_b| \leq r$. Note that the simplex-plot is just a convenient representation of the posterior distribution.

We computed the simplex-plot (Figure 12) with the 400 samples of X_A and X_B from Case 2 obtained in Section IV-B. Two samples were considered equal when their difference is lower than $r = 10^{-3}$, and we used the prior proposed by Benavoli *et al.* [16]. We can deduce from this figure that the hypothesis $X_B > X_A$ is much more likely than $X_A = X_B$ or $X_B > X_A$.

The simplex plot summarizes the data through the probabilities of $X_B > X_A$, $X_A = X_B$ or $X_B > X_A$, but does not offer any additional information: we cannot deduce from these probabilities in which intervals the values of a random variable are lower than the other. In this sense, the cumulative difference-plot is a more detailed visualization. Specifically, the observation that the 1% lowest values of X_A are lower than the 1% lowest values of X_B cannot be deduced from the simplex-plot, while it is easy to see in the cumulative difference-plot. Also, the cumulative difference-plot shows a comparison of the cumulative distributions through the dominance rate, while the simplex-plot does not.

4) Other plots in the interval $[0, 1]$*:* The probability-probability plot is defined as

$$PP : [0, 1] \rightarrow [0, 1]^2 : p \rightarrow (p, G_A(G_B^{-1}(p))).$$

As proposed by Schmid et al. [20], it can be interpreted via the integral of the non-negative part, which represents the amount of violation against the hypothesis that X_A dominates X_B.

The quantile-quantile plot [32], [44] is defined as

$$QQ : [0, 1] \rightarrow N^2 : p \rightarrow (G_A^{-1}(p), G_B^{-1}(p)),$$

and it is a natural way to visualize the differences in quantiles of X_A and X_B in N (the domain of definition of the random variables).

The quantile-quantile plot also allows a comparison between quantiles, just like the *cumulative difference-plot*. However, the cumulative difference-plot proposed in this paper is distinct from the two plots above in three aspects: i) the proposed cumulative difference-plot is defined directly from the observed samples. Because of its definition, it has a confidence band built-in, which allows the uncertainty associated with the estimation to be directly interpreted within the plot. ii) The proposed cumulative difference-plot contains several statistics simultaneously. Specifically, the estimated $\mathcal{C}_\mathcal{D}$, $\mathcal{C}_\mathcal{P}$ and the comparison of the quantiles can be visually interpreted. iii) The proposed plot is just the difference of two cumulative distributions (G_{Y_A} and G_{Y_B}), and thus, unlike in the pp-plot and qq-plot mentioned above, it can be defined without the need of the inverse function. The random variables Y_A and Y_B have the same $\mathcal{C}_\mathcal{D}, \mathcal{C}_\mathcal{P}$ as the kernel density estimations of the original distributions, and therefore, we can think of the cumulative difference-plot as the difference between the cumulative distribution function of two simpler versions of the original random variables.

VI. EXPERIMENTATION WITH THE CUMULATIVE DIFFERENCE-PLOT

To illustrate the applicability of the proposed methodology, in the following, we re-evaluate the experimentation of a recently published work. In a recent paper, Santucci *et al.* [45] introduced a gradient-based optimizer for solving problems defined in the space of permutations (from now on *PL-GS*). In their experimentation, they compared it with an estimation of distribution algorithm [46] (from now on *PL-EDA*). These two algorithms were tested in a set of 50 problem instances of the linear ordering problem [47]. The performance of each algorithm in each instance was estimated with the median relative deviation from the best-known objective value, with $n = 20$ repetitions. From now on, we call *score* to the relative deviation from the best-known objective value and note that a low score is better than a high score, as it means that the objective value found is closer to the best-known.

In the work by Santucci *et al.* [45], when the score of one of the algorithms was at least 10^{-4} higher than the other, it was considered that one of the algorithms performed better than the other in that instance. Santucci *et al.* [45] concluded that both algorithms performed equally in the instance *N-t70n11xx*, as the median scores were exactly the same for both algorithms in this instance. In the following, we take a closer look at the performance of *PL-EDA* and *PL-GS* in this problem instance by comparing $n = 10^3$ measurements of the score from each algorithm. With this increased sample size, we obtained more accurate estimates of the median scores— *PL-GS* $= 0.00407$, *EDA* $= 0.00433$, lower is better—and *PL-GS* obtains a better value by a difference higher than 10^{-4}.
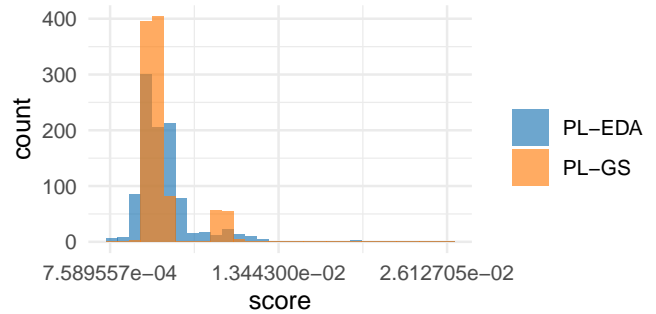
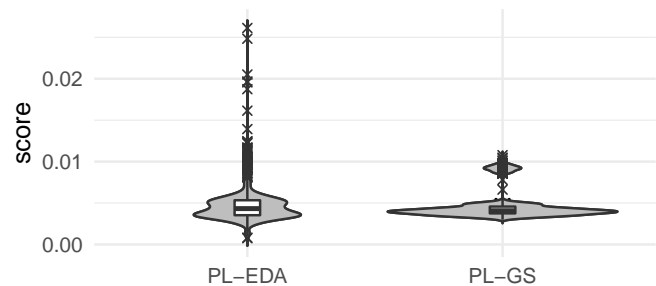Figure 13. Histogram of the scores obtained in the instance *N-t70n11xx*. Lower is better.

Figure 14. Box-plot and violin-plot of the scores obtained in the instance *N-t70n11xx*. Lower is better.

Step 1: Visualization

Figure 13 shows the histogram of the scores. It can be deduced from the figure that neither algorithm clearly produces better scores. In particular, neither algorithm dominates the other: *PL-EDA* has a longer tail both to the right and to the left. Also, notice that the score of the algorithms is not normally distributed: *PL-GS* has a bimodal shape, and *PL-EDA* has a very long tail to the right (while the tail to the left is shorter).

Figure 14 shows the box-plot and the violin-plot of the data. Both algorithms have a similar median, but due to the high number of outliers [18], it is difficult to compare the scores of the algorithms with the box-plot. The same happens with the violin-plot.

Step 2: Comparing PL-GS with PL-EDA

Sometimes, visualization is enough to compare the performance of two algorithms: if one of the algorithms always performs better than the other, there is no need for further analysis. However, in this case, the three visualization methods considered (histogram, box-plot, and violin-plot) have not been able to summarize the scores obtained with the algorithms in a way that enables an easy comparison. In the following, we further study the scores of the algorithms with statistical tests, the simplex-plot, and the cumulative difference-plot.

1) Mann-Whitney test: Applying the Mann-Whitney test we obtain a p-value of $p = 0.035$, lower than the usually used 0.05
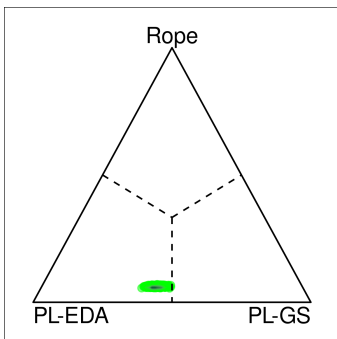
Figure 15. Simplex-plot of *PL-GS* and *PL-EDA* in the instance *N-t70n11xx*. A closer position in the plot to *PL-EDA* indicates a higher probability of measuring a higher score in *PL-EDA* than in *PL-GS*. A low score is preferred to a high score.

Figure 16. The cumulative difference-plot of 95% confidence of the objective values obtained in the instance *N-t70n11xx*.

threshold. With $p < 0.05$, we reject the null hypothesis and accept the alternate hypothesis: the random variable associated with the score of *PL-GS* dominates *PL-EDA*. Note that neither rejecting the null hypothesis nor a small *p*-value reflect the magnitude of the difference in score of the algorithms. In addition, as stated when we studied the histogram, we known that it is unlikely that *PL-GS* dominates *PL-EDA*.

2) Simplex-plot: We show the simplex-plot [16] (as explained in Section V-3) of the scores in Figure 15. Following the criterion by Santucci *et al.* [45], we considered that two scores are equal when they differ by less than $r = 10^{-4}$. Unlike in the statistical test, one can deduce the probability that one of the algorithms has a better score than the other from simplex-plot: it is more likely that *PL-GS* takes a lower value than *PL-EDA*. Specifically, from the simplex-plot shown in Figure 15, we can deduce that given two samples x_{gs} and x_{eda} of the scores of *PL-GS* and *PL-EDA* respectively,

$$\mathcal{P}(x_{eda} < x_{gs}) < \mathcal{P}(x_{gs} < x_{eda}).$$

However, the difference in these probabilities is small. Also, the probability that $\mathcal{P}(x_{gs} = x_{eda})$ is low (two samples are considered equal when their difference score is lower than 10^{-4}).

