Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–24
https://doi.org/10.1093/jrsssb/qkaf018

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Convexity and measures of statistical association

Emanuele Borgonovo[1], Alessio Figalli[2], Promit Ghosal[3], Elmar Plischke[4] and Giuseppe Savaré[1]

[1]Department of Decision Sciences and BIDSA, Bocconi University, Milan, Italy
[2]Department of Mathematics, ETH Zürich, Zurich, Switzerland
[3]Department of Statistics, University of Chicago, Chicago, USA
[4]Institute of Resource Ecology, HZDR, Dresden, Germany

Address for correspondence: Emanuele Borgonovo, Department of Decision Sciences and BIDSA, Bocconi University, Via Roentgen 1, Milan 20136, Italy. Email: emanuele.borgonovo@unibocconi.it

Abstract

Recent investigations on the measures of statistical association highlight essential properties such as zero-independence (the measure is zero if and only if the random variables are independent), monotonicity under information refinement, and max-functionality (the measure of association is maximal if and only if we are in the presence of a deterministic (noiseless) dependence). An open question concerns the reasons why measures of statistical associations satisfy one or more of those properties but not others. We show that convexity plays a central role in all properties. Convexity plus a form of strictness (that we are to define) are necessary and sufficient for zero-independence, and convexity and strict convexity on Dirac masses are necessary and sufficient for max-functionality. We apply the findings to study the families of measures of statistical association based on Csiszár divergences, optimal transport, kernels, as well as Chatterjee's new correlation coefficient. We further discuss the role of convexity in guaranteeing the asymptotic unbiasedness of given data estimators, prove a central limit theorem for those estimators under independence, and show the rate of convergence under arbitrary dependence. We demonstrate the findings with numerical simulations in a multivariate response context.

Keywords: correlation coefficient, convexity, data science, statistical association

1 Introduction

The need to determine the importance of features within large and complex datasets is renewing interest in the study of measures of statistical association. The works of Chatterjee (2021), Deb et al. (2020), and Wiesel (2022) introduce new indicators that possess properties such as zero-independence, max-functionality, and continuity (or monotonicity). These properties have been studied early on in the statistical literature (Hotelling, 1936), formulated as postulates in the seminal work of Rényi (1959) and revisited in many subsequent works (Mori & Szekely, 2019). (We introduce them qualitatively below and present them formally in the next section.)

(i) *Zero-independence* (Postulate D in Rényi, 1959, revised as Axiom (i) in Mori & Szekely, 2019) requires that a measure of statistical association between two non-constant random variables, say Y and X, is null if and only if they are independent.
(ii) *Max-functionality* (Postulate E in Rényi, 1959, extended by Axiom (iii) of Mori & Szekely, 2019) requires that a measure of statistical association is maximal if and only if Y is a deterministic function of the sole X, i.e. $Y = g(X)$ for some measurable function g.

Received: November 13, 2023. **Revised:** March 8, 2025. **Accepted:** March 17, 2025

(iii) *Continuity* requires that, given a sequence of random variables X_n, Y_n that converge in law to X, Y, respectively, the measure of statistical association of Y_n with X_n tends to the measure of statistical association of Y with X. This property is not stated as a postulate in Rényi (1959), but as Axiom (iv) in Mori and Szekely (2019).

A further important property concerns monotonic dependence on information, although it does not appear in the original postulates:

(iv) *Information monotonicity* requires that if X refines the information provided by Z (that is, if X generates a finer σ-algebra $\sigma(X) \supset \sigma(Z)$), then the measure of statistical association of Y with X is larger than the corresponding measure of the association of Y with Z.

This property can be seen as a version of the data-processing inequality, which states that if X refines the information of Z then the mutual information between Y and Z cannot be higher than the mutual information between Y and X (Mézard & Montanari, 2009, Ch. 1).

Furthermore, because the continuity condition (iii) is quite restrictive, it seems of great importance to study weaker conditions, such as lower semicontinuity under convergence in law, as follows:

(iii′) *Lower semicontinuity*: given a sequence of random variables X_n such that the joint law of Y, X_n converges weakly to the joint law of Y, X, the inferior limit of the measure of statistical association of Y with X_n is bounded below by the measure of statistical association of Y with X.

Popular measures of statistical association may possess only some of these properties. Pearson's correlation coefficient (Pearson, 1895) and correlation ratio (Pearson, 1905) do not satisfy zero-independence. The mutual information (Soofi, 1994), the maximal correlation coefficient of Hirschfeld (1935) and Rényi (1959), and the maximal information coefficient of Reshef et al. (2011) satisfy zero-independence and are maximal in the case of deterministic dependence. However, as Chatterjee (2021) underlines, they can attain their maximum also for a noisy dependence. The literature has also proposed measures of statistical association built on kernel functions (Barr & Rabitz, 2022; Da Veiga, 2021; Gretton et al., 2005), on copulae (Dette et al., 2013), on energy distances with distance-covariance (Székely & Rizzo, 2013; Székely et al., 2007), distance correlation (Székely & Rizzo, 2014) (see also Pan et al., 2020; Shen et al., 2020), as well as measures based on the Kolmogorov–Smirnov (Borgonovo et al., 2014), Hellinger (Geenens & de Micheaux, 2022), and the Wasserstein metrics (Wiesel, 2022). Whether they possess specific properties has, to date, been established on a case-by-case basis. This prompts the question of which geometric attributes render a measure of statistical association compatible with one or more of Rényi's postulates, and, more generally, with the various structural properties introduced thus far.

