Appendix for 'Comparing two samples through stochastic dominance: a graphical approach'.

Etor Arza
BCAM - Basque Center for Applied Mathematics
and
Josu Ceberio
University of the Basque Country UPV/EHU
and
Ekhiñe Irurozki
Télécom Paris
and
Aritz Pérez
BCAM - Basque Center for Applied Mathematics

May 8, 2022

Appendices

1 A literature review of measures

1.1 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[1] function $f : (0, +\infty) \to \mathbb{R}$ with $f(1) = 0$, and two continuous random variables X_A and X_B, the f-divergence Liese and Vajda (2006); Rényi et al. (1961) is defined as

$$D_f(X_A, X_B) = \int_{\mathbb{R}} g_B(x) f\left(\frac{g_A(x)}{g_B(x)}\right) dx \tag{1}$$

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 Polyanskiy and Wu (2012) 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 Kullback and Leibler (1951) 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 Papadopoulos (2017) 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 Goodfellow et al. (2016). It will not satisfy Property (2), as it is not antisymmetric either. This also makes the

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

interpretation less intuitive. The Kullback–Leibler divergence is often used to measure the difference between two random variables Goodfellow et al. (2016), 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 Papadopoulos (2017).

In Figure 1, 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 1b), one can clearly see that $X_A \succ X_B$, $X_B \lesssim X_C$ and $X_B \succ X_D$. However, as shown in Table 1.1, $D_{KL}(X_B, X_A) = D_{KL}(X_B, X_C) = D_{KL}(X_B, X_D) = 15.4$ 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 $X_A \lesssim X_B$. 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.

Jensen-Shannon divergence: The Jensen-Shannon divergence Polyanskiy and Wu (2012) is very similar to the Kullback-Leibler divergence, and is another the particular case of the f-divergence for $f(x) = x \cdot \ln(\frac{2x}{x+1}) + \ln(\frac{2}{x+1})$. It is also known as the symmetrized version of the Kullback–Leibler divergence Polyanskiy and Wu (2012), because

$$D_{JS}(X_A, X_B) = D_{KL}\left(X_A, X_{\mathcal{M}}\right) + D_{KL}\left(X_B, X_{\mathcal{M}}\right)$$

where the probability density function of $X_{\mathcal{M}}$ is $g_{\mathcal{M}}(x) = 0.5(g_A(x) + g_B(x))$. 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 $g_{\mathcal{M}}$. The Jensen-Shannon divergence also fails to identify (see Table 1.1) the dominance relationships between X_B

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

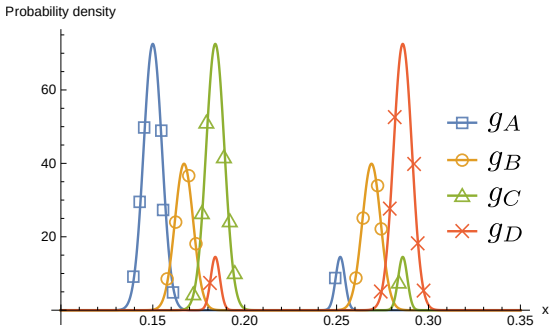
$\mathcal{C}_{\mathcal{P}}$

		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

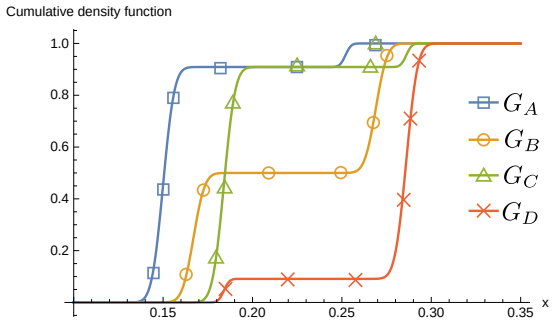
$\mathcal{C}_{\mathcal{D}}$

		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

Table 1: $\mathcal{C}(\mathrm{RV}_1, \mathrm{RV}_2)$ for the random variables X_A, X_B, X_C and X_D shown in Figure 1.