3) Cumulative difference-plot: We show the 95% confidence cumulative difference-plot in Figure 16. From this plot, we can deduce the following:

1) $\mathcal{P}(x_{eda} < x_{gs})$ and $\mathcal{P}(x_{gs} < x_{eda})$ have similar probabilities, as $\mathcal{C}_{\mathcal{P}}(PL\text{-}EDA, PL\text{-}GS) \approx 0.5$. However, The area under $\text{diff}(x) = 0$ is a little larger than the area over $\text{diff}(x) = 0$, hence $\mathcal{P}(x_{eda} < x_{gs})$ is a little smaller than $\mathcal{P}(x_{gs} < x_{eda})$.
2) Neither algorithm dominates the other one, and what is more, $\mathcal{C}_{\mathcal{D}}(PL\text{-}EDA, PL\text{-}GS) \approx 0.5$.
3) The difference is positive when $x < 0.3$, and therefore, if we only consider the best 30% values of both algorithms, *PL-EDA* dominates *PL-GS*.
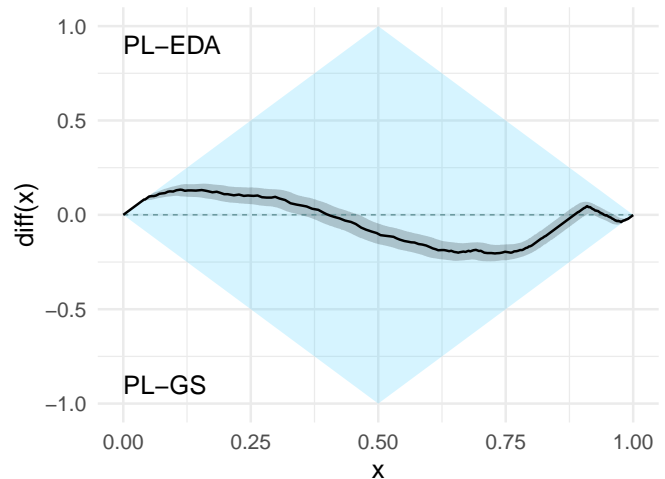4) The difference is negative when $x > 0.98$. In this case, we conclude that if we only consider the worst 2%

values of *PL-EDA* and *PL-GS*, then *PL-GS* dominates *PL-EDA*.
5) These "worst" 2% values are much less likely than the "best" 30% values mentioned in 3), as the estimated probability of these "best" and "worst" values is 0.3 and 0.02 respectively.
6) The difference is negative at $x = 0.5$ and at $x = 0.75$. This can be interpreted as PL-GS having a better median and a better 75% quantile.

Summarizing the above points, we conclude that the performance of the algorithms is quite similar, and *PL-EDA* takes both better and worse scores than *PL-GS*. The probability that *PL-EDA* takes these better values is much higher than the probability that it takes worse values. Therefore, if we are in a setting in which repeating the execution of the algorithms is reasonable, *PL-EDA* is a much better algorithm. On the other hand, if it is critical to avoid really bad values, then *PL-GS* would be preferred. With an increased number of samples, it might be possible to better compare the algorithms (it would reduce the uncertainty associated with the size of the confidence band).

VII. ASSUMPTIONS AND LIMITATIONS

In the following, we briefly summarize the assumptions that the cumulative difference-plot requires and comment on a few caveats.

A. Assumptions

Correctly using the proposed cumulative difference-plot requires that the following three assumptions are satisfied. The first assumption is that all samples of both X_A and X_B are i.i.d, consequently, it should not be used with paired data. This is also an assumption made by the Mann-Whitney test.

The second assumption is that the values of the random variables represent a minimization setting: lower values are preferred to higher values. To apply the proposed method in a maximization setting, it is enough to redefine the objective function by multiplying it by -1.

The third assumption is that X_A and X_B are continuous random variables defined in a connected subset of $\mathbb{R}$. This also implies that the cumulative distribution functions of X_A and X_B are continuous and that their probability density functions are bounded. Although having a bounded density means that there should never be two identical samples—the probability of observing two independent equal samples is 0 with a bounded density—, in reality, the proposed cumulative difference-plot can deal with repeated samples. To do so, when defining the kernel density estimations of Y_A and Y_B in Section IV, repeated samples were assigned the same rank. Then, the size of the uniform distributions was adjusted (with the γ function) ensuring that the sum of the estimated densities of Y_A and Y_B remains constant even in the case of repeated observations.

B. Limitations and future work

Just like with other methods, the number of samples determines in part the stability of the results. With a small sample size, the confidence band of the cumulative difference-plot will be larger. There are three reasons why a larger sample size increases the stability of the plot: i) we are doing a kernel density estimation, and a higher sample size [48] implies that the estimation is closer to the real distribution, ii) the bootstrap method also requires several samples to be meaningful [49], [50] and iii) the sample size needs to be reasonable with respect to the quantiles being estimated. For example, it would not make sense to use 10 samples to estimate a 1% quantile. In all of these cases, however, determining what is a *too small* sample size is a highly debated question, and is beyond the scope of this paper. To be on the safe side, we recommend using a sample size of at least $n = 100$. It is worth noting that this was arbitrarily chosen, and a suitable sample size should be chosen depending on the desired conclusions (for example, comparing small and big quantiles requires more data). With $n = 100$ we ensure that the comparison of 1% quantiles in the cumulative difference-plot is meaningful.

The most obvious limitation of the proposed approach is in its applicability: it should only be used in case of doubt between two random variables, and when none of the random variables dominates the other one. Otherwise, there are more suitable alternatives such as Bayesian analysis [42], [39], or directly comparing box-plots. For instance, if we take 10^3 samples of X_A and X_B and all samples of X_A are lower than all samples of X_B, then there is no need for further statistical comparison, as the results speak for themselves.

The proposed approach assumes X_A and X_B are continuous random variables and that all samples of both X_A and X_B are i.i.d, and consequently, it cannot be used with paired data. As future work, the proposed methodology could be extended for paired data and ordinal random variables. Also, the bootstrap method is the slowest part of the *cumulative difference-plot*, especially as the number of samples increases. To increase the computation speed, this slow part was written in C++ (the rest of the package was written in *R* [51]). However, its speed can probably be further improved with a better implementation.

VIII. CONCLUSION

In this paper, we approached the problem of choosing between two random variables in terms of which of them takes lower values. We proposed eight desirable properties for dominance measures: functions that compare random variables in the context of quantifying the dominance. Among the measures in the literature, we found out that *the probability that one of the random variables takes lower values than the other* was the one that satisfies the most properties. However, it fails to satisfy Property 1, hence it cannot be used to determine when one of the random variables stochastically dominates the other. To overcome this limitation, we introduced a new dominance measure: the dominance rate, which quantifies how much higher one of the cumulative distribution function is than the other.

Based on the above, we proposed a cumulative difference-plot that allows two random variables to be compared in terms of which of them takes lower values. This cumulative difference-plot contains a comparison of the quantiles, in addition to allowing a graphical estimation of the dominance rate and the probability that one of the random variables takes lower values than the others. It also models the uncertainty associated with the estimate through a confidence band. Finally, in Section VI we showed that the proposed methodology is suitable to compare two random variables, especially when they take similar values and other methods fail to give detailed and clear answers.

Alongside this paper, we created a *R* [51] package called RVCompare, to make it convenient for other researchers to use the introduced methodology. It can be downloaded from CRAN. With this package, users can compute the $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ of two distributions, given their probability density functions. Furthermore, it can be used to produce the proposed cumulative difference-plot, given the observed data. In addition, the code to generate every figure in this paper is available in the online repository https://github.com/EtorArza/SupplementaryPaperRVCompare.

REFERENCES

[1] J. Matejka and G. Fitzmaurice, "Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1290–1294.

[2] S. Chatterjee and A. Firat, "Generating Data with Identical Statistics but Dissimilar Graphics: A Follow up to the Anscombe Dataset," *The American Statistician*, vol. 61, no. 3, pp. 248–254, 2007.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[4] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: JMLR Workshop and Conference Proceedings, May 2010, pp. 249–256.

[5] F. Chollet *et al.*, "Keras," 2015.

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

[7] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent," *Scientific Reports*, vol. 7, no. 1, p. 46450, Jun. 2017.

[8] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An Introduction to Deep Reinforcement Learning," *Foundations and Trends in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," *arXiv:1312.5602 [cs]*, Dec. 2013.

[10] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, Mar. 2021.

[11] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, Jul. 2018.

[12] O. Regnier-Coudert, J. McCall, M. Ayodele, and S. Anderson, "Truck and trailer scheduling in a real world, dynamic and heterogeneous context," *Transportation research part E: logistics and transportation review*, vol. 93, pp. 389–408, 2016.

[13] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[14] W. J. Conover and W. J. Conover, "Practical nonparametric statistics," 1980.

[15] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, Dec. 1945.

[16] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, "Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2653–2688, 2017.

[17] J. P. Quirk and R. Saposnik, "Admissibility and measurable utility functions," *The Review of Economic Studies*, vol. 29, no. 2, pp. 140–146, Feb. 1962.

[18] A. Carreño, I. Inza, and J. A. Lozano, "Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework," *Artificial Intelligence Review*, vol. 53, no. 5, pp. 3575–3594, Jun. 2020.

[19] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, Jan. 2011.

[20] F. Schmid and M. Trede, "Testing for First-Order Stochastic Dominance: A New Distribution-Free Test," *The Statistician*, vol. 45, no. 3, p. 371, 1996.

[21] C. J. Bennet, "Inference for Dominance Relations," *International Economic Review*, vol. 54, no. 4, pp. 1309–1328, 2013.

[22] R. C. H. Cheng and T. C. Iles, "Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables," *Technometrics*, vol. 25, no. 1, pp. 77–86, Feb. 1983.

[23] G. P. Steck, "Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions," *The Annals of Mathematical Statistics*, vol. 42, no. 1, pp. 1–11, 1971.

[24] R. C. Cheng and T. Lies, "One-sided confidence bands for cumulative distribution functions," *Technometrics*, vol. 30, no. 2, pp. 155–159, 1988.

[25] J. Wang, F. Cheng, and L. Yang, "Smooth simultaneous confidence bands for cumulative distribution functions," *Journal of Nonparametric Statistics*, vol. 25, no. 2, pp. 395–407, Jun. 2013.

[26] J. J. Faraway and Myoungshic Jhun, "Bootstrap choice of bandwidth for density estimation," *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 1119–1122, 1990.

[27] P. J. Bickel and A. M. Krieger, "Confidence bands for a distribution function using the bootstrap," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 95–100, 1989.

[28] P. Hall and J. Horowitz, "A simple bootstrap method for constructing nonparametric confidence bands for functions," *The Annals of Statistics*, vol. 41, no. 4, Aug. 2013.

[29] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, ser. Monographs on Statistics and Applied Probability. New York: Chapman & Hall, 1993, no. 57.

[30] J. J. Goeman and A. Solari, "Multiple hypothesis testing in genomics," *Statistics in Medicine*, vol. 33, no. 11, pp. 1946–1978, May 2014.