To address this question, we consider the class of measures defined as expectations of mappings between probability distributions. Following Glick (1975), we term these mappings 'separation measurements' and delve into how their properties relate to Rényi's postulates. We posit lower semicontinuity as a minimal property of the separation measurement.

We then investigate the role of convexity of the separation measurement. We show that convexity is, in fact, necessary and sufficient (equivalent) to information monotonicity (iv) (Theorem 2.6). In this convex framework, we show that zero-independence is equivalent to a new condition, that we call *strictness* (Theorem 2.9), which corresponds to strict mid-point convexity on particular class of measures. Concerning max-functionality, convexity ensures that a measure of statistical association is maximal in the case of a deterministic dependence (Theorem 2.10). Together with strict convexity on Dirac-δ masses, convexity yields necessary and sufficient condition for max-functionality (Theorem 2.12). This finding clarifies why certain measures of statistical association, such as mutual information, may attain their maximum value, yet this maximum does not indicate deterministic dependence. Furthermore, convexity yields continuity along information refinements and lower semicontinuity of the corresponding measure of statistical association (Theorems 2.13 and 2.14). We apply these findings to examine the properties of several existing measures of association, based on the families of Csiszár divergences, on the theory of optimal transport, on kernel separations, as well as on Chatterjee's new correlation coefficient.

Shifting our focus to estimation, we begin by examining estimators grounded in a nearest-neighbour approach, establishing their asymptotic unbiasedness (Theorem 4.1). We also supply insights into convergence rates and derive a central limit theorem (CLT), particularly under independence (Propositions 4.2 and 4.3). This derivation draws from the proof methodology presented in Deb et al. (2020). Subsequently, we formalize the one-sample binning estimation design pioneered by Pearson (1905). Pearson's design is widely employed in practical applications (Strong & Oakley, 2013; Strong et al., 2012). However, a rigorous theoretical framework for these estimators has been lacking. We demonstrate that Pearson's design yields estimators that approach asymptotic unbiasedness for expansive families of measures of statistical association (Theorem 4.5). Furthermore, we propose a CLT for Pearson's estimators under conditions of independence and provide a result detailing the rate of convergence under general distributions (Theorem 4.6). To exemplify these findings, we present numerical experiments in the context of a multivariate response case study.

2 Geometric properties and statistical association

This section is divided into two parts. In Section 2.1, we define the background, the notation, and the family of measures of association, and present the main theoretical results. In Section 2.2, we apply the findings to analyse several existing measures of statistical association.

2.1 Separation measurements and properties of measures of statistical association

Let X and Y be random variables on $(\Omega, \mathcal{A}, \mathbb{P})$, with values in Polish spaces X and Y and with laws $\mathbb{P}_X \in \mathscr{P}(\mathrm{X})$ and $\mathbb{P}_Y \in \mathscr{P}(\mathrm{Y})$, respectively, where $\mathscr{P}(\mathrm{X})$ (resp. $\mathscr{P}(\mathrm{Y})$) is the set of all Borel probability measures on X (resp. Y), endowed with the usual weak topology. In the remainder, to avoid triviality, we will assume Y non-constant. We denote by $\mathcal{B}(\mathrm{X})$ the Borel σ-algebra of X. If $\mathcal{F} \subset \mathcal{B}(\mathrm{X})$ is a countably generated σ-algebra, we recall that a system of regular conditional laws of Y generated by $(X, \mathcal{F})$ is a measure-valued map $x \mapsto \mathbb{P}^{\mathcal{F}}_{Y|X=x} \in \mathscr{P}(\mathrm{Y})$ satisfying the following two conditions: (a) for every $B \in \mathcal{B}(\mathrm{Y})$, the map $x \mapsto \mathbb{P}^{\mathcal{F}}_{Y|X=x}(B) \in \mathscr{P}(\mathrm{Y})$ is $\mathcal{F}$-measurable; and (b) for every $A \in \mathcal{F}$ and $B \in \mathcal{B}(\mathrm{Y})$, we have $\mathbb{P}[X \in A,\ Y \in B] = \int_A \mathbb{P}^{\mathcal{F}}_{Y|X=x}(B)\, \mathbb{P}_X(\mathrm{d}x)$. In the remainder, to simplify notation, we use $\mathbb{P}_{Y|X=x}$ instead of $\mathbb{P}^{\mathcal{F}}_{Y|X=x}$ when the σ-algebra is the reference Borel one, and we will also use the shorter $\mu = \mathbb{P}_X$ and $\nu = \mathbb{P}_Y$. Since Y is Polish, a conditional law exists and it is unique $\mathbb{P}_X$-a.e., see, e.g. Schwartz (1973, Section 2). When $\mathrm{X} = \mathbb{R}^{d_1}$, $\mathrm{Y} = \mathbb{R}^{d_2}$, for given measures $\mu = \mathbb{P}_X$, $\nu = \mathbb{P}_Y$, we denote by F_X, F_Y the corresponding cumulative distribution function and by f_X, f_Y the probability densities with respect to the Lebesgue measure, whenever $\mathbb{P}_X$ or $\mathbb{P}_Y$ is absolutely continuous. We start defining the notion of separation measurement.

