Comparing Two Samples Through Stochastic Dominance: A Graphical Approach

Etor Arza\(^a\), Josu Ceberio\(^b\), Ekhiñe Irurozki\(^c\), and Aritz Pérez\(^a\)

\(^a\)BCAM – Basque Center for Applied Mathematics, Bilbao, Spain; \(^b\)University of the Basque Country UPV/EHU, Leioa, Spain; \(^c\)Télécom Paris, Palaiseau, France

ABSTRACT
Nondeterministic 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 article, we propose an alternative framework to visually compare two samples according to their estimated 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 stochastically dominates the other one. Then, we present a graphical method that decomposes in quantiles (i) the proposed dominance measure and (ii) the probability that one of the random variables takes lower values than the other. With illustrative purposes, we reevaluate 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, the software package \texttt{RVC}\ Compare was created as a convenient way of applying and experimenting with the proposed framework.

ARTICLE HISTORY
Received December 2021
Accepted May 2022

KEYWORDS
Cumulative distribution function; Data visualization; First-order stochastic dominance; Random variables

1. Introduction

The objective value obtained by an optimization algorithm may be nondeterministic. 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 nondeterministic 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 (Matejka and Fitzmaurice 2017; Chatterjee and Firat 2007).

1.1. 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 error rates (Goodfellow, Bengio, and Courville 2016) 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 (Glorot and Bengio 2010).

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 dataset: adam and RMSProp (Goodfellow, Bengio, and Courville 2016). 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 estimation 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.

1.2. 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...
1.3. 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 (Mann and Whitney 1947; Conover and Conover 1980; Wilcoxon 1945) to analyze observed samples and choose one of the random variables accordingly. Nonetheless, as claimed in Benavoli et al. (2017), 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 (Benavoli et al. 2017).

1.4. Contribution

In this article, 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 eight 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 (Quirk and Saposnik 1962). 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 reevaluating 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 article. With this package, the cumulative difference-plot can be conveniently computed. The source code of the package and the supplementary material for the article are available at github.com/EtorArza.2

The rest of the article is organized as follows: in the next section, we propose eight desirable properties for dominance measures. Then, in Section 3, we study two dominance measures that satisfy most of these properties. Section 4 introduces a

---

2The source of the package RVCompare can be found at github.com/EtorArza/RVCompare. The code to reproduce every figure in the article is available at github.com/EtorArza/SupplementaryPaperRVCompare.
2. Desirable Properties for Dominance Measures

2.1. 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 (Conover and Conover 1980) of a continuous random variable \( X_A \), denoted as \( m_A \), is the value that satisfies \( P(X_A < m_A) = P(X_A > m_A) \). 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 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 0.07, 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 \( \text{E}[X_A] < \text{E}[X_B] \) 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 \( \text{E}[X_A] > \text{E}[X_B] \), hence, the comparison of the medians is justified; the intuition that \( X_A \) takes lower values than \( X_B \), while the expected values are not. In the presence of outliers (Carreño, Inza, and Lozano 2020), the median is considered more robust than the expected value (Rousseeuw and Hubert 2011).

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 \( X_B \) (Mann and Whitney 1947). Depending on the field of study, it can also be referred to (Schmid and Trede 1996; Bennet 2013; Quirk and Sapounik 1962) 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 (Schmid and Trede 1996; Quirk and Sapounik 1962), although, for the sake of brevity, we will call it stochastic dominance throughout the article.

**Definition 1 (Stochastic dominance).** Let \( X_A \) and \( X_B \) be two continuous random variables defined in a connected subset \( N \subseteq \mathbb{R} \). We say that \( X_A \) stochastically 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.

³Without loss of generality, minimization is assumed in this article.
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 stochastically dominates the other. In the context of this article, we are interested in measures that, given two random variables, quantify through the stochastic 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 \( C \) that maps two random variables into a real value \( C(X_A, X_B) \).

It is desirable that \( C(X_A, X_B) \) quantifies the stochastic dominance. Hence, we want \( C(X_A, X_B) \) to be proportional to the portion of the support of \( X_A \) and \( X_B \) in which \( G_A(x) < G_B(x) \). Formally, this intuitive idea can be represented as:

**Property 1.** \( C \) is defined in the \([0, 1]\) interval, where:

(i) \( C(X_A, X_B) = 1 \iff X_A > X_B \)

(ii) \( C(X_A, X_B) = 0 \iff X_B > X_A \)

(iii) \( C(X_A, X_B) \in (0, 1) \iff X_B \leq X_A \).