(a) Probability density.



(b) Cumulative distribution function.

Figure 1: The probability density function and cumulative distribution of the four random variables. The distances between these random variables are listed in Table 1.1.

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

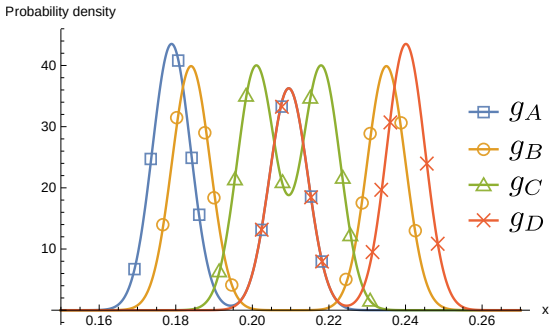
Total variation: The total variation Polyanskiy and Wu (2012) is also a particular f-divergence, for $f(x) = \frac{1}{2}|x-1|$. Unlike the Kullback–Leibler divergence, the total variation is symmetric. In fact, it is a properly defined distance Tsybakov (2009); Polyanskiy and Wu (2012). In addition, it is defined between 0 and 1.

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

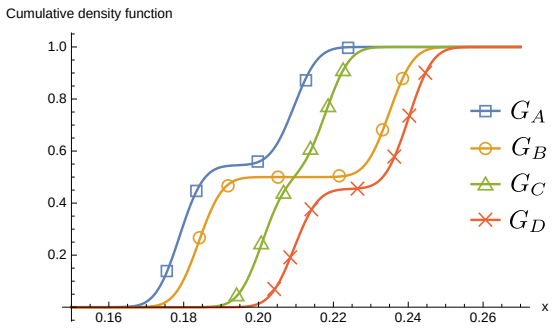
$$TV(X_A, X_B) = \sup_{C \subseteq \mathbb{R}} |\mathcal{P}_A(C) - \mathcal{P}_B(C)|,$$

where $\mathcal{P}_A$ and $\mathcal{P}_B$ are the probability distributions[2] of X_A and X_B respectively. Since

[2]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, $A(U) = \mathcal{P}(X_A \in U)$ Vapnik (1998).



(a) Probability density.



(b) Cumulative distribution function.

Figure 2: 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.

the subset C that takes the supremum is $C = \{x \in \mathbb{R} \mid g_A(x) > g_B(x)\}$Devroye et al. (2020), 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 Polyanskiy and Wu (2012), 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 1.1) the dominance relationships between X_B and the rest of the random variables in Figure 1.

Hellinger distance and the Bhattacharyya distance: The Hellinger distance is the square root of the f-divergence for $f(x) = (1 - \sqrt{x})^2$ Polyanskiy and Wu (2012). It is related to the Bhattacharya coefficient, since $D_H(X_A, X_B) = 2(1 - \text{BhattCoef}(X_A, X_B))$ Xi (2017); Polyanskiy and Wu (2012), where $\text{BhattCoef}(X_A, X_B)$ is the Bhattacharyya coefficient Kailath (1967); Bhattacharyya (1943). This coefficient is defined as $\text{BhattCoef}(X_A, X_B) = \int_{\mathbb{R}} \sqrt{g_A(x) g_B(x)} dx$, and has proven useful on signal processing Kailath (1967). 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 $D_{Bhatt}(X_A, X_B) = -\ln(\text{BhattCoef}(X_A, X_B))$.

The Hellinger distance and the Bhattacharyya distance also fail to identify (see Table 1.1) the dominance relationships between X_B and the rest of the random variables in Figure 1.