Definition 2.1 A **separation measurement** is a map $\zeta : \mathscr{P}(\mathrm{Y}) \times \mathscr{P}(\mathrm{Y}) \to (-\infty, +\infty]$ such that for all $\nu \in \mathscr{P}(\mathrm{Y})$ the map $\nu' \mapsto \zeta(\nu, \nu')$ is lower semicontinuous and it is grounded, i.e.

$$\zeta(\nu, \nu) = 0 \quad \textit{for every } \nu \in \mathscr{P}(\mathrm{Y}). \tag{1}$$

Lower semicontinuity guarantees a minimal regularity also useful to define the integral in the definition of the corresponding measure of statistical association, see Definition 2.2. The grounded condition in (1) is not restrictive as well: if $\zeta(\nu, \nu) \neq 0$, we can always replace ζ with

$$\zeta_0(\nu, \nu') := \zeta(\nu, \nu') - \zeta(\nu, \nu), \tag{2}$$

obtaining a grounded separation measurement. We next introduce the family of measures of statistical association under investigation in this work. The family we are to consider encompasses many popular measures of statistical association. However, it does not exhaust the way in which measures of statistical association can be constructed. For instance, the construction in Nies et al. (2023) falls outside our framework. We offer additional remarks in Section 5.

Definition 2.2 Given a separation measurement $\zeta : \mathscr{P}(\mathrm{Y}) \times \mathscr{P}(\mathrm{Y}) \to (-\infty, +\infty]$, we define the **measure of statistical association** of Y with X, subject to $\mathcal{F}$ as

$$\xi^{\zeta}(Y, X \mid \mathcal{F}) := \mathbb{E}\Big[\zeta(\mathbb{P}_Y, \mathbb{P}^{\mathcal{F}}_{Y \mid X})\Big] = \int_{\mathrm{X}} \zeta(\mathbb{P}_Y, \mathbb{P}^{\mathcal{F}}_{Y \mid X=x})\, \mathbb{P}_X(\mathrm{d}x), \tag{3}$$

provided the negative part of $\zeta(\mathbb{P}_Y, \mathbb{P}^{\mathcal{F}}_{Y \mid X=x})$ is integrable.

If the context is clear, we drop the dependence on $\mathcal{F}$ in (3). If ζ does not satisfy (1), the shift (2) affects (3) by a constant:

$$\xi^{\zeta_0}(Y, X \mid \mathcal{F}) := \mathbb{E}\Big[\zeta_0(\mathbb{P}_Y, \mathbb{P}^{\mathcal{F}}_{Y \mid X})\Big] = \xi^{\zeta}(Y, X \mid \mathcal{F}) - \zeta(\mathbb{P}_Y, \mathbb{P}_Y), \tag{4}$$

so that, without loss of generality, we can incorporate the grounded condition (1) in Definition 2.1. Note that ξ^{ζ} depends only on the restriction of $\zeta(\nu, \cdot)$ on the probability measures of $\mathscr{P}(\mathrm{Y})$ with support contained in $\mathrm{supp}(\nu)$.

Our main structural assumption on ζ is convexity with respect to its second argument.

Definition 2.3 ((Convexity)). We say that a separation measurement ζ is **convex** (w.r.t. the second argument) if for every $\nu \in \mathscr{P}(\mathrm{Y})$ and every $\nu', \nu'' \in \mathscr{P}(\mathrm{Y})$ with support contained in $\mathrm{supp}(\nu)$

$$\zeta\left(\nu, \frac{1}{2}\nu' + \frac{1}{2}\nu''\right) \le \frac{1}{2}\zeta(\nu, \nu') + \frac{1}{2}\zeta(\nu, \nu''). \tag{5}$$

Equivalently, for every $\nu \in \mathscr{P}(\mathrm{Y})$, the map $\vartheta \mapsto \zeta(\nu, \vartheta)$ is convex on the set $\{\nu' \in \mathscr{P}(\mathrm{Y}) : \mathrm{supp}(\nu') \subset \mathrm{supp}(\nu)\}$.

Note that, because $\zeta(\nu, \cdot)$ is lower semicontinuous, its convexity ensures that the integral in (3) exists, possibly taking the value $+\infty$. We start showing that the above convexity property is intimately connected with information monotonicity. The simplest situation is when we receive information about some random variable Z which is a deterministic, possibly non-injective, function of X. Since we lose information passing from X to Z, we could expect that the strength of the statistical association of Y with Z is not greater than the one of Y with the non-distorted X.

Definition 2.4 ((Information monotonicity)). We say that ξ satisfies the information monotonicity property if for every pair of random variables X and Z (with values in Polish spaces X and Z, respectively) such that $\sigma(Z) \subset \sigma(X)$ (i.e. the information on X is finer than on Z) then $\xi(Y, X) \ge \xi(Y, Z)$.