**Proof.** By definition, \( X_B \leq X_A \iff X_A \neq X_B \) and \( X_B \neq X_A \). Property 1 (i) and (ii) implies that \( X_A \neq X_B \) and \( X_B \neq X_A \) if \( C(X_A, X_B) \neq 1 \) and \( C(X_A, X_B) \neq 0 \). From Property 1 (i) also \( C(X_A, X_B) \in [0, 1] \), thus \( X_B \leq X_A \) iff \( C(X_A, X_B) \in (0, 1) \).

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

\[ C(X_A, X_B) = 1 - C(X_B, X_A). \]

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 \( C \) equals the inversion of \( C \):

\[ C(-1 \cdot X_A, -1 \cdot X_B) = 1 - C(X_A, X_B). \]

**Property 4.** When \( X_A \) and \( X_B \) are equal, \( C \) is symmetric.

\[ X_A = X_B \implies C(X_A, X_B) = C(X_B, X_A). \]

---

4Note that \( X_A \neq X_B \) is not equivalent to \( X_B > X_A \).
Assuming Property 2 holds, we can rewrite the previous property as:

\[ X_A = X_B \implies C(X_A, X_B) = 0.5. \]

Note that the opposite is not true:

\[ C(X_A, X_B) = C(X_B, X_A) \Rightarrow X_A = X_B. \]

**Property 5 (Invariance to translation).** Moving the domain of definition of \( X_A \) and \( X_B \) by the same amount does not change \( C \).\(^5\)

for all \( \lambda \in \mathbb{R} , \ C(X_A + \lambda, X_B + \lambda) = C(X_A, X_B). \)

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

\[ \text{for all } \lambda > 0, \ C(\lambda \cdot X_A, \lambda \cdot X_B) = C(X_A, X_B). \]

In the following lines, we give an intuition for Property 7. In Case 2, shown in Figure 4, we saw that for all \( x \in (0.075, 0.2) \), \( G_A(x) < G_B(x) \). However, notice that most of the mass of \( X_A \) and \( X_B \) is in the interval \((0.2, 0.23)\), where \( G_A(x) > G_B(x) \). This means that most of the observed points of \( X_A \) and \( X_B \) will be in that interval. Therefore, it makes sense that \( G_A(x) > G_B(x) \) has a higher weight than \( G_A(x) < G_B(x) \) in the computation \( C(X_A, X_B) \). In other words, the small mass of \( X_B \) centered in 0.05 can only account for a small part of \( C(X_A, X_B) \). 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 the computation of \( C \) produced by the distribution of small weight in the mixture can be, at most, its weight in the mixture.

**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 \( \tau \), respectively, and let \( X_A \) be another random variable. Then,

\[ |C(X_A, X_B) − C(X_A, X_{B1})| ≤ \tau. \]

Property 8 explains that, under certain circumstances, \( C(X_A, X_B) \) is invariant to the translation/dilatation of one of the random variables. Specifically, it states that the distribution of one of the random variables \( X_B \) can change without affecting the value of \( C(X_A, X_B) \) as long as the changed part does not overlap with the support of the other random variable \( X_A \). Let us assume that the random variable \( X_B \) is defined as mixture distribution \( \mathcal{M}_{[1−\rho,\rho]}(X_{B1}, X_{B2}) \) where the supports of \( X_{B2} \) and \( X_A \) do not overlap, with \( \rho \in (0, 1) \). Property 8 states that a translation and/or dilatation can be applied to \( X_{B1} \), as long as: (i) this transformation does not cause an overlap of the supports of \( X_A \) and \( X_{B2} \), and (ii) partial transformations will also not cause an overlap (hence the need for \( \xi_1 \) and \( \xi_2 \) in Property 8). In the following, we formalize this property:

\[ \text{Figure 6. Case 3. The probability density functions of } X_A \text{ and } X_B: g_A = g_{\mathcal{R}_1(0.2, 0.2)} \text{ and } g_B = 0.925 - g_{\mathcal{R}_1(0.19, 0.2)} + 0.075 - g_{\mathcal{R}_1(0.04, 0.05)} \text{, respectively, where } \mathcal{R}_1(0.2, 0.2) \text{ is the uniform distribution in the interval } (0.2, 0.2). \]

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

\[ C(X_A, X_B) = C(X_A, \mathcal{M}_{[1−\rho,\rho]}(X_{B1}, 1 \cdot X_{B2} + \lambda_2)). \]

This property can be applied to the distributions in Case 3 shown in Figure 6. For example, the probability mass in the interval \((0.04, 0.05)\) could have been centered in 0.1 or 0.15 instead of 0.045, without any changes to \( C(X_A, X_B) \). In addition to the position, the shape of the mass can also be altered as long as its weight in the mixture stays the same and does not overlap with \( X_A \).