[31] P. Bauer, "Multiple testing in clinical trials," *Statistics in Medicine*, vol. 10, no. 6, pp. 871–890, Jun. 1991.

[32] O. Thas, *Comparing Distributions*. Springer, 2010.

[33] scikit-learn developers, "Density Estimation," https://scikit-learn.org/stable/modules/density.html.

[34] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[35] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations," *European Journal of Epidemiology*, vol. 31, no. 4, pp. 337–350, Apr. 2016.

[36] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, Apr. 2016.

[37] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson, "Redefine statistical significance," *Nature Human Behaviour*, vol. 2, no. 1, pp. 6–10, Jan. 2018.

[38] J. P. Ioannidis, "The proposal to lower P value thresholds to. 005," *Jama*, vol. 319, no. 14, pp. 1429–1430, 2018.

[39] B. Calvo, O. M. Shir, J. Ceberio, C. Doerr, H. Wang, T. Bäck, and J. A. Lozano, "Bayesian performance analysis for black-box optimization benchmarking," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '19*. Prague, Czech Republic: ACM Press, 2019, pp. 1789–1797.

[40] A. Gelman, *Bayesian Data Analysis*, third edition ed., ser. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press, 2014.

[41] J. M. Bernardo and A. F. Smith, *Bayesian Theory*. John Wiley & Sons, 2009, vol. 405.

[42] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri, "A Bayesian Wilcoxon signed-rank test based on the Dirichlet process," in *International Conference on Machine Learning*. PMLR, 2014, pp. 1026–1034.

[43] B. Calvo and G. Santafé Rodrigo, "Scmamp: Statistical comparison of multiple algorithms in multiple problems," *The R Journal, Vol. 8/1, Aug. 2016*, 2016.

[44] M. B. Wilk and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," *Biometrika*, vol. 55, no. 1, p. 1, Mar. 1968.

[45] V. Santucci, J. Ceberio, and M. Baioletti, "Gradient search in the space of permutations: An application for the linear ordering problem," in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*. Cancún Mexico: ACM, Jul. 2020, pp. 1704–1711.

[46] P. Larrañaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Springer Science & Business Media, 2001, vol. 2.

[47] T. Schiavinotto and T. Stützle, "The Linear Ordering Problem: Instances, Search Space Analysis and Algorithms," *Journal of Mathematical Modelling and Algorithms*, vol. 3, no. 4, pp. 367–402, 2004.

[48] Danica and power, "What is the minimum number of data points required for kernel density estimation?"

[49] M. R. Chernick, *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, 2011, vol. 619.

[50] P. Hall, *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media, 2013.

[51] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[52] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[53] A. Rényi *et al.*, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[54] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *MIT (6.441), UIUC (ECE 563), Yale (STAT 664)*, 2012.

[55] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[56] A. Papadopoulos, "Interpretation of the Kullback-Leibler divergence," https://stats.stackexchange.com/questions/111445/analysis-of-kullback-leibler-divergence.

[57] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York ; London: Springer, 2009.

[58] V. N. Vapnik, *Statistical Learning Theory*, ser. Adaptive and Learning Systems for Signal Processing, Communications, and Control. New York: Wiley, 1998.

[59] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional Gaussians," *arXiv:1810.08693 [math, stat]*, May 2020.

[60] a. Xi, "Differences between Bhattacharyya distance and KL divergence."

[61] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communications*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[62] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[63] D. Schuhmacher, "Compute The Wasserstein Distance Between Two Univariate Samples," https://www.rdocumentation.org/packages/transport/versions/0.12-2/topics/wasserstein1d.

[64] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 405–431, 2019.

APPENDICES

APPENDIX A
A LITERATURE REVIEW OF MEASURES

A. f-divergences

The *f-divergence* is a family of functions that can be used to measure the difference between two random variables. Given a strictly convex[13] function $f : (0, +\infty) \to \mathbb{R}$ with $f(1) = 0$, and two continuous random variables X_A and X_B, the *f*-divergence [52], [53] is defined as

$$D_f(X_A, X_B) = \int_{\mathbb{R}} g_B(x) f\left(\frac{g_A(x)}{g_B(x)}\right) dx \quad (4)$$

where g_A and g_B are the probability density functions of the random variables X_A and X_B respectively. Since $g_B(x)$ can be 0, we assume [54] that $0 \cdot f(0/0) = 0$ and $0 \cdot f(a/0) = \lim_{x \to 0+} x \cdot f(a/x)$. Notice that if g_A and g_B are the same probability density functions, then $D_f(X_A, X_B) = 0$.

Kullback–Leibler divergence: The Kullback–Leibler divergence [55] is a particular case of the *f*-divergence, for $f(x) = x \cdot \ln(x)$. Given two random variables X_A and X_B, $D_{KL}(X_A, X_B)$ can be interpreted [56] as the amount of entropy increased by using g_B to model data that follows the probability density function g_A.

The Kullback–Leibler divergence is non-negative, and non symmetric $D_{KL}(X_A, X_B) \neq D_{KL}(X_B, X_A)$, and therefore, it is not actually a distance [3]. It will not satisfy Property (2), as it is not antisymmetric either. This also makes the interpretation less intuitive. The Kullback–Leibler divergence is often used to measure the difference between two random variables [3], but since $D_{KL}(X_A, X_B) \neq D_{KL}(X_B, X_A)$, it may be better to interpret the Kullback–Leibler divergence as stated above [56].

In Figure 17, we show the probability density functions and cumulative distribution functions of four random variables X_A, X_B, X_C and X_D. Looking at their cumulative distributions (Figure 17b), one can clearly see that $X_A \succ X_B$, $X_B \lesseqgtr X_C$ and $X_B \succ X_D$. However, as shown in Table II, $D_{KL}(X_B, X_A) = D_{KL}(X_B, X_C) = D_{KL}(X_B, X_D) = 15.4$ and $D_{KL}(X_A, X_B) = D_{KL}(X_C, X_B) = D_{KL}(X_C, X_D) = 6.2$. This means that, given any two random variables X_A and X_B, the Kullback-Leibler is not able to distinguish if $X_A \succ X_B$, $X_B \succ X_A$ or $X_A \lesseqgtr X_B$. We can interpret this as the Kullback–Leibler divergence only caring about the difference between two random variables, and not if this difference is related to one of the random variables taking lower values than the other. Hence, it cannot satisfy Property 1, even if we try to transform it to be defined in the $[0, 1]$ interval. We conclude that the Kullback–Leibler divergence is not suitable to gain information regarding which of the random variables takes lower values.

---

[13]A function $f : (0, +\infty) \to \mathbb{R}$ is strictly convex if for all $t \in [0, 1]$, for all $x_1, x_2 \in (0, +\infty)$, $f(tx_1 + (1-t)x_1) < tf(x_1) + (1-t)f(x_2)$

(a) Probability density.

(b) Cumulative distribution function.

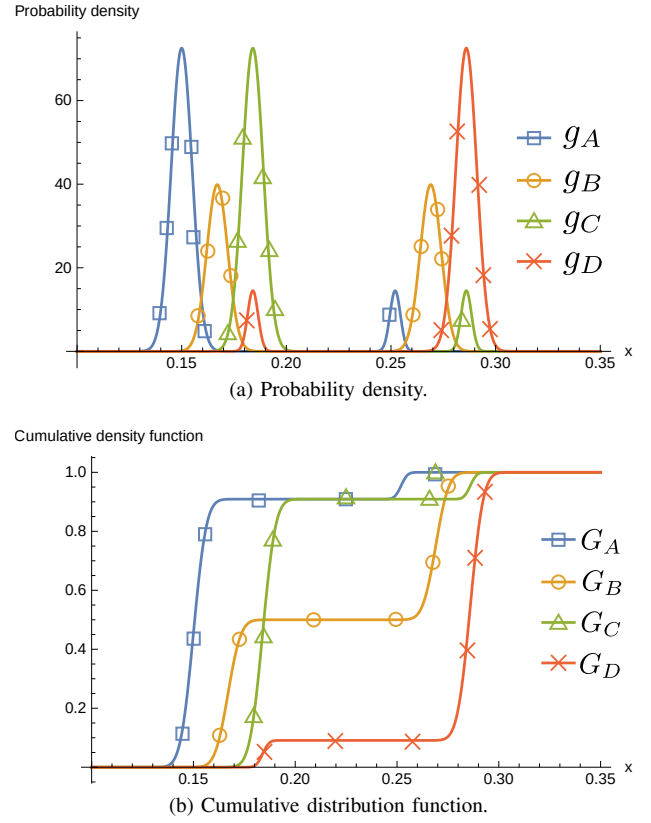Figure 17. The probability density function and cumulative distribution of the four random variables. The distances between these random variables are listed in Table II.

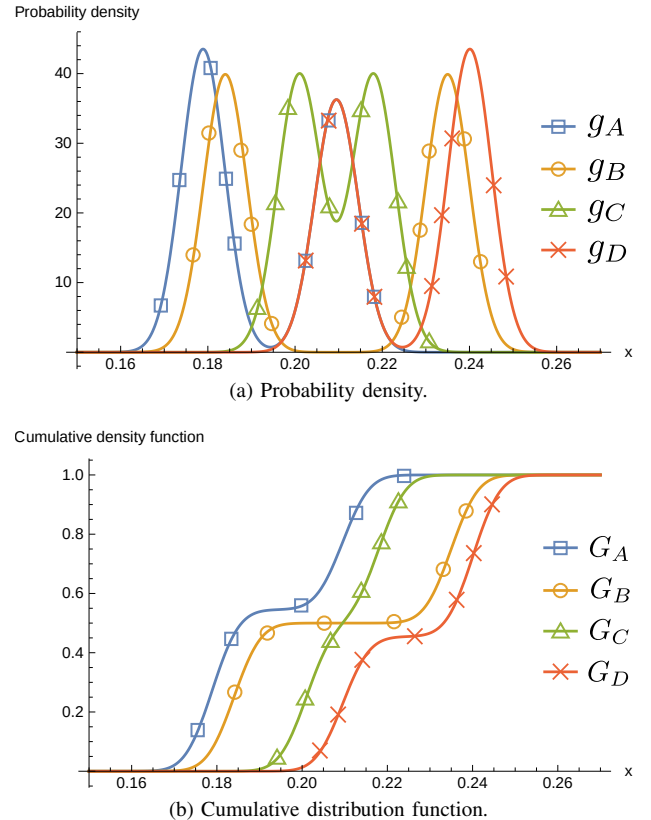(a) Probability density.