Recall that $\sigma(X) \subset \mathcal{A}$ is the σ algebra generated by the sets $X^{-1}(B)$, $B \in \mathcal{B}(\mathrm{X})$. Since Z is Polish, by the Doob-Dynkin Lemma the condition $\sigma(X) \subset \sigma(Z)$ is equivalent to the existence of a Borel map $g : \mathrm{X} \to \mathrm{Z}$ such that $Z = g(X)$.

The second claim of the next result shows that information monotonicity is satisfied by ξ^{ζ} whenever ζ is convex, according to Definition 2.3.

Theorem 2.5 ((Information monotonicity)). Let ζ be a convex separation measurement.

1. If $\mathcal{F} \subset \mathcal{B}(\mathrm{X})$ is a given σ-algebra, then $\xi^{\zeta}(Y, X) \ge \xi^{\zeta}(Y, X \mid \mathcal{F})$.
2. If $(\mathrm{Z}, \mathcal{G})$ is a measurable space and $g : \mathrm{X} \to \mathrm{Z}$ is a $(\mu, \mathcal{F})$-measurable map with $Z := g \circ X$, then we have

$$\begin{aligned} &\xi^{\zeta}(Y, X \mid \mathcal{F}) \ge \xi^{\zeta}(Y, Z \mid \mathcal{G}) = \xi^{\zeta}(Y, X \mid \mathcal{F}') \\ &\quad \textit{where } \mathcal{F}' = g^{-1}(\mathcal{G}) \subset \mathcal{F}. \end{aligned} \tag{6}$$

3. If the σ-algebra generated by Z coincides with the σ-algebra generated by X(in particular if $Z = g(X)$ and g is μ-essentially injective), we have $\xi^{\zeta}(Y, Z) = \xi^{\zeta}(Y, X)$.

The next result shows that, under minimal regularity assumptions, the convexity property 2.3 of $\zeta(\nu, \cdot)$ is **equivalent** to the information monotonicity of ξ^ζ. For every $\nu \in \mathscr{P}(\mathrm{Y})$, we denote by $\mathscr{P}_\nu(\mathrm{Y})$ the set of probability measures absolutely continuous with respect to ν whose density is bounded:

$$\mathscr{P}_\nu(\mathrm{Y}) := \{\nu' \in \mathscr{P}(\mathrm{Y}) : \nu' \le C\nu \text{ for some } C \ge 1\}. \tag{7}$$

Theorem 2.6 ((Information monotonicity and convexity)). *If* $\xi^\zeta(Y, X)$ *satisfies information monotonicity*

$$\xi^\zeta(Y, X) \ge \xi^\zeta(Y, Z) \quad \textit{for every } Y, X \textit{ and } Z = g(X), \quad g : \mathrm{X} \to \mathrm{Z}, \tag{8}$$

then for every $\nu \in \mathscr{P}(\mathrm{Y})$, *the function* $\nu' \mapsto \zeta(\nu, \nu')$ *is convex on the set* $\mathscr{P}_\nu(\mathrm{Y})$.

If moreover $\zeta(\nu, \cdot)$ *satisfies the following regularity property:*

$$\begin{aligned}&\textit{for every } \nu' \in \mathscr{P}(\mathrm{Y}) \textit{ with } \mathrm{supp}(\nu') \subset \mathrm{supp}(\nu), \textit{ there exists a sequence}\\ &\nu'_n \in \mathscr{P}_\nu(\mathrm{Y}) \textit{ such that}\\ &\nu'_n \to \nu' \textit{ in } \mathscr{P}(\mathrm{Y}), \quad \zeta(\nu, \nu'_n) \to \zeta(\nu, \nu') \quad \textit{as } n \to \infty,\end{aligned} \tag{9}$$

then ζ is convex according to Definition 2.3.

Condition (9) is surely satisfied if $\zeta(\nu, \cdot)$ is continuous on the set of measures with support contained in $\mathrm{supp}(\nu)$.

We will now investigate the two extreme cases, when ξ^ζ attains its minimum or maximum values. Under convexity, the minimum is related to statistical independence.

Definition 2.7 ((Zero-independence)). We say that a measure of statistical association $\xi(Y, X)$ satisfies zero-independence if the following condition holds:

$$\xi(Y, X) = 0 \textit{ if and only if } Y \textit{ and } X \textit{ are statistically independent}. \tag{10}$$

Definition 2.8 ((Strictness)). We say that a convex separation measurement ζ is **strict** if the mid-point convexity inequality (5) is strict when $\nu = (1/2)\nu' + (1/2)\nu''$, $\nu' \ne \nu''$. Equivalently

$$\frac{1}{2}\nu' + \frac{1}{2}\nu'' = \nu, \quad \frac{1}{2}\zeta(\nu, \nu') + \frac{1}{2}\zeta(\nu, \nu'') = \zeta(\nu, \nu) \Rightarrow \nu' = \nu''. \tag{11}$$

Because we assumed ζ convex, it is immediate to verify that (11) and $\nu' \ne \nu''$, $\nu = (1/2)(\nu' + \nu'')$, yield $(1/2)\zeta(\nu, \nu') + (1/2)\zeta(\nu, \nu'') > \zeta(\nu, \nu) = 0$. The simplest situation where a *strict* separation measurement arises is when ζ is nonnegative, $\zeta(\nu, \nu') = 0$ occurs if an only if $\nu = \nu'$. In this case, if $\nu' \ne \nu''$, then $\nu = (1/2)\nu' + (1/2)\nu''$ is different from ν' and ν'' so that the left-hand side of (5) vanishes and the right-hand side is strictly positive. The next result shows the interplay between strictness as defined in (11), minimality and zero-independence.