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 \( X_A > X_{B1} \) and at the same time \( X_B > X_A \) with \( \tau < 0.5. \)\(^7\) We formalize and prove this claim in the following proposition:

**Proposition 2.** Let \( C \) be a dominance measure.

(i) If \( C \) satisfies Property 1, then it fails to satisfy Property 7.

(ii) If \( C \) 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) \( X_A > X_{B1} \),

(iii) \( X_B > X_A \).

If four random variables can be found that satisfy these three statements, then with Property 1 we obtain that \( C(X_A, X_B) = 1 \)

\(^5\)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 \( \lambda \) to this observation. We define \( \lambda \cdot X_A \) in a similar way.

\(^6\)The probability density function of \( \mathcal{M}_{[1−\tau,\tau]}(X_{B1}, X_{B2}) \) is defined as \((1−\tau) \cdot g_{B1}(x) + \tau \cdot g_{B2}(x)\). 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.
and $C(X_A, X_B) = 0$. This contradicts Property 7, because $|C(X_A, X_B) - C(X_A, X_{B1})| \neq 0.1$. The same is true the other way around, Property 7 states that $|C(X_A, X_B) - C(X_A, X_{B1})| \leq 0.1$ and this contradicts Property 1, with $C(X_A, X_{B1}) < 1$ or $C(X_A, X_B) > 0$.

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

- $X_A = U(0, 1)$,
- $X_B = U([0,0.01], U(0.1, 1), U(-0.5, 0))$,
- $X_{B1} = U(0.1, 1)$,
- $X_{B2} = U(-0.5, 0)$.

The cumulative distribution functions of $X_A$, $X_B$ and $X_{B1}$ are shown in Figure 7, where it is clear that $X_B > X_A$ and $X_A > X_{B1}$.

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 (refer to the online Appendices 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 1).

### 3. Dominance Measures

Most of the measures in the literature fail to satisfy the eight properties introduced in Section 2.2. However, there is a dominance measure in the literature that overcomes this limitation: the probability that $X_A < X_B$ (Conover and Conover 1980).

![Figure 7. The cumulative distribution functions of $X_A$, $X_B$, and $X_{B1}$.](image)

![Figure 8. Implications between the values of $C_P$ and $C_D$.](image)

\[
C_P(X_A, X_B) = 1
\]

\[
C_D(X_A, X_B) = 1
\]

\[
C_P(X_A, X_B) > 0.5 \quad C_D(X_A, X_B) > 0.5
\]

### Table 1. Which of the properties in Section 2.2 does each measure satisfy?

| Property | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----------|---|---|---|---|---|---|---|---|
| Kullback–Leibler divergence | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Jensen–Shannon divergence | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Total-Variation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Hellinger distance | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Wasserstein distance | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $C_P$: Probability of $X_A < X_B$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| $C_D$: Dominance rate of $X_A$ over $X_B$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

A checkmark ✓ indicates that the measure satisfies the property.

### 3.1. $C_P$: 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:

\[
P(X_A < X_B) = \int_N g_B(x)G_A(x)dx.
\]

(1)

When we consider $P(X_A < X_B)$ as a dominance measure, we will denote it as $C_P(X_A, X_B)$.

One of the advantages of $C_P$ is its easy interpretation. In addition, $C_P$ 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:

\[
C_P(X_A, X_B) = 1 \implies X_A > X_B \implies
\]

\[
C_P(X_A, X_B) \in (0.5, 1]
\]

and

\[
C_P(X_A, X_B) = 0 \implies X_B > X_A \implies
\]

\[
C_P(X_A, X_B) \in [0, 0.5).
\]

Note that, when $X_A > X_B$, $C_P(X_A, X_B) \neq 1$ 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 $X_A > X_B$ but $C_P(X_A, X_B) < 1$.

So far, we have seen that $C_P$ satisfies most of the properties. Unfortunately, since it does not satisfy Property 1, not all cases of $X_A > X_B$ can be identified by $C_P$. We now propose a dominance measure that satisfies Property 1 and, thus, can be used to identify cases in which $X_A > X_B$. 
3.2. \( C_D \): 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:

\[
D_{X_A, X_B}(x) = \begin{cases} 
  g_A(x) \cdot k_A & \text{if } G_A(x) > G_B(x) \\
  -g_B(x) \cdot k_B & \text{if } G_A(x) < G_B(x) \\
  0 & \text{otherwise,}
\end{cases}
\]

where \( k_A = \left( \int_{\mathbb{R}} [G_A(x) \neq G_B(x)] \ g_A(t) \, dt \right)^{-1} \) is the normalization constant and \( k_B \) is defined likewise.