(b) Cumulative distribution function.

Figure 18. The probability density function and cumulative distribution of four other random variables. The Wasserstein distance between X_B and each of the other random variables is 0.017.

Table II
$\mathcal{C}(\mathrm{RV}_1, \mathrm{RV}_2)$ FOR THE RANDOM VARIABLES X_A, X_B, X_C AND X_D SHOWN IN FIGURE 17.

Kullback–Leibler

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.0	6.2	28.6	88.8
	X_B	15.4	0.0	15.4	15.4
	X_C	29.4	6.2	0.0	2.6
	X_D	88.8	6.2	2.6	0.0

Jensen-Shannon

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.0	1.2	1.4	1.4
	X_B	1.2	0.0	1.2	1.2
	X_C	1.4	1.2	0.0	0.8
	X_D	1.4	1.2	0.8	0.0

Total variation

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.000	0.934	0.999	1.000
	X_B	0.934	0.000	0.934	0.934
	X_C	0.999	0.934	0.000	0.818
	X_D	1.000	0.934	0.818	0.000

Hellinger

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.00	1.28	1.41	1.41
	X_B	1.28	0.00	1.28	1.28
	X_C	1.41	1.28	0.00	0.99
	X_D	1.41	1.28	0.99	0.00

Wasserstein

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.000	0.06	0.03	0.12
	X_B	0.06	0.00	0.04	0.06
	X_C	0.03	0.04	0.000	0.083
	X_D	0.12	0.06	0.083	0.000

$\mathcal{C}_{\mathcal{P}}$

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.50	0.95	0.92	0.99
	X_B	0.05	0.50	0.54	0.95
	X_C	0.08	0.46	0.50	0.91
	X_D	0.01	0.05	0.09	0.50

$\mathcal{C}_{\mathcal{D}}$

		RV$_2$			
		X_A	X_B	X_C	X_D
RV$_1$	X_A	0.50	1.00	1.00	1.00
	X_B	0.00	0.50	0.59	1.00
	X_C	0.00	0.41	0.50	1.00
	X_D	0.00	0.00	0.00	0.50

Jensen-Shannon divergence: The Jensen-Shannon divergence is very similar to the Kullback-Leibler divergence, and is another the particular case of the f-divergence for $f(x) = x \cdot \ln(\frac{2x}{x+1}) + \ln(\frac{2}{x+1})$ [54]. It is also known as the symmetrized version of the Kullback–Leibler divergence [54], because

$$D_{JS}(X_A, X_B) = D_{KL}(X_A, X_M) + D_{KL}(X_B, X_M)$$

where the probability density function of X_M is $g_M(x) = 0.5(g_A(x) + g_B(x))$. Thus, we can interpret this divergence as the sum of the Kullback–Leibler divergences of g_A and g_B with respect to the average probability density function g_M. The Jensen-Shannon divergence also fails to identify (see

Table II) the dominance relationships between X_B and the rest of the random variables in Figure 17, thus, it cannot satisfy Property 1. In addition, the Jensen-Shannon divergence also fails to satisfy Properties 2 and 3. See Table I for a detailed list of the properties that each measure satisfies.

Total variation: The total variation is also a particular f-divergence, for $f(x) = \frac{1}{2}|x - 1|$ [54]. Unlike the Kullback–Leibler divergence, the total variation is symmetric. In fact, it is a properly defined distance [57], [54]. In addition, it is defined between 0 and 1.

Given two random variables X_A, X_B, the total variation can also be defined as:

$$TV(X_A, X_B) = \sup_{C \subseteq \mathbb{R}} |\mathcal{P}_A(C) - \mathcal{P}_B(C)|,$$

where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the probability distributions[14] of X_A and X_B respectively. Since the subset C that takes the supremum is $C = \{x \in \mathbb{R} \mid g_A(x) > g_B(x)\}$[59], we can interpret the total variation as the "size" of the difference in the density functions in all points where g_A is more likely than g_B. Following this intuition, when $TV(X_A, X_B) = 1$, g_A and g_B have disjoint supports [54], and thus X_A and X_B are at their maximum difference with respect to this metric. On the other hand, when $TV(X_A, X_B) = 0$ the random variables are identical.

The Total-Variance also fails to identify (see Table II) the dominance relationships between X_B and the rest of the random variables in Figure 17.

Hellinger distance and the Bhattacharyya distance: The Hellinger distance is the square root of the f-divergence for $f(x) = (1 - \sqrt{x})^2$ [54]. It is related to the Bhattacharya coefficient, since $D_H(X_A, X_B) = 2(1 - \mathrm{BhattCoef}(X_A, X_B))$ [60], [54], where $\mathrm{BhattCoef}(X_A, X_B)$ is the Bhattacharyya coefficient [61], [62]. This coefficient is defined as $\mathrm{BhattCoef}(X_A, X_B) = \int_{\mathbb{R}} \sqrt{g_A(x)g_B(x)}dx$, and has proven useful on signal processing [61]. Given two probability density functions g_A and g_B, the Bhattacharyya coefficient can be interpreted as the integral of the geometric mean of the probability density functions. The Bhattacharyya coefficient is also related to the Bhattacharyya distance, as $D_{Bhatt}(X_A, X_B) = -\ln(\mathrm{BhattCoef}(X_A, X_B))$.

The Hellinger distance and the Bhattacharyya distance also fail to identify (see Table II) the dominance relationships between X_B and the rest of the random variables in Figure 17.

B. Wasserstein distance

The Wasserstein distance is another type of distance between probability random variables. Given two continuous random

---

[14] Given the random variable X_A defined in $\mathbb{R}$, its probability distribution, noted as $\mathcal{P}_A$, is a mapping that, for all $U \subseteq \mathbb{R}$ that is measurable, $A(U) = \mathcal{P}(X_A \in U)$ [58].

variables X_A, X_B, the Wasserstein distance (of order 1) is defined as [63], [64]

$$D_W(X_A, X_B) = \int_{\mathbb{R}} |G_A(x) - G_B(x)| dx$$

In Figure 18, we show a different set of four random variables X_A, X_B, X_C and X_D. In this case, it is also clear that $X_A \succ X_B$, $X_B \lesseqgtr X_C$ and $X_B \succ X_D$ (Figure 18b), but $D_W(X_B, X_A) = D_W(X_B, X_C) = D_W(X_B, X_D) = 0.017$. Therefore, in this case, the Wasserstein distance does not give any insights about the dominance between X_B and the rest of the random variables, thus, it cannot satisfy Property 1 even with a transformation. It also does not satisfy Properties 2, 3, 6, 7, 8.

However, with a small change, the Wasserstein distance can comply with Properties 2 and 3. This change also improves its correlation with the dominance, even though it still does not comply with Property 1. We remove the absolute value, such that the *signed Wasserstein* distance is defined as

$$D_{SW}(X_A, X_B) = \int_{\mathbb{R}} G_A(x) - G_B(x) dx.$$

For the random variables in Figure 18, the signed Wasserstein distance has different values: $D_{SW}(X_B, X_A) = 0.17$, $D_{SW}(X_B, X_C) = 0$ and $D_{SW}(X_B, X_D) = -0.017$. Observe that $X_A \succ X_B \implies D_{SW}(X_B, X_A) > 0$ and $X_B \succ X_A \implies D_{SW}(X_B, X_A) < 0$, but unfortunately, when $X_A \lesseqgtr X_B$, $D_{SW}(X_B, X_A)$ could be positive or negative. This implies that $D_{SW}(X_B, X_A)$ still can not determine if $X_A \succ X_B$, $X_B \succ X_A$, or $X_A \lesseqgtr X_B$.

C. Heuristic derivation of the first-order stochastic dominance

A measure similar to the Wasserstein distance has been proposed in the literature [20] in the context of comparing random variables. Specifically, this measure is part of the heuristic derivation of a distribution-free statistical test for first-order stochastic dominance [20]. Given two random variables X_A, X_B, this measure is defined as

$$\mathcal{C}_I(X_A, X_B) = \int_{\mathbb{R}} max(0, G_A(x) - G_B(x)) dG_B(x).$$

Note that the values of I range between 0 and 0.5. When $\mathcal{C}_I(X_A, X_B) = 0.5$, we know that $X_A \succ X_B$. Unfortunately, when $\mathcal{C}_I(X_A, X_B) \in (0, 0.5)$, it could be that $X_A \succ X_B$ or $X_A \not\succ X_B$. Consequently, $\mathcal{C}_I(X_A, X_B)$ cannot satisfy Property 1.

APPENDIX B
QUANTILE RANDOM VARIABLES

A. Computing the probability density functions of Y_A and Y_B

In Section IV-A we introduced the quantile random variables Y_A and Y_B. We now describe how to compute the probability density functions of g_{Y_A} and g_{Y_B} step by step, with the pseudocode shown in Algorithm 1. We define a function r that returns the position of an observation according to its rank in the sorted list of the observation $A_n \cup B_n$ (lines 1–4). The ranks go from 0 (for the smallest observation) to r_{max} (for the largest), where r_{max} is the number of unique observation in $A_n \cup B_n$ minus 1. Repeated observations are assigned the same rank, and no ranks are skipped: there is at least a value in $\boldsymbol{a} \cup \boldsymbol{b}$ corresponding to each rank from 0 to r_{max}. For each observation in $\{a_1, ..., a_n\}$, a uniform distribution defined in the interval $(\frac{r(a_i) + \gamma(r(a_i) - 1)}{2n}, \frac{r(a_i) + \gamma(r(a_i))}{2n})$ is added to the mixture (lines 10–19), where $\gamma(k)$ (lines 7–9) counts the number of ranks in $A_n \cup B_n$ that are lower than or equal to k (since the lowest rank is 0, $\gamma(-1) = 0$) . The kernel density estimation for Y_B is defined similarly, but with the observations $\{b_1, ..., b_n\}$ instead.