Theorem 2.9 ((Minimality and independence)). Let ζ be a *convex* separation measurement and $Y \sim \nu \in \mathscr{P}(\mathrm{Y})$.

1. (Nonnegativity and minimality). The measure of statistical association $\xi^\zeta(Y, X)$ is nonnegative and it takes its minimal value $\zeta(\nu, \nu) = 0$ when X and Y are statistically independent.
2. (Zero-independence).
   ξ^ζ satisfies the zero-independence property if and only if ζ is also strict.

By Theorem 2.9, convexity guarantees that the minimum of $\xi^{\zeta}(Y, X)$ is attained under independence. This minimality property becomes a necessary condition when ζ is strict.

Let us now study the maximum value of $X \mapsto \xi^{\zeta}(Y, X)$, which should be linked to deterministic dependence. More precisely, we will be interested in determining whether Y and X are related by a Borel measurable mapping $g : \mathsf{X} \to \mathsf{Y}$, $Y = g(X)$. If this is the case, fixing X uniquely determines the value of Y and we shall say that the dependence of Y on X is deterministic. To clarify, consider the case when $g : \mathbb{R}^2 \to \mathbb{R}$, $Y = g(X_1, X_2)$ with X_1 and X_2 having joint law $\mathbb{P}_X$. Then, Y is a function of each X_i $(i = 1, 2)$. However, there is a deterministic dependence only between Y and the random vector $X = (X_1, X_2)$.

We start introducing the (candidate maximum) quantity

$$\mathbb{M}^{\zeta}[Y] := \int_{\mathsf{Y}} \zeta(\mathbb{P}_Y, \delta_y)\, \mathrm{d}\nu(y). \tag{12}$$

Note that $\mathbb{M}^{\zeta}[Y]$ is a weighted average of the separations between the law of Y and Dirac-δ masses centred at all realizations y of Y in Y. In some instances, we will also write $\mathbb{M}^{\zeta}(\nu)$ for $\mathbb{M}^{\zeta}[Y]$ when the emphasis is on the law ν of Y. By Equation (3), it is

$$\int_{\mathsf{Y}} \zeta(\mathbb{P}_Y, \delta_y)\, \mathrm{d}\nu(y) = \int_{\mathsf{Y}} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y\,|\,Y=y})\, \mathbb{P}_Y(\mathrm{d}y) = \mathbb{E}\big[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y\,|\,Y})\big] = \xi^{\zeta}(Y, Y). \tag{13}$$

That is, our candidate maximum $\mathbb{M}^{\zeta}[Y]$ equals the value of the measure of statistical association when we learn Y itself.

Theorem 2.10 ((Maximality and deterministic dependence)). Suppose that ζ is a convex separation measurement. For all random variables X, Y we have $\xi^{\zeta}(Y, X) \le \mathbb{M}^{\zeta}[Y]$. In particular, $\xi^{\zeta}(Y, X)$ is finite if $\mathbb{M}^{\zeta}[Y] < \infty$. In this case, if $Y = g(X)$ $\mathbb{P}$-a.e. for some Borel map $g : \mathsf{X} \to \mathsf{Y}$, then $\xi^{\zeta}(Y, X) = \mathbb{M}^{\zeta}[Y]$ so that the maximum value is attained. Finally, if ζ is also strict then $\mathbb{M}^{\zeta}[Y] > 0$, as we assumed Y non-constant.

Then, given a convex separation measurement ζ, for every random variable Y satisfying $0 < \mathbb{M}^{\zeta}[Y] < +\infty$, we can introduce the normalized index

$$\iota^{\zeta}(Y, X) := \frac{\xi^{\zeta}(Y, X)}{\mathbb{M}^{\zeta}[Y]}. \tag{14}$$

Note that $0 \le \iota^{\zeta}(Y, X) \le 1$, and it inherits the properties of $\xi^{\zeta}(Y, X)$.

The converse implication in Theorem 2.10 does not hold in general (see the examples in the next section). To obtain a necessary and sufficient condition, we need an additional property.

Definition 2.11 We say that a convex separation measurement ζ is **strictly convex on Dirac-$\boldsymbol{\delta}$ masses** if for every ν with $\mathbb{M}^{\zeta}(\nu) < \infty$, there exists a ν-negligible set $N \subset \mathsf{Y}$ such that

$$\zeta\left(\nu, \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}\right) < \frac{1}{2}\zeta(\nu, \delta_{y_1}) + \frac{1}{2}\zeta(\nu, \delta_{y_2}) \\ \textit{for every } y_1, y_2 \in \operatorname{supp}(\nu) \setminus N \textit{ with } y_1 \neq y_2. \tag{15}$$

Notice that ζ is surely strictly convex on Dirac δ-masses if for every $y_1 \neq y_2$ in $\operatorname{supp}(\nu)$ with $\zeta(\nu, \delta_{y_i}) < \infty$, we have

$$\zeta\left(\nu, \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}\right) < \frac{1}{2}\zeta(\nu, \delta_{y_1}) + \frac{1}{2}\zeta(\nu, \delta_{y_2}). \tag{16}$$

Theorem 2.12 ((Max-functionality)). Let us suppose that ζ is a convex separation measurement. $\xi^{\zeta}(Y, X)$ satisfies the max-functionality property, i.e.

$$\xi^{\zeta}(Y, X) = \mathbb{M}^{\zeta}[Y] < \infty \Leftrightarrow \textit{there exists a measurable function } g \\ \textit{such that } Y = g(X), \tag{17}$$

if and only if ζ is strictly convex on Dirac masses.