1.2 Wasserstein distance

The Wasserstein distance is another type of distance between probability random variables. Given two continuous random variables X_A, X_B, the Wasserstein distance (of order 1) is defined as Schuhmacher (2021); Panaretos and Zemel (2019)

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

In Figure 2, 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 2b), 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 2, the signed Wasserstein distance has different values: $D_{SW}(X_B, X_A) = 0.17$, $D_{SW}(X_B, X_C) = 0$ and $D_{SW}(X_B, X_D) = -0.017$. Notice that

$$X_A \succ X_B \implies D_{SW}(X_B, X_A) > 0 \text{ and } X_B \succ X_A \implies D_{SW}(X_B, X_A) < 0,$$

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 $X_A \lesseqgtr X_B$.

1.3 Heuristic derivation of the first-order stochastic dominance

A measure similar to the Wasserstein distance has been proposed in the literature Schmid and Trede (1996) 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 Schmid and Trede (1996). 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.

2 Quantile random variables

2.1 Computing the probability density functions of Y_A and Y_B

In Section 4.1 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.

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 $A_n \cup B_n$. The lowest value has rank 0. Assign the same rank to ties without skipping any rank. */
for i = *1,...,n* **do**
  $r(a_i) \leftarrow$ rank of a_i in $A_n \cup B_n$
  $r(b_i) \leftarrow$ rank of b_i in $A_n \cup B_n$
end
$R \leftarrow \{r(a_1), ..., r(a_n), r(b_1), ..., r(b_n)\}$
$r_{max} \leftarrow \max(R)$
for k = *-1,0,1,...,*r_{max} **do**
  $\gamma(k) \leftarrow$ the number of items in R lower than or equal to k
end
/* The probability density function of g_{Y_A} is represented as a mixture of n uniform distributions. $g_{Y_A}[s]$ is the probability density of Y_A in the interval $[\frac{s}{2n}, \frac{s+1}{2n})$. */
$g_{Y_A} \leftarrow$ array of zeros of length $2n$
$g_{Y_B} \leftarrow$ array of zeros of length $2n$
for $x_i = a_1, ..., a_n, b_1, ..., b_n$ **do**
  $A_{mult} \leftarrow$ number of times that x_i is in A_n
  $B_{mult} \leftarrow$ number of times that x_i is in B_n
  for *mult = 1,...,*$(A_{mult} + B_{mult})$ **do**
    $g_{Y_A}[\gamma(r(a_i) - 1) + mult - 1] \leftarrow (n \cdot A_{mult})^{-1}$
    $g_{Y_B}[\gamma(r(b_i) - 1) + mult - 1] \leftarrow (n \cdot B_{mult})^{-1}$
  end
end
return g_{Y_A}, g_{Y_B}

2.2 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 4.1, we claimed that when a "small enough" uniform scikit-learn developers (2021) 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 2.2, 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.

3 $\mathcal{C}_{\mathcal{P}}$ 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.

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

Definition 1. *(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 2. *(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

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

Definition 3. *(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 $\gamma(c_d) = \dfrac{\hat{G}_B(c_{d-1}) - \hat{G}_A(c_{d-1})}{[B_n = c_d] - [A_n = c_d]}$. *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 Steck (1971) 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 2. As mentioned in Section 4.3, 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

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

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

□

Proposition 1. *($\mathcal{C}_\mathcal{P}$ 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

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

With Lemma 1, it is enough to prove that

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

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

$$G_{Y_A}(\frac{j}{2n}) = \frac{[A_n < c_j] + \frac{[A_n = c_j][k \leq j | c_k = c_j]}{[C_{2n} = c_j]}}{n} \text{ and}$$
$$G_{Y_B}(\frac{j}{2n}) = \frac{[B_n < c_j] + \frac{[B_n = c_j][k \leq j | c_k = c_j]}{[C_{2n} = c_j]}}{n}$$

where $[A_n < c_j]$ counts the number of items in A_n lower than c_j, and $[k \leq j | c_k = c_j]$ 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

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

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

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

Now we group the terms in Equation (2) 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 $c_j = c_d$.