Note that the dominance density function is not correctly defined when \( \int_{\mathbb{R}} |g_A(x) - g_B(x)| \, dx = 0 \).

**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

\[
C_D(X_A, X_B) = \begin{cases} 
  0.5, & \text{if } \int_{\mathbb{R}} |g_A(x) - g_B(x)| \, dx = 0 \\
  0.5 \int_{\mathbb{R}} D_{X_A, X_B}(t) \, dt + 0.5, & \text{otherwise.}
\end{cases}
\]

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, that is, \( \int_{\mathbb{R}} D_{X_A, X_B}(t) \, dt = \frac{E_A[\mathbb{I}[G_A(x) > G_B(x)]] - E_B[\mathbb{I}[G_A(x) < G_B(x)]]}{E_A[\mathbb{I}[G_A(x) \neq G_B(x)]]} \).

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

From

i) \( C_D(X_A, X_B) = 1 \iff X_A > X_B \) and

ii) \( C_D(X_A, X_B) = 0 \iff X_B > X_A \),

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 \( \int_{\mathbb{R}} \mathbb{I}[x \in N \cap (G_A(0) > G_B(0))] \ g_A(x) \, dx > 0 \), and this point is guaranteed to exist when \( X_A > X_B \) 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 \( C_P \) and \( C_D \) satisfy most of the properties listed in Section 2.2. As we will see in the next section, their values are related.

3.3. The Relationship Between \( C_P \) and \( C_D \)

In Section 2.1 we stated that \( C_P = 1 \) is a stronger condition than \( C_D = 1 \), because \( C_P(X_A, X_B) = 1 \) implies that for all \( x \in N \) that \( G_A(x) < 1, G_B(x) = 0 \). On the other hand, \( C_D = 1 \) implies that \( X_A > X_B \) (the two conditions in Definition 1), which is weaker. In Figure 8, we show the values of \( C_P \) and \( C_D \) that imply other values of \( C_P \) and \( C_D \). Each arrow can be interpreted as an implication. The implications are transitive: i.e. \( C_D(X_A, X_B) = 1 \) implies \( C_P(X_A, X_B) > 0.5 \).

3.4. Estimating \( C_P \) and \( C_D \)

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 \( C_P \) and \( C_D \) from the observed samples. With this purpose, we propose the following empirical estimates of \( C_P \) and \( C_D \).

**Definition 5 (Estimation of \( C_P \)).** 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

\[
\hat{C}_P(A_n, B_n) = \sum_{i=1}^{n} \frac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2}.
\]

This estimator is well known in the literature because it is the U statistic of the Mann-Whitney test Mann and Whitney (1947).

**Definition 6 (Estimation of \( C_D \)).** 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 \( a_i \neq b_k \) for all \( i = 1, ..., n \). We define the estimation of the dominance rate as

\[
\hat{C}_D(A_n, B_n) = \sum_{j=1}^{2n} \frac{\mathbb{I}(\hat{G}_A(c_j) > \hat{G}_B(c_j) \land c_j \in A_n)}{2n} - \frac{\sum_{j=1}^{2n} \mathbb{I}(\hat{G}_A(c_j) < \hat{G}_B(c_j) \land c_j \in B_n)}{2n} + \frac{1}{2}.
\]

where \( \mathbb{I} \) 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 the online Appendices for details).

4. Cumulative Difference-Plot

In this section, we propose a graphical method called cumulative difference-plot that shows the estimations of \( C_P \) and \( C_D \) decomposed by quantiles: \( C_P \) and \( C_D \) 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.
4.1. 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 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 9 for an example. In the online Appendices, we show how to compute the probability densities of $Y_A$ and $Y_B$ step by step.

A more simple approach would be to estimate and define the quantile random variables through the empirical cumulative distribution functions of the observed samples of $X_A$ and $X_B$. However, the quantile random variables defined through uniform kernels have some interesting properties: they have the same $C_P$ and $C_D$ as the kernel density estimations of $X_A$ and $X_B$ (shown in the online Appendices). In addition, $g_{Y_A}(x) + g_{Y_B}(x) = 2$ for all $x \in [0, 1]$. As we will later see, these properties are essential for the interpretation of the cumulative difference-plot.

4.2. 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 Cheng and Iles (1983), Steck (1971), Cheng and Lies (1988), Wang, Cheng, and Yang (2013), Faraway and Myounghsic (1990), Bickel and Krieger (1989), and Hall and Horowitz (2013) on how to estimate the confidence bands of cumulative distributions, and, in this work, we use a simple bootstrap approach.