B. The quantile random variables have the same $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ as the kernel density estimates of X_A and X_B.

In Section IV-A, we claimed that when a "small enough" uniform [33] kernel is used in the kernel density estimations of X_A and X_B, these estimations will have the same $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ as the quantile random variables Y_A and Y_B. Specifically, the size of the uniform kernels needs to be smaller than $\min_{i,j \in \{1...n\} | a_i \neq b_j} 2|a_i - b_j|$, where $A_n = \{a_1, ..., a_n\}$ and $B_n = \{b_1, ..., b_n\}$ are the n observed samples of X_A and X_B respectively. As a result, the $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ of the kernel density estimations will not change when the size of the kernels is reduced below its initial size. This can be deduced from Property 8 in Section II-B, which both $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ satisfy.

The quantile random variables Y_A and Y_B can also be obtained by applying a sequence of transformations to the kernel density estimations (with small uniform kernels) of X_A and X_B. Three consecutive transformations are required, none of which modify the $\mathcal{C}_\mathcal{D}$ and $\mathcal{C}_\mathcal{P}$ due to Property 8. The first transformation involves further *reducing* the size of the kernels to $1/(4n)$. Secondly, each kernel k is moved into the position $r(k)/(2n) + (4n)^{-1}$, where $r(k)$ is the rank of the sample in k in $A_n \cup B_n$. In the case of ties, r assigns the same rank to all kernels and this same rank is the average of the previous and the next rank. Since each of the possible positions are at distance $1/(2n)$ from each other, this transformation will not change the $\mathcal{C}_\mathcal{D}$ and $\mathcal{C}_\mathcal{P}$. Finally, the length of the kernels is increased to *mult*/$(4n)$, where *mult* is the number of times that the sample defining the kernel is repeated in $A_n \cup B_n$. Note that this increase in the length will in no case cause an overlap of kernels.

Algorithm 1: Kernel density estimation of Y_A and Y_B

Input:

$A_n = \{a_1, ..., a_n\}$: The n observed samples of X_A.
$B_n = \{b_1, ..., b_n\}$: The n observed samples of X_B.

Output:

g_{Y_A}: The probability density of Y_A.
g_{Y_B}: The probability density of Y_B.

```
/* Compute the ranks of An ∪ Bn. The lowest
   value has rank 0. Assign the same rank to
   ties without skipping any rank.          */
```
1 for $i = 1,...,n$ **do**
2 $r(a_i) \leftarrow$ rank of a_i in $A_n \cup B_n$
3 $r(b_i) \leftarrow$ rank of b_i in $A_n \cup B_n$
4 end
5 $R \leftarrow \{r(a_1), ..., r(a_n), r(b_1), ..., r(b_n)\}$
6 $r_{max} \leftarrow \max(R)$
7 for $k = -1,0,1,...,r_{max}$ **do**
8 $\gamma(k) \leftarrow$ the number of items in R lower than or equal to k
9 end
```
/* The probability density function of gYA is
   represented as a mixture of n uniform
   distributions. gYA[s] is the probability
   density of YA in the interval [s/2n, s+1/2n).  */
```
10 $g_{Y_A} \leftarrow$ array of zeros of length $2n$
11 $g_{Y_B} \leftarrow$ array of zeros of length $2n$
12 for $x_i = a_1, ..., a_n, b_1, ..., b_n$ **do**
13 $A_{mult} \leftarrow$ number of times that x_i is in A_n
14 $B_{mult} \leftarrow$ number of times that x_i is in B_n
15   **for** $mult = 1,...,(A_{mult} + B_{mult})$ **do**
16 $g_{Y_A}[\gamma(r(a_i) - 1) + mult - 1] \leftarrow (n \cdot A_{mult})^{-1}$
17 $g_{Y_B}[\gamma(r(b_i) - 1) + mult - 1] \leftarrow (n \cdot B_{mult})^{-1}$
18   **end**
19 end
20 return g_{Y_A}, g_{Y_B}

APPENDIX C
$\mathcal{C}_\mathcal{P}$ AND $\mathcal{C}_\mathcal{D}$ IN THE CUMULATIVE DIFFERENCE-PLOT

In this section, we mathematically prove and experimentally verify that the cumulative difference-plot can be used to deduce $\mathcal{C}_\mathcal{D}$ and $\mathcal{C}_\mathcal{P}$. First, we describe which estimators are used when these dominance measures are visually estimated from the cumulative difference-plot. Then, we show that these estimators converge to $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ as the number of samples increases.

A. Estimating $\mathcal{C}_\mathcal{P}$ and $\mathcal{C}_\mathcal{D}$ from the cumulative difference-plot

Definition 7. *(observations of random variables)*
Let X_A be a continuous random variable. We define n observations of X_A as the realizations of the i.i.d random variables $\{X_A^i\}_{i=1}^n$ that are distributed as X_A, denoted as $A_n = \{a_i\}_{i=1}^n$.

Definition 8. *(estimation of $\mathcal{C}_\mathcal{P}$)*
Let X_A and X_B be two continuous random variables and A_n and B_n their n observations respectively. We define the estimation of the probability that $X_A < X_B$ as

$$\widetilde{\mathcal{C}_\mathcal{P}}(A_n, B_n) = \sum_{i,k=1...n} \frac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2}.$$

Definition 9. *(estimation of $\mathcal{C}_\mathcal{D}$)*
Let X_A and X_B be two continuous random variables and A_n and B_n their n observations respectively. Let $\{c_j\}_{j=1}^{2n}$ the sorted list of all the observations of A_n and B_n where c_1 is the smallest observation and c_{2n} the largest. Let $\{c_d\}_{d=1}^{d_{max}}$ be the sorted list of unique values in $\{c_j\}_{j=1}^{2n}$. We define the estimation of the dominance rate as

$$\widetilde{\mathcal{C}_\mathcal{D}}(A_n, B_n) = \frac{\sum_{j=1}^{2n} \frac{\psi(c_j)}{2n} + 1}{2} \cdot k_c^{-1}$$

$k_c = \frac{\sum_{j=1}^{2n} \mathcal{I}[\psi(c_j) \neq 0]}{2n}$ *is the normalization constant and ψ_j is defined as*

$$\psi(c_d) = \begin{cases} 0 & \text{if } \hat{G}_A(c_{d-1}) = \hat{G}_B(c_{d-1}) \\ & \text{and } \hat{G}_A(c_d) = \hat{G}_B(c_d) \\ 1 & \text{if } \hat{G}_A(c_{d-1}) \geq \hat{G}_B(c_{d-1}) \\ & \text{and } \hat{G}_A(c_d) > \hat{G}_B(c_d) \\ 1 & \text{if } \hat{G}_A(c_{d-1}) > \hat{G}_B(c_{d-1}) \\ & \text{and } \hat{G}_A(c_d) \geq \hat{G}_B(c_d) \\ -1 & \text{if } \hat{G}_B(c_{d-1}) \geq \hat{G}_A(c_{d-1}) \\ & \text{and } \hat{G}_B(c_d) > \hat{G}_A(c_d) \\ -1 & \text{if } \hat{G}_B(c_{d-1}) > \hat{G}_A(c_{d-1}) \\ & \text{and } \hat{G}_B(c_d) \geq \hat{G}_A(c_d) \\ 1 - 2\gamma(c_d) & \text{if } \hat{G}_B(c_{d-1}) > \hat{G}_A(c_{d-1}) \\ & \text{and } \hat{G}_A(c_d) > \hat{G}_B(c_d) \\ 2\gamma(c_d) - 1 & \text{if } \hat{G}_A(c_{d-1}) > \hat{G}_B(c_{d-1}) \\ & \text{and } \hat{G}_B(c_d) > \hat{G}_A(c_d) \end{cases}$$

with $\gamma(c_d) = \frac{\hat{G}_B(c_{d-1}) - \hat{G}_A(c_{d-1})}{[B_n = c_d] - [A_n = c_d]}$. Note that $[A_n = c_d]$ counts the number of items in A_n equal to c_d and $\hat{G}_A$ is the empirical distribution [23] estimated from A_n. To improve the readability, we abuse the notation and assume that $\hat{G}_A(c_0) = 0$.

We now show that these estimates can be directly computed from the cumulative difference plot. First, we show that the estimation of $\mathcal{C}_\mathcal{P}$ from the cumulative difference-plot is equivalent to the estimation in Definition 8. As mentioned in Section IV-C, the $\mathcal{C}_\mathcal{P}$ estimated from the cumulative difference-plot is $0.5 + \int_0^1 \text{diff}(x)dx$ where diff is the difference function introduced in Equation (2). Specifically, the difference function was defined as $\text{diff}(x) = G_{Y_A}(x) - G_{Y_B}(x)$.