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

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

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

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

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

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

Focusing on the first sum, we have that

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

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

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

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

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

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,

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

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

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

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

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

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

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

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

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

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

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

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

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,

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

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

□

Proposition 2. *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 4.3, 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:

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

$$\sum_{j=1}^{2n}\int_{\frac{j-1}{2n}}^{\frac{j}{2n}} \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0]dx \tag{3}$$

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 (3) 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 $j_{d\downarrow} = j_{d\uparrow} = j$).

$$\sum_{d=1}^{d_{max}} \int_{\frac{j_{d\downarrow}-1}{2n}}^{\frac{j_{d\uparrow}}{2n}} \mathcal{I}[\text{diff}(x) > 0] - \mathcal{I}[\text{diff}(x) < 0] dx \tag{4}$$

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 (4) 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 (4) 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 3.

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 $\psi(c_d) = 0$.

□

3.2 Convergence of the estimators

Proposition 3. *Let X_A and X_B be two continuous 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 $\widetilde{\mathcal{C}_{\mathcal{P}}}(A_n, B_n) = \sum_{i,k=1\ldots n} \frac{\text{sign}(b_k - a_i)}{2n^2} + \frac{1}{2}$, which means that $\widetilde{\mathcal{C}_{\mathcal{P}}}(A_n, B_n)$ is also an estimator of $\mathcal{P}(X_A < X_B)$.

□

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 $|\operatorname{diff}(x)| < \delta$, the difference is displayed as 0.

In practice, we do not even need to account for the case that $\operatorname{diff}(x) = 0$. The cumulative difference-plot models the uncertainty with a confidence band, and when $\operatorname{diff}(x) = 0$ is inside the confidence band, then so are $\operatorname{diff}(x) > 0$ and $\operatorname{diff}(x) < 0$. If we assume that the difference is positive, negative or zero every time that $\operatorname{diff}(x) = 0$ 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 4. *(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 $\operatorname{diff}^{\delta}(x) = \operatorname{diff}(x)$ when $\operatorname{diff}(x) \geq \delta$, and $\operatorname{diff}^{\delta}(x) = 0$ 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

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

3.3 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 2 and 3 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 3 through 8. 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 9 and 10 respectively. In every case, the estimations with the three methods match, except for $\mathcal{C}_{\mathcal{D}}$ in Example 4 (Figure 6). 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 3 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.