When the dependence between Y and X is deterministic, Theorem 2.12 establishes that having perfect information about X makes the measure of statistical association maximal. In this case, fixing X makes Y certain and it is desirable that $\xi^{\zeta}(Y, X)$ reaches its maximum value. The next result shows a continuity property with respect to an increasing family of σ-algebras in X.

Theorem 2.13 ((Information continuity)). Under the assumptions of Theorem 2.5, let $(\mathcal{F}^n)_{n\in\mathbb{N}}$ be an increasing family of sub-σ-algebras in $\mathcal{F}$ with $\mathcal{F} = \bigvee_{n=1}^{\infty} \mathcal{F}^n$. We have

$$\lim_{n\to\infty} \xi^{\zeta}(Y, X \mid \mathcal{F}^n) = \xi^{\zeta}(Y, X \mid \mathcal{F}). \tag{18}$$

The information continuity property in (18) is proven, independently, for kernel-based, Wasserstein- and optimal transport-based, and value-of-information based measures of statistical association, respectively, in Deb et al. (2020), Wiesel (2022), and Fissler and Pesenti (2023). Theorem 2.13 provides a unification of those results showing that continuity holds for measures of statistical association whose separation measurement is convex.

It is easy to check that continuity of ξ^{ζ} w.r.t. convergence in law of a sequence of random variables (Y_n, X_n) is a demanding condition for measures of association in (3), as such a convergence does not typically imply convergence of the conditional laws of Y_n w.r.t. X_n, even if ζ is weakly continuous. Convexity however is sufficient to recover at least lower semicontinuity.

Theorem 2.14 ((Lower semicontinuity)). Let ζ be a convex separation measurement.

1. If $X_n, X, n \in \mathbb{N}$ are random variables such that (Y, X_n) converges in law to (Y, X) as $n \to \infty$, then $\liminf_{n\to\infty} \xi^{\zeta}(Y, X_n) \geq \xi^{\zeta}(Y, X)$.
2. If ζ is also bounded from below, jointly convex, i.e. for $\nu_1, \nu_2, \nu_1', \nu_2' \in \mathscr{P}(\mathrm{Y})$, it is

$$\zeta\left(\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2, \frac{1}{2}\nu_1' + \frac{1}{2}\nu_2'\right) \leq \frac{1}{2}\zeta(\nu_1, \nu_1') + \frac{1}{2}\zeta(\nu_2, \nu_2'), \tag{19}$$

and lower semicontinuous in $\mathscr{P}(\mathrm{Y}) \times \mathscr{P}(\mathrm{Y})$ and the sequence (Y_n, X_n) converges in law to (X, Y) as $n \to \infty$, then $\liminf_{n\to\infty} \xi^{\zeta}(Y_n, X_n) \geq \xi^{\zeta}(Y, X)$.

As an immediate corollary of the previous result, we gain a sort of continuity in two extreme situations. First, suppose that ζ is a strict convex separation measurement. If (X_n) is a sequence such that (Y, X_n) converges in law to (Y, X) as $n \to \infty$ and satisfies $\lim_{n\to\infty} \xi^{\zeta}(Y, X_n) = 0$, then $\xi^{\zeta}(Y, X) = 0$ and X and Y are independent. (Note that $\xi^{\zeta}(Y, X) \geq 0$ by Theorem 2.9.) Clearly, if X_n and Y are independent for all n, then X and Y are independent. The statement in the first item, however, tells us that even if X_n and Y are not statistically independent for all n, but the limit of $\xi(Y, X_n)$ is zero, then X and Y are independent. Second, let ζ be a convex separation measurement. If $Y = g(X)$ and (X_n) is a sequence converging in law to X, then $\lim_{n\to\infty} \xi^{\zeta}(Y, X_n) = 1$.

Figure 1 synthesizes the relationships uncovered thus far. The left part refers to the structural properties of the separation measurement, and the right part to the corresponding properties of the measure of statistical association. Definition 2.1 asks for lower semicontinuity as a minimal requirement for the separation measurement ζ to be a sensible mathematical object. With