Lemma 1. *Let X_A and X_B be two continuous random variables and A_n and B_n their n observations respectively. Then,*

$$\int_0^1 \text{diff}(x)dx = \sum_{j=1}^{2n} \frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n}$$

Proof. Considering that the density functions of Y_A and Y_B are constant in each interval $[\frac{j}{2n}, \frac{j+1}{2n})$ for $j = 0, ..., (2n-1)$, we get that

$$\int_{\frac{j}{2n}}^{\frac{j+1}{2n}} \text{diff}(x)dx = \frac{\text{diff}(\frac{j}{2n}) + \text{diff}(\frac{j+1}{2n})}{4n} =$$

$$\frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n}) + G_{Y_A}(\frac{j+1}{2n}) - G_{Y_B}(\frac{j+1}{2n})}{4n}$$

Taking into account that $G_{Y_A}(0) = G_{Y_B}(0) = 0$ and $G_{Y_A}(1) = G_{Y_B}(1) = 1$,

$$\int_0^1 \text{diff}(x)dx = \sum_{j=0}^{2n-1} \int_{\frac{j}{2n}}^{\frac{j+1}{2n}} \text{diff}(x)dx =$$

$$\frac{G_{Y_A}(\frac{0}{2n}) - G_{Y_B}(\frac{0}{2n}) + G_{Y_A}(\frac{2n}{2n}) - G_{Y_B}(\frac{2n}{2n})}{4n} +$$

$$\sum_{j=1}^{2n-1} \frac{2 \cdot G_{Y_A}(\frac{j}{2n}) - 2 \cdot G_{Y_B}(\frac{j}{2n})}{4n} =$$

$$\sum_{j=1}^{2n-1} \frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n}$$

Finally, since $G_{Y_A}(1) = G_{Y_B}(1) = 1$, we have that

$$\sum_{j=1}^{2n-1} \frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n} =$$

$$\sum_{j=1}^{2n} \frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n}$$

$\square$

Proposition 3. *($\mathcal{C}_\mathcal{P}$ estimated from the cumulative difference-plot)*
s Let X_A and X_B be two random variables and A_n and B_n their n observations respectively. Let diff *be the difference function obtained from the samples A_n and B_n as defined in Equation (2). Then,*

$$\widetilde{\mathcal{C}_\mathcal{D}}(A_n, B_n) = \int_0^1 \text{diff}(x)dx + \frac{1}{2}$$

Proof. Given the observations A_n and B_n, we need to prove that

$\sum_{i,k=1...n} \dfrac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2} = \int_0^1 \text{diff}(x)dx + \frac{1}{2}$

With Lemma 1, it is enough to prove that

$\sum_{i,k=1...n} \dfrac{\text{sign}(b_k - a_i)}{2n^2} = \sum_{j=1}^{2n} \dfrac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n}$

Let $C_{2n} = \{c_j\}_{j=1}^{2n}$ be the list of all the sorted observations of A_n and B_n where c_1 is the smallest observation and c_{2n} the largest. Then, we have that

$$G_{Y_A}\left(\frac{j}{2n}\right) = \frac{[A_n < c_j] + \dfrac{[A_n = c_j][k \le j | c_k = c_j]}{[C_{2n} = c_j]}}{n} \quad \text{and}$$

$$G_{Y_B}\left(\frac{j}{2n}\right) = \frac{[B_n < c_j] + \dfrac{[B_n = c_j][k \le j | c_k = c_j]}{[C_{2n} = c_j]}}{n}$$

where $[A_n < c_j]$ counts the number of items in A_n lower than c_j, and $[k \le j | c_k = c_j]$ counts the number of items in C_{2n} equal to c_j but with a lower or equal position in C_{2n}. Therefore, we have that

$$\sum_{j=1}^{2n} \frac{G_{Y_A}(\frac{j}{2n}) - G_{Y_B}(\frac{j}{2n})}{2n} =$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] + \dfrac{[A_n = c_j][k \le j | c_k = c_j]}{[C_{2n} = c_j]} - [B_n < c_j] - \dfrac{[B_n = c_j][k \le j | c_k = c_j]}{[C_{2n} = c_j]}}{2n^2}$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j] + \dfrac{([A_n = c_j] - [B_n = c_j])[k \le j | c_k = c_j]}{[C_{2n} = c_j]}}{2n^2} \tag{5}$$

Now we group the terms in Equation (5) into d_{max} groups such that each group contains all the terms with the same c_j, and each group d contains $[C_{2n} = c_d]$ terms, with $c_j = c_d$.

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{d=1}^{d_{max}} \sum_{c_j} \frac{\dfrac{([A_n = c_j] - [B_n = c_j])[k \le j | c_k = c_j]}{[C_{2n} = c_j]}}{2n^2} =$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{d=1}^{d_{max}} \frac{\dfrac{([A_n = c_d] - [B_n = c_d])(([C_{2n} = c_d] + 1) \cdot [C_{2n} = c_d]/2)}{[C_{2n} = c_d]}}{2n^2} =$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{d=1}^{d_{max}} \frac{([A_n = c_d] - [B_n = c_d])([C_{2n} = c_d] + 1)/2}{2n^2} =$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{d=1}^{d_{max}} \frac{([A_n = c_d] - [B_n = c_d])([C_{2n} = c_d])/2 + ([A_n = c_d] - [B_n = c_d])/2}{2n^2} =$$

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{j=1}^{2n} \frac{([A_n = c_j] - [B_n = c_j])/2}{2n^2} + \underbrace{\sum_{d=1}^{d_{max}} \frac{([A_n = c_d] - [B_n = c_d])/2}{2n^2}}_{=0} =$$

$$\underbrace{\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2}}_{\text{first sum}} + \underbrace{\sum_{j=1}^{2n} \frac{([A_n = c_j] - [B_n = c_j])/2}{2n^2}}_{\text{second sum}}$$

Focusing on the first sum, we have that

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} =$$

$$\frac{\sum_{j=1}^{2n}[A_n < c_j] - \sum_{j=1}^{2n}[B_n < c_j]}{2n^2} =$$

$$\frac{\sum_{j=1}^{2n} \sum_{i=1}^{n}[\{a_i\} < c_j] - \sum_{j=1}^{2n} \sum_{i=1}^{n}[\{b_i\} < c_j]}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n}[\{a_i\} < a_k] + \sum_{k=1}^{n} \sum_{i=1}^{n}[\{a_i\} < b_k]}{2n^2} -$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n}[\{b_i\} < a_k] + \sum_{k=1}^{n} \sum_{i=1}^{n}[\{b_i\} < b_k]}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n}[\{a_i\} < a_k] + [\{a_i\} < b_k] - [\{b_i\} < a_k] - [\{b_i\} < b_k]}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n}[\{a_i\} < b_k] - [\{b_i\} < a_k] + [\{a_i\} < a_k] - [\{b_i\} < b_k]}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n} \text{sign}(b_k - a_i) + [\{a_i\} < a_k] - [\{b_i\} < b_k]}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n} \text{sign}(b_k - a_i)}{2n^2} + \frac{\sum_{k=1}^{n}[A_n < a_k] - [B_n < b_k]}{2n^2}$$

From the second sum, we obtain

$$\sum_{j=1}^{2n} \frac{([A_n = c_j] - [B_n = c_j])/2}{2n^2} = \sum_{k=1}^{n} \frac{([A_n = a_k] - [B_n = a_k] + [A_n = b_k] - [B_n = b_k])/2}{2n^2}$$

Combining these summations,

$$\sum_{j=1}^{2n} \frac{[A_n < c_j] - [B_n < c_j]}{2n^2} + \sum_{j=1}^{2n} \frac{([A_n = c_j] - [B_n = c_j])/2}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n} \text{sign}(b_k - a_i)}{2n^2} + \frac{\sum_{k=1}^{n}[A_n < a_k] - [B_n < b_k]}{2n^2} + \frac{\sum_{k=1}^{n}([A_n = a_k] - [B_n = a_k] + [A_n = b_k] - [B_n = b_k])/2}{2n^2} =$$

$$\frac{\sum_{k=1}^{n} \sum_{i=1}^{n} \text{sign}(b_k - a_i)}{2n^2} + \frac{\sum_{k=1}^{n}[A_n \leq a_k] - [B_n \leq b_k]}{2n^2} + \frac{\sum_{k=1}^{n}(-[A_n = a_k] - [B_n = a_k] + [A_n = b_k] + [B_n = b_k])/2}{2n^2} =$$

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{k=1}^{n}[A_n \leq a_k]-[B_n \leq b_k]}{2n^2} + \frac{\sum_{k=1}^{n}(-[C_{2n}=a_k]+[C_{2n}=b_k])}{4n^2} =$$

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{n(n+1)/2+\sum_{d=1}^{d_{max}}\frac{[A_n=c_d]^2-[A_n=c_d]}{2} - (n(n+1)/2+\sum_{d=1}^{d_{max}}\frac{[B_n=c_d]^2-[B_n=c_d]}{2})}{2n^2} +$$
$$\frac{\sum_{k=1}^{n}(-[C_{2n}=a_k]+[C_{2n}=b_k])}{4n^2} =$$

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}\frac{[A_n=c_d]^2-[A_n=c_d]}{2} - \sum_{d=1}^{d_{max}}\frac{[B_n=c_d]^2-[B_n=c_d]}{2}}{2n^2} + \frac{\sum_{k=1}^{n}(-[C_{2n}=a_k]+[C_{2n}=b_k])}{4n^2} =$$

considering that $\sum_{d=1}^{d_{max}}\frac{[B_n=c_d]-[A_n=c_d]}{2} = 0$, we simplify the previous equation to

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}\frac{[A_n=c_d]^2-[B_n=c_d]^2}{2}}{2n^2} + \frac{\sum_{k=1}^{n}(-[C_{2n}=a_k]+[C_{2n}=b_k])}{4n^2} =$$

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}[A_n=c_d]^2-[B_n=c_d]^2}{4n^2} + \frac{\sum_{k=1}^{n}(-[C_{2n}=a_k]+[C_{2n}=b_k])}{4n^2} =$$
$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}[A_n=c_d]^2-[B_n=c_d]^2}{4n^2} + \frac{\sum_{d=1}^{d_{max}}(-[C_{2n}=c_d][A_n=c_d]+[C_{2n}=c_d][B_n=c_d])}{4n^2} =$$
$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}[A_n=c_d]^2-[B_n=c_d]^2}{4n^2} + \underbrace{\frac{\sum_{d=1}^{d_{max}}[C_{2n}=c_d]([B_n=c_d]-[A_n=c_d])}{4n^2}}_{\text{third sum}} =$$

We expand the third sum,

$$\frac{\sum_{d=1}^{d_{max}}[C_{2n}=c_d]([B_n=c_d]-[A_n=c_d])}{4n^2} =$$

$$\frac{\sum_{d=1}^{d_{max}}([B_n=c_d]+[A_n=c_d])([B_n=c_d]-[A_n=c_d])}{4n^2} = \frac{\sum_{d=1}^{d_{max}}([B_n=c_d]^2-[A_n=c_d]^2)}{4n^2}$$