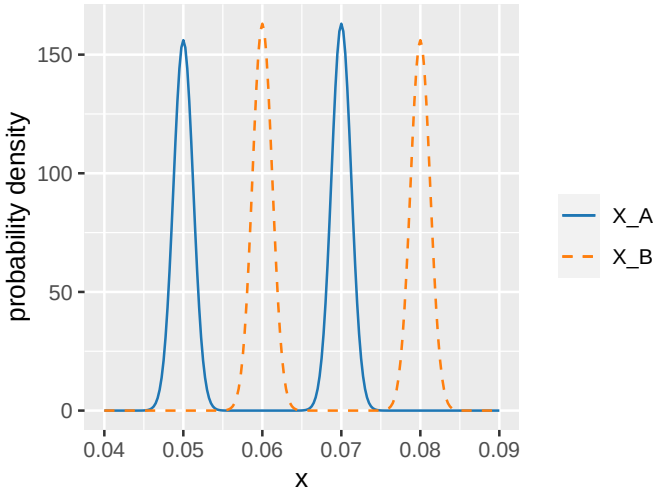


Figure 3: Probability density functions of Example 1

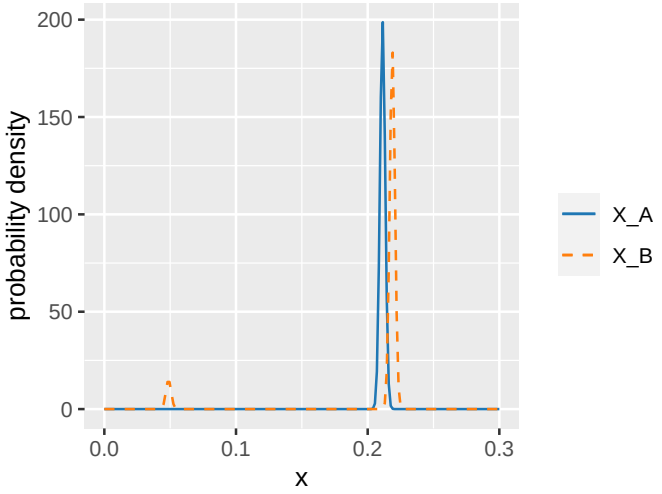


Figure 4: Probability density functions of Example 2

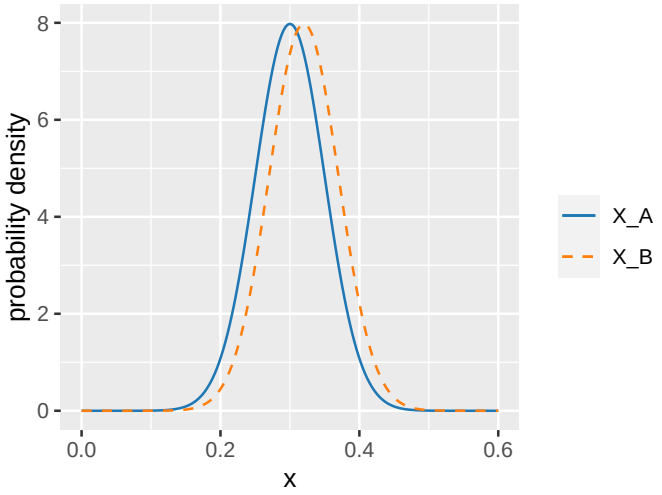


Figure 5: Probability density functions of Example 3

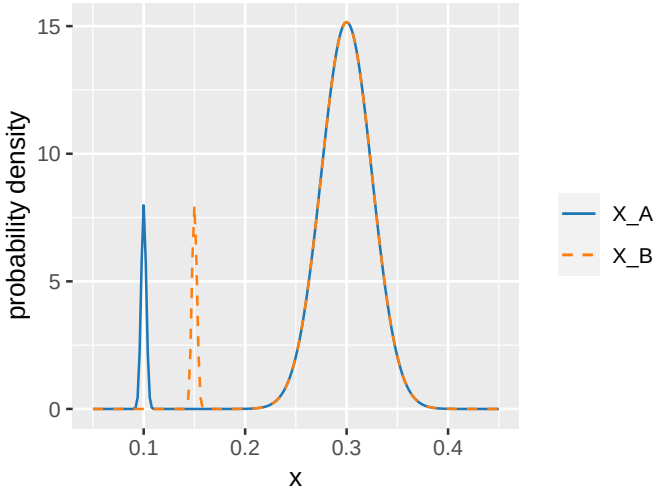


Figure 6: Probability density functions of Example 4

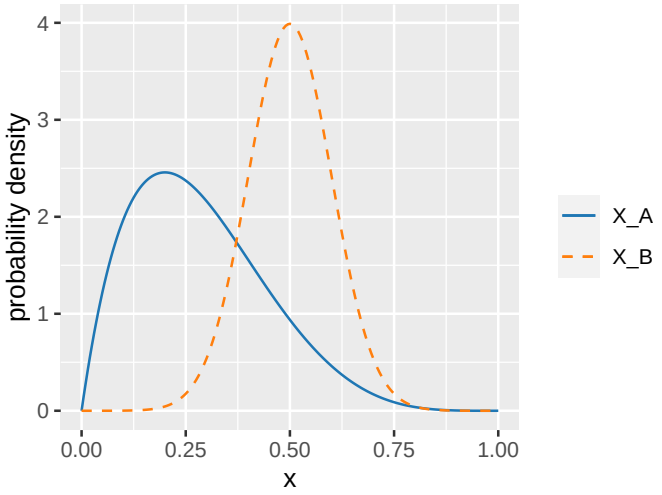