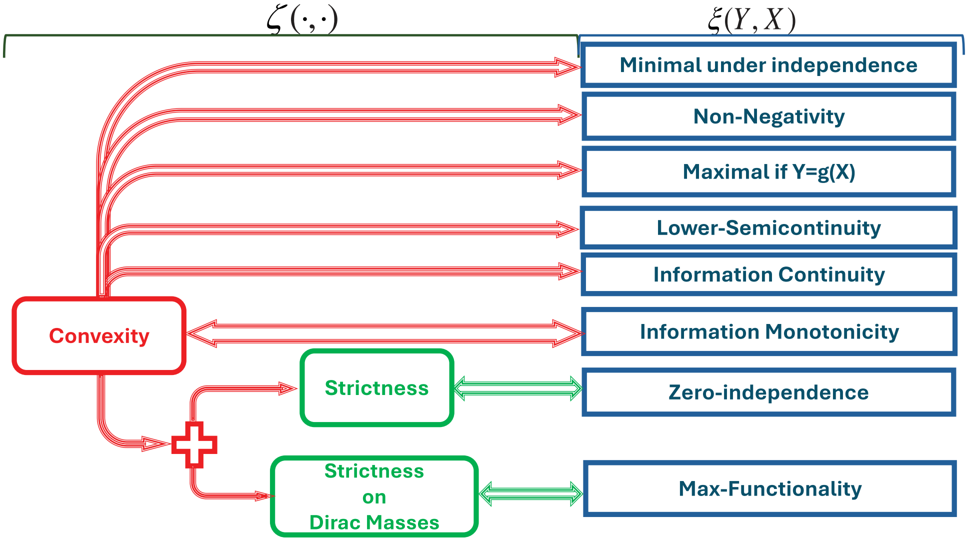

Figure 1. From properties of the separation measurement (left) to properties of the measure of statistical association (right).

convexity, it ensures that the integral in (3) exists. Convexity suffices to make the measure of statistical association nonnegative, minimal when Y and X are independent, maximal when Y is determined by X, and lower semicontinuous. Moreover, convexity is intrinsically linked to monotonicity (as visualized by the double implication arrow in Figure 1). Convexity and strictness are necessary and sufficient for zero-independence. Convexity and strictness on Dirac-δ masses are necessary and sufficient for max-functionality. Notably, two distinct forms of convexity characterize zero-independence and max-functionality.

2.2 Analysing the measures of statistical association

The results in Section 2.1 provide us with tools to investigate the properties of the measures of statistical association in the family of Definition (2.2). While some properties for some indicators have been reported in the previous literature, our results allow for a unified view. Table 1 lists the measures of statistical association analysed here and three main properties.

2.2.1 Pearson's correlation ratio and correlation coefficient

Let $\mathrm{Y} = \mathbb{R}$, $Y \sim \mathbb{P}_Y, Y' \sim \mathbb{P}_{Y'}$ and let $\mathbb{E}[Y]$, $\mathbb{E}[Y']$ denote their respective mean values and let $\mathbb{V}[Y]$ denote the variance of Y. Then, the quantity

$$\zeta^{\text{Pearson}}(\mathbb{P}_Y, \mathbb{P}_{Y'}) = (\mathbb{E}[Y] - \mathbb{E}[Y'])^2 \tag{20}$$

is a separation measurement, where we assume that $\zeta^{\text{Pearson}} = +\infty$ if Y or Y' are not integrable. As corresponding measure of statistical association, inserting (20) in (3), we obtain (the numerator of) Pearson's correlation ratio (Pearson, 1905)

$$\xi^{\text{Pearson}}(Y, X) = \mathbb{E}[\mathbb{V}\{Y\} - \mathbb{V}\{Y \mid X\}] = \mathbb{V}[\mathbb{E}[Y \mid X]]. \tag{21}$$

Proposition 2.15 $\xi^{\text{Pearson}}(Y, X)$ is null when Y and X are independent, it is nonnegative, it satisfies monotonicity, lower semicontinuity, and max-functionality, with maximum value equal to $\mathbb{V}[Y]$.

Table 1. Measures of association studied in this work and three main properties

Association measure	Zero-independence	Max-functionality	Monotonicity
Correlation ratio		✓	✓
Csiszár divergences	✓		✓
Chatterjee correlation	✓	✓	✓
Optimal transport-based	✓	✓	✓
Kernel-based	[✓]	[✓]	[✓]

Note. For kernel-based indices, the results depend on the properties of the kernel (Theorem 2.21).

By max-functionality, we can apply the normalization $\eta^2(Y, X) = \mathbb{V}[Y]^{-1}\xi^{\text{Pearson}}(Y, X)$ as originally proposed by Pearson (1905). With this normalization, in simulation experiments, $\eta^2(Y, X)$ coincides with first-order variance-based sensitivity indices (Saltelli & Tarantola, 2002). However, $\xi^{\text{Pearson}}(Y, X)$ does not satisfy zero-independence. In fact, the separation measurement in (20) is strictly convex on Dirac-δ masses, but it is not strict. The strictness condition in (11) fails, as $2(\mathbb{E}[Y] - (1/2)(\mathbb{E}[Y_1] + \mathbb{E}[Y_2]))^2 = (\mathbb{E}[Y] - \mathbb{E}[Y_1])^2 + (\mathbb{E}[Y] - \mathbb{E}[Y_1])^2$ only implies $\mathbb{E}[Y_1] = \mathbb{E}[Y_2]$ and not $Y_1 \sim Y_2$.