Finally,

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2} + \frac{\sum_{d=1}^{d_{max}}[A_n=c_d]^2-[B_n=c_d]^2}{4n^2} + \frac{\sum_{d=1}^{d_{max}}([B_n=c_d]^2-[A_n=c_d]^2)}{4n^2} =$$

$$\frac{\sum_{k=1}^{n}\sum_{i=1}^{n}\mathrm{sign}(b_k-a_i)}{2n^2}$$

$\square$

Proposition 4. *Let X_A and X_B be two random variables and A_n and B_n their n observations respectively. The $\mathcal{C}_\mathcal{D}$ estimated from the cumulative difference-plot is $\widetilde{\mathcal{C}_\mathcal{D}}$.*

Proof. In Section IV-C, we defined the $\mathcal{C}_\mathcal{D}$ estimated from the cumulative difference-plot as

$$\mathcal{C}_\mathcal{D} = \frac{\frac{\int_0^1 \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx}{2} + \frac{1}{2}}{\int_0^1 \mathcal{I}[\text{diff}(x) \neq 0]dx},$$

where $\mathcal{I}$ is the indicator function. This proposition claims that

$$\frac{\frac{\int_0^1 \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx}{2} + \frac{1}{2}}{\int_0^1 \mathcal{I}[\text{diff}(x) \neq 0]dx} =$$

$$\frac{\sum_{j=1}^{2n} \frac{\psi(c_j)}{2n} + 1}{2} \cdot k_c^{-1}.$$

To prove it, we show that

i) $\int_0^1 \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx = \sum_{j=1}^{2n} \frac{\psi(c_j)}{2n}$

and

ii) $\int_0^1 \mathcal{I}[\text{diff}(x) \neq 0]dx = k_c.$

Let us focus our attention in *i)*. We split the integral into $2n$ parts:

$$\int_0^1 \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx =$$

$$\sum_{j=1}^{2n} \int_{\frac{j-1}{2n}}^{\frac{j}{2n}} \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx \quad (6)$$

Let $C_{2n} = \{c_j\}_{j=1}^{2n}$ be the list of all the sorted observations of A_n and B_n where c_1 is the smallest observation and c_{2n} the largest and let $\{c_d\}_{d=1}^{d_{max}}$ be the sorted list of unique values in C_{2n}. We group the terms in the sum of Equation (6) into d_{max} groups such that for every j in a group, $c_j = c_d$.

$$\sum_{d=1}^{d_{max}} \sum_{j} \int_{\frac{j-1}{2n}}^{\frac{j}{2n}} \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx$$

Now we join the integrals for every j in each group, such that the j of the integral goes from $j_{d\downarrow} - 1$ to $j_{d\uparrow}$ (if the sample c_d is unique in C_{2n}, then $j_{d\downarrow} = j_{d\uparrow} = j$).

$$\sum_{d=1}^{d_{max}} \int_{\frac{j_{d\downarrow}-1}{2n}}^{\frac{j_{d\uparrow}}{2n}} \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx \quad (7)$$

In the interval $(\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$, diff evaluates to one of these four possibilities:

1) $\text{diff}(x) = 0$ for all $x \in (\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$
2) $\text{diff}(x) > 0$ for all $x \in (\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$
3) $\text{diff}(x) < 0$ for all $x \in (\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$
4) $\text{diff}(x) = 0$ in one point in the interval $(\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$ and $\text{diff}(x) > 0$ or $\text{diff}(x) < 0$ for every other x in the interval. However, we can safely ignore this point as the value of the integral is invariant to the value of the function in sets of zero measure.

By looking at the empirical distributions $\hat{G}_A(x)$ and $\hat{G}_B(x)$ estimated from A_n and B_n respectively, we can guess which of these possibilities corresponds to each interval.

$$\begin{cases}
1) & \text{if } \hat{G}_A(c_{d-1}) = \hat{G}_B(c_{d-1}) \text{ and } \hat{G}_A(c_d) = \hat{G}_B(c_d) \\
2) & \text{if } \hat{G}_A(c_{d-1}) \geq \hat{G}_B(c_{d-1}) \text{ and } \hat{G}_A(c_d) > \hat{G}_B(c_d) \\
2) & \text{if } \hat{G}_A(c_{d-1}) > \hat{G}_B(c_{d-1}) \text{ and } \hat{G}_A(c_d) \geq \hat{G}_B(c_d) \\
3) & \text{if } \hat{G}_B(c_{d-1}) \geq \hat{G}_A(c_{d-1}) \text{ and } \hat{G}_B(c_d) > \hat{G}_A(c_d) \\
3) & \text{if } \hat{G}_B(c_{d-1}) > \hat{G}_A(c_{d-1}) \text{ and } \hat{G}_B(c_d) \geq \hat{G}_A(c_d) \\
4) & \text{if } \hat{G}_B(c_{d-1}) > \hat{G}_A(c_{d-1}) \text{ and } \hat{G}_A(c_d) > \hat{G}_B(c_d) \\
4) & \text{if } \hat{G}_A(c_{d-1}) > \hat{G}_B(c_{d-1}) \text{ and } \hat{G}_B(c_d) > \hat{G}_A(c_d)
\end{cases}$$

The value of the integral in Equation (7) corresponding to these possibilities are the following:

1) 0
2) $[C_{2n} = c_d] \cdot \frac{1}{2n}$
3) $-[C_{2n} = c_d] \cdot \frac{1}{2n}$
4) $[C_{2n} = c_d] \cdot (2 \cdot l_d - 1) \cdot \frac{1}{2n}$

where $[C_{2n} = c_d]$ counts the number of items in C_{2n} equal to c_d and l_d is the proportion in which $\text{diff}(x) > 0$ in the interval $(\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$. For example, $l_d = 0.75$ would represent that $\text{diff}(x) > 0$ in 75% of the total length of the interval, and $\text{diff}(x) < 0$ in the other 25%.

With this, we can rewrite Equation (7) as

$$\sum_{d=1}^{d_{max}} [C_{2n} = c_d] \cdot \psi(c_d) \cdot \frac{1}{2n} = \sum_{j=1}^{2n} \frac{\psi(c_j)}{2n},$$

where ψ is the function introduced in Definition 9.

Now, we only need to prove *ii)*. Specifically, we need to show that

$$\int_0^1 \mathcal{I}[\text{diff}(x) \neq 0]dx = k_c.$$

We have that

$$\int_0^1 \mathcal{I}[\text{diff}(x) \neq 0]dx = \sum_{d=1}^{d_{max}} \int_{\frac{j_{d\downarrow}-1}{2n}}^{\frac{j_{d\uparrow}}{2n}} \mathcal{I}[\text{diff}(x) \neq 0]dx,$$

and

$$k_c = \frac{\sum_{j=1}^{2n} \mathcal{I}[\psi(c_j) \neq 0]}{2n} =$$

$$\sum_{d=1}^{d_{max}} [C_{2n} = c_d] \frac{\mathcal{I}[\psi(c_d) \neq 0]}{2n}.$$

Finally, it is easy to see that

$$\int_{\frac{j_{d\downarrow}-1}{2n}}^{\frac{j_{d\uparrow}}{2n}} \mathcal{I}[\text{diff}(x) \neq 0]dx = [C_{2n} = c_d] \frac{\mathcal{I}[\psi(c_d) \neq 0]}{2n},$$

because $\text{diff}(x) = 0$ in the interval $(\frac{j_{d\downarrow}-1}{2n}, \frac{j_{d\uparrow}}{2n})$ if and only if $\psi(c_d) = 0$.

$\square$

B. Convergence of the estimators

Proposition 5. *Let X_A and X_B be two <u>continuous</u> random variables and $\{a_i\}_{i \in \mathbb{N}}$ and $\{b_i\}_{i \in \mathbb{N}}$ be two infinite sequences of their observations respectively. Let A_n and B_n be the two finite subsequences that contain the first n elements of $\{a_i\}_{i \in \mathbb{N}}$ and $\{b_i\}_{i \in \mathbb{N}}$ respectively. Then,*

$$\mathcal{C}_{\mathcal{P}}(X_A, X_B) = \lim_{n \to \infty} \widetilde{\mathcal{C}_{\mathcal{P}}}(A_n, B_n)$$

Proof. Let $\{P_s\}_{s \in \mathbb{N}}$ be a sequence of estimators with every estimator is determined randomly with the following procedure:

1) generate two random permutations σ_s and τ_s of size n.

2) define each estimation as

$$P_s(A_n, B_n) = \sum_{i=1}^n \frac{\text{sign}(b_{\sigma_s(i)} - a_{\tau_s(i)})}{2n} + \frac{1}{2}.$$

It is easy to see that each P_s is an estimator of $\mathcal{P}(X_A < X_B)$ (since X_A, X_B are continuous, we know that $\mathcal{P}(X_A = X_B) = 0$). Now observe that the sequence $\left\{ \frac{\sum_{t=1}^s P_t(A_n, B_n)}{s} \right\}_{n \in \mathbb{N}}$ converges to $\widetilde{\mathcal{C}_{\mathcal{P}}}(A_n, B_n) = \sum_{i,k=1...n} \frac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2}$, which means that $\widetilde{\mathcal{C}_{\mathcal{P}}}(A_n, B_n)$ is also an estimator of $\mathcal{P}(X_A < X_B)$.

$\square$

Unfortunately, the estimator $\widetilde{\mathcal{C}_{\mathcal{D}}}$ will not always converge: $\mathcal{C}_{\mathcal{D}}$ fails to satisfy Property 7, and this means that a a few points can still have a big impact in the estimation of $\mathcal{C}_{\mathcal{D}}$. Specifically, given the continuous random variables X_A and X_B defined in N, $\widetilde{\mathcal{C}_{\mathcal{D}}}$ will converge iff $\int_N \mathcal{I}[G_A(x) = G_B(x)] \cdot (g_A + g_B)dx = 0$.

Luckily, this lack of convergence is not a problem when the estimation of $\mathcal{C}_{\mathcal{D}}$ is carried out visually in the cumulative difference-plot. Since the visual representation of the cumulative difference-plot involves rendering the plot with pixels, there exists an small $\delta > 0$ such that when $|\text{diff}(x)| < \delta$, the difference is displayed as 0.