Figure 7: Probability density functions of Example 5

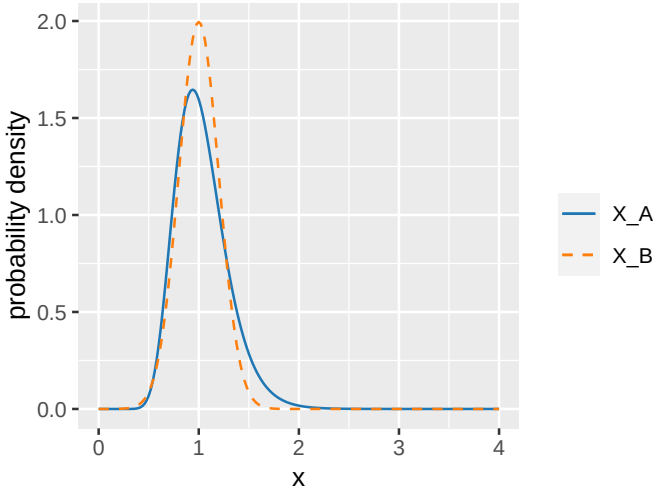


Figure 8: Probability density functions of Example 6

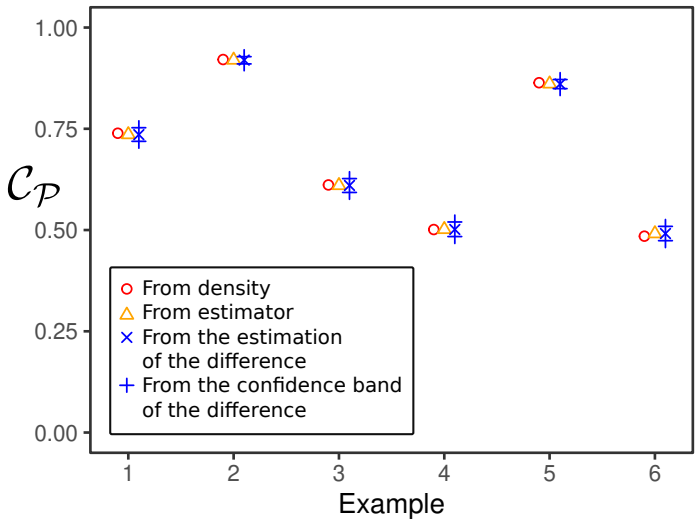


Figure 9: The $\mathcal{C}_{\mathcal{P}}$ values obtained in the six examples with the three methods.

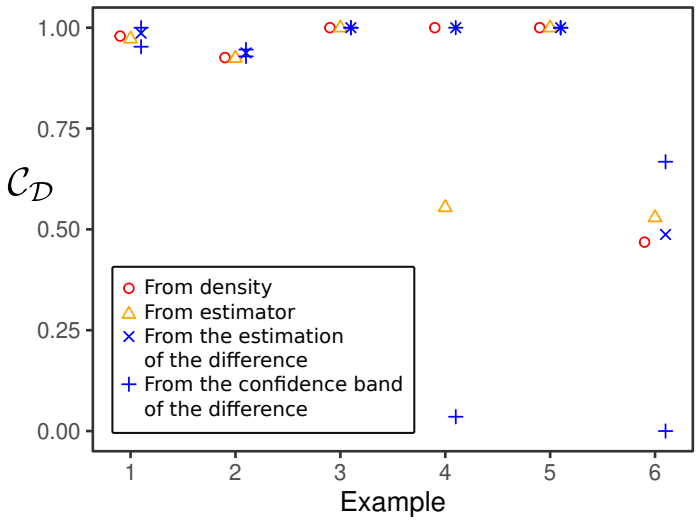


Figure 10: The $\mathcal{C}_{\mathcal{D}}$ values obtained in the six examples with the three methods.