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 2.1). We show the 95% confidence bands of the cumulative distribution functions of $Y_A$ and $Y_B$ in Figure 10. The estimated cumulative distribution functions of $Y_A$ and $Y_B$ resemble the cumulative distribution functions of $X_A$ and $X_B$ from Figure 5(b). However, there are several relevant differences. In Figure 10, 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

---

**Figure 9.** An example of the probability density functions of $Y_A$ and $Y_B$ given the observed samples $A_n \cup B_n$.  

**Figure 10.** An example of the probability density functions of $Y_A$ and $Y_B$ given the observed samples $A_n \cup B_n$.  

---

8The bootstrapping Efron and Tibshirani (1993) 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 $\alpha$ 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 (Goeman and Solari 2014; Bauer 1991). This means that we have to use a higher confidence level for each band: $\sqrt{(1 - \alpha)}$.  

\(X_A\) 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.

### 4.3. 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 example shown in Figure 11, the estimation is \(C_D = \text{Length}_2 \approx 0.75\). 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.

#### 4.3.1. \(C_P\) and \(C_D\) in the Cumulative Difference-Plot

\(C_P\) and \(C_D\) can be directly obtained from the proposed plot. The integral of the difference between \(Y_A\) and \(Y_B\) is \(C_P = 0.5\) (we prove this in the online Appendices). Formally, \(C_P = 0.5 + \int_0^1 \text{diff}(x)\,dx\). However, in practice, \(C_P\) can be visually estimated by adding 0.5 to the difference in the areas over and under 0. For the example shown in Figure 11, \(C_P = 0.5 - \text{Area}_1 + \text{Area}_2\). 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 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, \(C_D\) 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,

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

where \(I\) is the indicator function (we prove this in the online Appendices). As an example, \(C_D\) is proportional to \(\text{Length}_2 - \text{Length}_1\) in Figure 11: it is higher than 0.5, because \(\text{Length}_2 \gg \text{Length}_1\). 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, \(C_D\) can also be estimated as the total length in which \(\text{diff}(x) > 0\). In the example in Figure 11, the estimation is \(C_D = \text{Length}_2 \approx 0.75\) but this is an easy case to identify, as the difference is constantly 0.

#### 4.3.2. Illustrative Example

Figure 12 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.

---

\(^9\)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 \(X_B\).
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 $C_D(X_A, X_B) > 0.75$, as deduced from $\text{diff}(x) > 0$ for all $x \in (0.25, 1)$. The difference is also near its maximum value, implying that its integral is high and thus $C_P$ (the probability that $X_A < X_B$) is also near 1.

5. 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.

5.0.1. 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 boxplots 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 6.1).

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 (Thas 2010; Scikit-Learn Developers 2021). A kernel density estimation with the uniform kernel—considered to be the moving window equivalent of the histogram—overcomes this limitation (Thas 2010), at the cost of using a more complex model. Similarly, the boxplot has a “noninjectivity” problem: very different data can still have the same boxplot (Matejka and Fitzmaurice 2017; Chatterjee and Firat 2007). The violinplot is an extension of the boxplot that overcomes the above limitation by combining the kernel density estimate of the random variables with the traditional boxplot (Hintze and Nelson 1998). 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.

5.0.2. 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 than the observed (Conover and Conover 1980) 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 (Greenland et al. 2016). Usually, this threshold is set at an arbitrary but well established (Wasserstein and Lazar 2016) $p = 0.05$, although recently, further reducing the threshold to $p = 0.005$ has been proposed (Benjamini et al. 2018; Ioannidis 2018).

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 nonparametric test (Conover and Conover 1980) is a suitable choice. Specifically, the Mann-Whitney (Mann and Whitney 1947) test is a good choice, as the samples observed from the random variables are iid for each random variable.\footnote{For paired data, the Wilcoxon signed-rank test (Wilcoxon 1945) or the sign test (Conover and Conover 1980) should be used. However, in the context of this article, the samples observed from the random variables are not paired. In this article, we consider the Mann–Whitney test as it is probably the most well-known nonparametric test for unpaired data, although take into account that more modern alternatives have been proposed (Ledwina and Wyłupek 2012; Baumgartner, WeiB, and Schindler 1998; Biswas and Ghosh 2014).}

With this test, the null hypothesis is that $P(X_A > X_B) = P(X_B > X_A)$, and a possible alternative hypothesis is that $X_A > X_B$ (Mann and Whitney 1947).