In practice, we do not even need to account for the case that $\text{diff}(x) = 0$. The cumulative difference-plot models the uncertainty with a confidence band, and when $\text{diff}(x) = 0$ is inside the confidence band, then so are $\text{diff}(x) > 0$ and $\text{diff}(x) < 0$. If we assume that the difference is positive, negative or zero every time that $\text{diff}(x) = 0$ is inside the confidence band, we obtain the estimations $\widetilde{\mathcal{C}_{\mathcal{D}}}^+, \widetilde{\mathcal{C}_{\mathcal{D}}}^-$ and $\widetilde{\mathcal{C}_{\mathcal{D}}}^0$ respectively. Now since $\widetilde{\mathcal{C}_{\mathcal{D}}}^+ > \widetilde{\mathcal{C}_{\mathcal{D}}}^= > \widetilde{\mathcal{C}_{\mathcal{D}}}^-$, the estimation of $\mathcal{C}_{\mathcal{D}}$ with the highest part of the confidence band is an upper bound of $\mathcal{C}_{\mathcal{D}}$. The same is true for the estimation with the lowest part of the confidence band: it is a lower bound of $\mathcal{C}_{\mathcal{D}}$.

Although $\widetilde{\mathcal{C}_{\mathcal{D}}}$ does not converge to $\mathcal{C}_{\mathcal{D}}$, for any $\epsilon > 0$ we can find a δ small enough such that the difference between $\widetilde{\mathcal{C}_{\mathcal{D}}}^\delta$ and $\mathcal{C}_{\mathcal{D}}$ is smaller than ϵ. We formalize this claim in Conjecture 1, and we leave the proof for future work.

Definition 10. *(δ-estimation of $\mathcal{C}_{\mathcal{D}}$)*
Let X_A and X_B be two continuous random variables and A_n and B_n their n observations respectively. Let $\{c_j\}_{j=1}^{2n}$ the sorted list of all the observations of A_n and B_n where c_1 is the smallest observation and c_{2n} the largest. Let $\{c_d\}_{d=1}^{d_{max}}$ be the sorted list of unique values in $\{c_j\}_{j=1}^{2n}$.

We define the δ-estimation of $\mathcal{C}_{\mathcal{D}}$, denoted as $\widetilde{\mathcal{C}_{\mathcal{D}}}^\delta$, as the same estimation as $\widetilde{\mathcal{C}_{\mathcal{D}}}$, but assuming that the empirical distributions computed from A_n and B_n are equal when $|\hat{G}_A(x) - \hat{G}_B(x)| < \delta$.

The previous definition can also be based in the δ-difference, defined as $\text{diff}^\delta(x) = \text{diff}(x)$ when $\text{diff}(x) \geq \delta$, and $\text{diff}^\delta(x) = 0$ otherwise.

Conjecture 1. *Let X_A and X_B be two <u>continuous</u> random variables and $\{a_i\}_{i \in \mathbb{N}}$ and $\{b_i\}_{i \in \mathbb{N}}$ be two infinite sequences of their observations respectively. Let A_n and B_n be the two finite subsequences that contain the first n elements of $\{a_i\}_{i \in \mathbb{N}}$ and $\{b_i\}_{i \in \mathbb{N}}$ respectively. Then, for all $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$\left| \mathcal{C}_{\mathcal{D}}(X_A, X_B) - \lim_{n \to \infty} \widetilde{\mathcal{C}_{\mathcal{D}}}^\delta(A_n, B_n) \right| < \epsilon$$

C. Experimental verification

In the following, we experimentally verify that the cumulative difference-plot can be used to deduce $\mathcal{C}_{\mathcal{D}}$ and $\mathcal{C}_{\mathcal{P}}$. To do so, we define six pairs of example random variables and measure the $\mathcal{C}_{\mathcal{P}}$ and $\mathcal{C}_{\mathcal{D}}$ with three different methods: the definition of $\mathcal{C}_{\mathcal{D}}$ and $\mathcal{C}_{\mathcal{P}}$ (Equation (1) and Definition 4), the estimators in Definitions 8 and 9 and from the *cumulative difference-plot*. The *cumulative difference-plot* has a confidence band in addition to the estimation, and this confidence band allows the lower and upper bounds of $\mathcal{C}_{\mathcal{D}}$ and $\mathcal{C}_{\mathcal{P}}$ to be computed.

The probability density functions of the six examples are shown in Figures 19 through 24. The probability density of these random variables is a mix of normal distributions, the beta distribution, and the log-normal distribution.

The difference plot and the estimations were carried out with 5000 samples from each random variable. The $\mathcal{C}_{\mathcal{P}}$ and $\mathcal{C}_{\mathcal{D}}$ values computed are shown in Figures 25 and 26 respectively. In every case, the estimations with the three methods match, except for $\mathcal{C}_{\mathcal{D}}$ in Example 4 (Figure 22). This is a deceptive example because, in most of the probability mass of X_A and X_B, the cumulative distribution functions are equal. Consequently, in this example, the estimator of $\mathcal{C}_{\mathcal{D}}$ introduced in Definition 9 is unstable: it is very likely that the estimated empirical distributions are different even though the cumulative distribution functions are identical. Overcoming this limitation involves choosing a small $\delta > 0$, such that when the difference between the empirical distributions is smaller than δ, they are considered equal.

We conclude that, in most cases, the three estimation methods (from densities, using the estimators and with the cumulative difference-plot) yield a similar result, which validates the statements in the previous section.
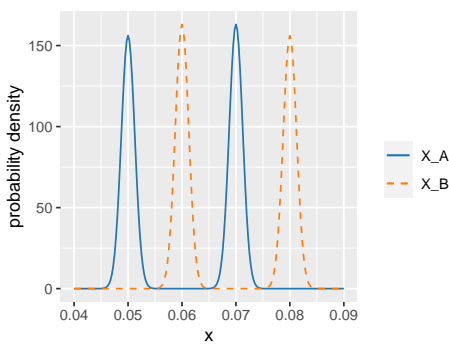

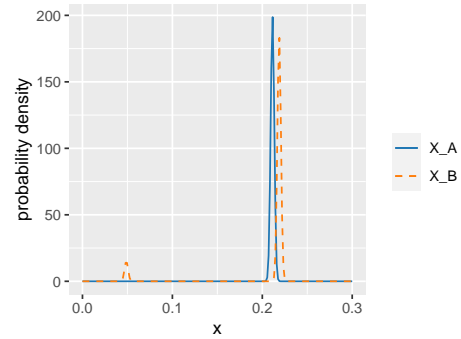Figure 19. Probability density functions of Example 1

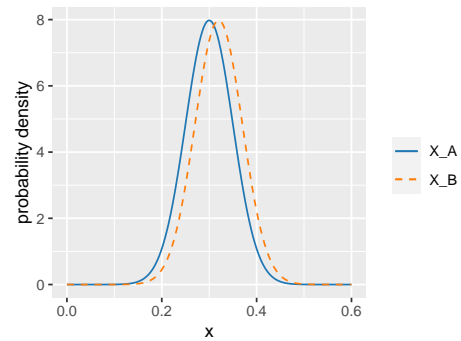Figure 20. Probability density functions of Example 2

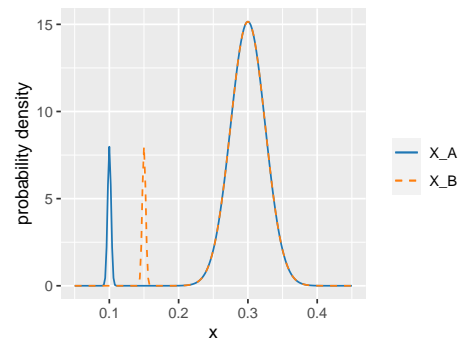Figure 21. Probability density functions of Example 3
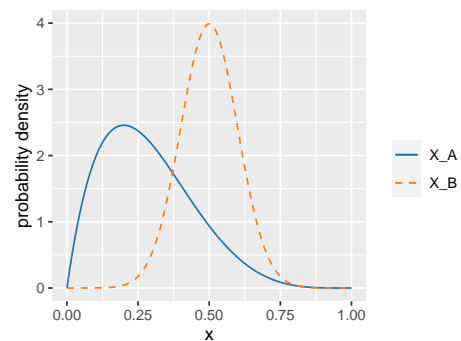

Figure 22. Probability density functions of Example 4

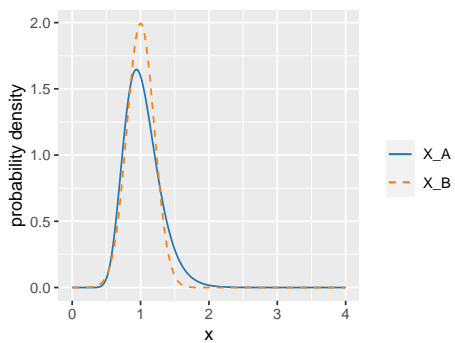

Figure 23. Probability density functions of Example 5

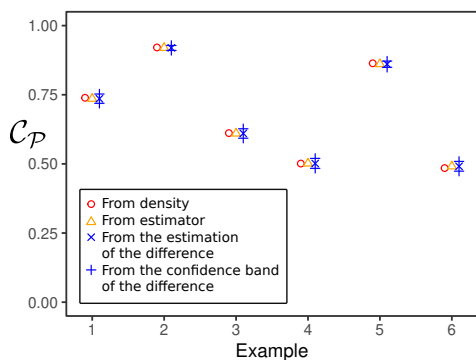Figure 24. Probability density functions of Example 6

Figure 25. The $\mathcal{C}_\mathcal{P}$ values obtained in the six examples with the three methods.
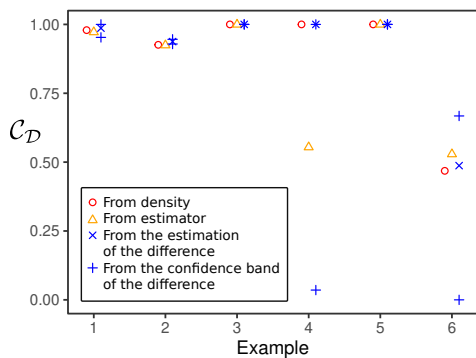

Figure 26. The $\mathcal{C}_\mathcal{D}$ values obtained in the six examples with the three methods.