Null hypothesis tests have some limitations: for example, the $p$-value does not separate between the effect size and the sample size (Benavoli et al. 2017; Calvo et al. 2019). 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

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{cumulative_difference_plot.png}
\caption{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).}
\end{figure}
every time\textsuperscript{12}. However, $X_A \not\sim X_B$ and $X_B \not\sim X_A$, implying that the alternative hypothesis is not true. Note that the proposed cumulative difference-plot (shown in Figure 12) 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.

5.0.3. Bayesian Analysis

As an alternative (Benavoli et al. 2017; Calvo et al. 2019) to the limitations of null hypothesis test, Bayesian analysis has been proposed. Bayesian analysis (Gelman 2014; Bernardo and Smith 2009) 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 and data. Recently a Bayesian version of the Wilcoxon signed-rank test (Benavoli et al. 2014, 2017) has been proposed. In this article, we will consider the simplex-plot of its posterior distribution. For convenience, in the rest of the article, 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, where “rope” or range of practical equivalence denotes hypothesis $X_A = X_B$ (when the difference is in the interval $[-r, r]$).

We computed the simplex-plot (Figure 13) with the 400 samples of $X_A$ and $X_B$ from Case 2 obtained in Section 4.2. Two samples were considered equal when their difference is lower than $r = 10^{-3}$, and we used the prior proposed by Benavoli et al. (2017). 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 comparative 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.

5.0.4. 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 and Trede (1996), it can be interpreted via the integral of the nonnegative part, which represents the amount of violation against the hypothesis that $X_A$ dominates $X_B$.

The quantile-quantile plot (Thas 2010; Wilk and Gnandesikan 1968) 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 article 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 $C_D, C_P$ 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 $C_D, C_P$ 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.

6. Experimentation with the Cumulative Difference-Plot

To illustrate the applicability of the proposed methodology, in the following, we reevaluate the experimentation of a recently published work. In a recent article, Santucci, Ceberio, and Baioletti (2020) 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 (Larrañaga and Lozano 2001) (from now on PL-EDA). These two algorithms were tested in a set of 50 problem instances of the linear ordering problem (Schiavinotto and Stützle 2004). 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, Ceberio, and Baioletti (2020), 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, Ceberio, and Baioletti (2020) 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. We increase the sample size from \( n = 20 \) to \( n = 10^3 \) because the difference between the performance of the algorithms is small. With a sample size of \( n = 20 \), the uncertainty is too high to come to any meaningful conclusion (regardless of the statistical methodology considered). 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 \( 10^{-4} \).

### 6.1. Step 1: Visualization

Figure 14 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 15 shows the boxplot and the violinplot of the data. Both algorithms have a similar median, but due to the high number of outliers (Carreño, Inza, and Lozano 2020), it is difficult to compare the scores of the algorithms with the boxplot. The same happens with the violinplot.

### 6.2. 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, boxplot, and violinplot) 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.

6.2.1. Mann–Whitney Test
Applying the Mann-Whitney test we obtain a p-value of \( p = 0.035 \), lower than the usually used 0.05 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.

6.2.2. Simplex-Plot
We show the simplex-plot (Benavoli et al. 2017) of the scores in Figure 16. Following the criterion by Santucci, Ceberio, and Baiolletti (2020), we considered that two scores are equal when they differ by less than \( 10^{-4} \). 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. 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.

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).

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

7.1. 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 iid, 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 4, repeated samples were assigned the same rank. Then, the size of the uniform distributions was adjusted (with the \( y \) function) ensuring

\[
\text{diff}(x) = 0 \text{ is a little larger than the area over dif}(x) = 0, \text{ hence } P(x_{eda} < x_{gs}) \text{ is a little smaller than } P(x_{gs} < x_{eda}).
\]

1. Neither algorithm dominates the other one, and what is more, \( C_{PL-EDA, PL-GS} \approx 0.5 \).
2. 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.
3. 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.
4. 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.
5. 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.

Figure 16. Simplex-plot of PL-GS and PL-EDA in the instance \( N=70n=13x \). 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.
that the sum of the estimated densities of $Y_A$ and $Y_B$ remains constant even in the case of repeated observations.

### 7.2. 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 (Danica and Power 2009) implies that the estimation is closer to the real distribution, (ii) the bootstrap method also requires several samples to be meaningful (Chernick 2011; Hall 2013) 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 article. 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 (e.g., 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 (Benavoli et al. 2014; Calvo et al. 2019), or directly comparing boxplots. 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 iid, 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, R Core Team 2020). However, its speed can probably be further improved with a better implementation.

### 8. Conclusion

In this article, 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 6 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.

### Supplementary Material

**RVCompare:** With this R package, users can compute the $C_P$ and $C_D$ of two distributions, given their probability density functions.
more, it can be used to produce the proposed cumulative difference-plot, given the observed data. (The package can be directly installed from CRAN and is also available in the GitHub repo https://github.com/EtorArza/RVCompare).

Reproducibility: Alongside the article, we provide the code to generate the figures in this article and replicate the experimental setup. For instructions on how to install the dependencies and replicate the results, refer to the README.md file in the repository. (GitHub repo https://github.com/EtorArza/SupplementaryPaperRVCompare).

Appendices: To keep the length of the article at a reasonable size, the appendices have been moved to another document. (The appendices are available for download at https://doi.org/10.5281/zenodo.6528669).

Acknowledgments

This work was funded in part by the Spanish Ministry of Science, Innovation and Universities through PID2019-106453GA-I00/AEI/10.13039/5011 00011033 and the BCAM Severo Ochoa excellence accreditation SEV-2017-0718; by the Basque Government through the Research Groups 2019-2021 IT1244-19, ELKARTEK Program (project code KK-2020/00049) and BERC 2018-2021 program.

Disclosure Statement

The authors report that there are no competing interests to declare.

ORCID

Etor Arza © https://orcid.org/0000-0002-8044-0334

References

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. R., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., and De Souza, A. F. (2021), “Self-Driving Cars: A Survey,” Expert Systems with Applications, 165, 113816. [552]

Bauer, P. (1991), “Multiple Testing in Clinical Trials,” Statistics in Medicine, 10(6), 871–890. [558]

Baumgartner, W., WeiB, P., and Schindler, H. (1998), “A Nonparametric Test for the General Two-Sample Problem,” Biometrics, 54, 1129. [560]

Benavoli, A., Corani, G., Demsár, J., and Zaffalon, M. (2017), “Time for Statistics in Psychology,” J. Math. Psychol., 61, 248–254. [561,564]

Benavoli, A., Corani, G., Mangili, F., Zaffalon, M., and Ruggeri, F. (2014), “A Bayesian Wilcoxon Signed-Rank Test Based on the Dirichlet Process,” in International Conference on Machine Learning, pp. 1026–1034, PMLR. [561,564]

Benjamini, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenland, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho T., Hojttink, H., Hruschka, D. I., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchner, M., Lai, Song, Y., Lai, S. T., Laverty, J., Le, H. Y., Little, R., Machin, T., Machin, D., Madeleine, F., McPherson, C., McShane, B., Mehta, P. D., Merkle, C., Micali, S., Muller, T., Nakagawa, S., Nishii, H., Nosek, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Ruscio, J., Savalei, V., Schonbrodt, F. D., Sellke, T., Sinclair, C., Tingley, D., Vaz, T., Vazirani, S., Watts, D. J., Winship, C., Wolpert, R., Xie, Y., Young, C., Zuckerman, B., Zuckerman, E., Zuckerman, V. E. (2018), “Redefine Statistical Significance,” Nature Human Behaviour, 2, 6–10. [560]

Bennet, C. J. (2013), “Inference for Dominance Relations,” International Economic Review, 54, 1309–1328. [553,554]

Bernardo, J. M., and Smith, A. F. (2009), Bayesian Theory, Vol. 405. Wiley. [561]

Bickel, P. J., and Krieger, A. M. (1989), “Confidence Bands for a Distribution Function Using the Bootstrap,” Journal of the American Statistical Association, 84, 95–100. [558]

Biswas, M., and Ghosh, A. K. (2014), “A Nonparametric Two-Sample Test Applicable to High Dimensional Data,” Journal of Multivariate Analysis, 123, 160–171. [560]

Calvo, B., and Santafé Rodrigo, G. (2016), “scmp: Statistical Comparison of Multiple Algorithms in Multiple Problems,” The R Journal, 8, 2016. [561]

Calvo, B., Shir, O. M., Ceberio, J., Doerr, C., Wang, H., Bäck, T., and Lozano, J. A. (2019), “Bayesian Performance Analysis for Black-Box Optimization Benchmarking,” in Proceedings of the Genetic and Evolutionary Computation Conference Companion on – GECCO’19, Prague, Czech Republic, pp. 1789–1797, ACM Press. [560,561,564]

Carreño, A., Inza, I., and Lozano, J. A. (2020), “Analyzing Rare Event, Anomaly, Novelty and Outlier Detection Terms Under the Supervised Classification Framework,” Artificial Intelligence Review, 53, 3575–3594. [553,562]

Chatterjee, S., and Firtat, A. (2007), “Generating Data with Identical Statistics but Dissimilar Graphics: A Follow up to the Anscombe Dataset,” The American Statistician, 61, 248–254. [551,560]

Cheng, R. C., and Li, T. (1988), “One-Sided Confidence Bands for Cumulative Distribution Functions,” Technometrics, 30, 155–159. [558]

Cheng, R. C. H., and Iles, T. C. (1983), “Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables,” Technometrics, 25, 77–86. [558]

Chernick, M. R. (2011), Bootstrap Methods: A Guide for Practitioners and Researchers, Vol. 619. Wiley. [564]

Chollet, F. et al. (2015), “Keras.” [551]

Conover, W. J., and Conover, W. J. (1980), Practical Nonparametric Statistics, New York: Wiley. [552,553,556,560]

Cruz-Roa, A., Gilmore, H., Basavanahally, A., Feldman, M., Ganesan, S., Shih, N. N., Tomaszewski, J., González, F. A., and Madabhushi, A. (2017), “Accurate and Reproducible Invasive Breast Cancer Detection in Whole-Slide Images: A Deep Learning Approach for Quantifying Tumor Extent,” Scientific Reports, 7, 46450. [552]

Danica and Power (2009), “What is the Minimum Number of Data Points Required for Kernel Density Estimation?” [564]

Efron, B., and Tibshirani, R. (1993), An Introduction to the Bootstrap, Number 57 in Monographs on Statistics and Applied Probability: New York: Chapman & Hall. [558]

Faraway, J. J., and Myongshic, J. (1990), “Bootstrap Choice of Bandwidth for Density Estimation,” Journal of the American Statistical Association, 85, 1119–122. [558]

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). “An Introduction to Deep Reinforcement Learning,” Foundations and Trends in Machine Learning, 11, 219–354. [552]

Gelman, A. (2014), Bayesian Data Analysis (3rd ed.), Chapman & Hall/CRC.

Glorot, X., and Bengio, Y. (2010), “Understanding the Difficulty of Training Deep Feedforward Neural Networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, eds. Y. W. Teh and M. Titterington, Volume 9 of Proceedings of Machine Learning Research, Chia Laguna Resort, Sardinia, Italy, pp. 249–256. JMLR Workshop and Conference Proceedings. [561]

Goodfellow, I., Bengio, Y., and Courville, A. (2016), Deep Learning, Cambridge, MA: MIT Press. [551]

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), “Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations,” European Journal of Epidemiology, 31, 337–350. [560]

Hall, P. (2013), The Bootstrap and Edgeworth Expansion, Springer Science & Business Media. [564]

Hall, P., and Horowitz, J. (2013), “A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions,” The Annals of Statistics, 41, 1892–1921. [558]

Hinzte, J. L., and Nelson, R. D. (1998), “Violin Plots: A Box Plot-Density Trace Synergism,” The American Statistician, 52, 181–184. [560]
Ioannidis, J. P. (2018), “The Proposal to Lower P Value Thresholds to .005,” Journal of the American Medical Association (JAMA), 319, 1429–1430.

Larrañaga, P., and Lozano, J. A. (2001), Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation, Vol. 2. Springer Science & Business Media.

Ledwina, T., and Wyłupek, G. (2012), “Nonparametric Tests for Stochastic Ordering,” TEST, 21, 730–756.

Mann, H. B., and Whitney, D. R. (1947), “On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other,” The Annals of Mathematical Statistics, 18, 50–60.

Matejka, J., and Fitzmaurice, G. (2017), “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, pp. 1290–1294.

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

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

Quirk, J. P., and Saposnik, R. (1962), "Admissibility and Measurable Utility Functions," The Review of Economic Studies, 29, 140–146.

R Core Team (2020), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Regnier-Coudert, O., McCall, J., Ayodele, M., and Anderson, S. (2016), "Truck and Trailer Scheduling in a Real World, Dynamic and Heterogeneous Context," Transportation Research Part E: Logistics and Transportation Review, 93, 389–408.

Rousseeuw, P. J., and Hubert, M. (2011), "Robust Statistics for Outlier Detection," WIREs Data Mining and Knowledge Discovery, 1, 73–79.

Santucci, V., Ceberio, J., and Baioletti, M. (2020), "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, Cancun Mexico, July, pp. 1704–1711. ACM.

Schiavinotto, T., and Stützle, T. (2004), “The Linear Ordering Problem: Instances, Search Space Analysis and Algorithms,” Journal of Mathematical Modelling and Algorithms, 3, 367–402.

Schmid, F., and Trede, M. (1996), “Testing for First-Order Stochastic Domination: A New Distribution-Free Test,” The Statistician, 45, 371.

Scikit-Learn Developers (2021), "Density Estimation."