Pearson's correlation ratio was introduced as a generalization of the linear correlation coefficient, $\rho(Y, X)$ (Pearson, 1895). It is well known that $\rho(Y, X)$ neither satisfies zero-independence nor max-functionality. In the framework of Definition 2.2, we have

$$\rho(Y, X) = \frac{\mathbb{E}[X\mathbb{E}[Y - \mathbb{E}[Y] \mid X]]}{\sqrt{\mathbb{V}[Y]\mathbb{V}[X]}}, \tag{22}$$

which is the expectation of the quantity $\zeta^{\rho}(Y, X) = (1/\sqrt{\mathbb{V}[Y]\mathbb{V}[X]})X\mathbb{E}[Y - \mathbb{E}[Y] \mid X]$ for which convexity cannot be guaranteed (and does not actually satisfy the definition of separation measurement).

2.2.2 *Csiszár divergences*

Consider a convex function $\mathrm{J} : [0, +\infty) \to [0, +\infty]$, finite in $(0, +\infty)$, with $\mathrm{J}(1) = 0 < \mathrm{J}(r)$ for every $r \neq 1$. Given two probability measures ν and ν' with densities f, f' with respect a common dominating measure ϑ, we define the Csiszár J-divergence of ν with respect to ν' as $\zeta^{\text{Csiszar}}(\nu, \nu') := \int \mathrm{J}_{\text{hom}}(f, f')\,\mathrm{d}\vartheta$ with

$$\mathrm{J}_{\text{hom}}(f, f') := \begin{cases} \mathrm{J}\left(\dfrac{f}{f'}\right)f' & \text{if } f, f' \neq 0, \qquad \mathrm{J}_0 = \lim\limits_{r \downarrow 0} \mathrm{J}(r) \quad \text{if } f = 0 \neq f', \\ \mathrm{J}'_\infty = \lim\limits_{r \uparrow +\infty} \mathrm{J}(r)/r & \text{if } f \neq 0 = f', \qquad\qquad\qquad 0 \quad \text{if } f = f' = 0. \end{cases} \tag{23}$$

When J has a superlinear growth, as in the case of the Kullback–Leibler divergence, then $\mathrm{J}'_\infty = +\infty$, so that $\zeta^{\text{Csiszar}}(\nu, \nu')$ is finite if and only if $\nu \ll \nu'$ is absolutely continuous with respect to ν'.

The family of measures of statistical association corresponding to Csiszár divergences is

$$\xi^{\text{Csiszar}}(Y, X) = \mathbb{E}\big[\zeta^{\text{Csiszar}}(\mathbb{P}_Y, \mathbb{P}_{Y \mid X})\big] = \iint \mathrm{J}_{\text{hom}}\left(\frac{d\mathbb{P}_{Y \mid X}}{d\mathbb{P}_Y}\right)\mathrm{d}(\mathbb{P}_Y \otimes \mathbb{P}_X). \tag{24}$$

Proposition 2.16 The measure of association $\xi^{\text{Csiszar}}(Y, X)$ is nonnegative, maximal if $Y = g(X)$, and satisfies zero-independence, data-processing monotonicity, and lower semicontinuity.

This result follows from the convexity of ζ^{Csiszar} and the fact that it is strict. However, $\zeta^{\mathrm{Csiszar}}(\nu, \nu')$ in (23) does not satisfy strict convexity on Dirac-δ masses (see Appendix Remark A.2, online supplementary material). The Kullback–Leibler, the total variation, and the (squared) Hellinger separation measurements are three notable representatives of the family of Csiszár divergences.

If f, f' are the densities of ν, ν' with respect to a common dominating measure ϑ, they are, respectively, given by

$$\zeta^{\mathrm{KL}}(\nu, \nu') = \int_{\mathrm{Y}} \log\left(\frac{f(y)}{f'(y)}\right) f(y) d\vartheta(y) \quad \text{with } \mathrm{J}_{\mathrm{KL}}(r) = r\log r - r + 1, \tag{25}$$

$$\zeta^{\mathrm{TV}}(\nu, \nu') = \frac{1}{2}\int_{\mathrm{Y}} |f(y) - f'(y)|\, d\vartheta(y) \quad \text{with } \mathrm{J}_{\mathrm{TV}}(r) = |r - 1|, \tag{26}$$

and

$$\zeta^{\mathrm{H}^2}(\nu, \nu') = 1 - \int \sqrt{f(y) \cdot f'(y)}\, d\vartheta(y) \quad \text{with } \mathrm{J}_{\mathrm{H}^2}(r) = \frac{1}{2}\left(\sqrt{r} - 1\right)^2. \tag{27}$$

For the total variation or Hellinger divergences $\mathrm{J}'_\infty < +\infty$, so that $\zeta^{\mathrm{Csiszar}}(\nu, \nu')$ is finite even if ν is not absolutely continuous with respect to ν'. The corresponding measures of statistical association are the mutual information between Y and X (Soofi, 1994), the δ-importance measure (Borgonovo et al., 2014), and the Hellinger correlation (Geenens & de Micheaux, 2022). By Proposition 2.16, they all satisfy zero-independence and possess the properties that follow directly by convexity (Figure 1). However, because they are not strictly convex on Dirac-δ masses, they cannot satisfy max-functionality. This offers an explanation as to why, for instance, the mutual information satisfies zero-independence but not max-functionality, as observed by Chatterjee (2021).

Remark 2.17 Let ζ denote the total variation (ζ^{TV}) or the Hellinger (ζ^{H^2}) separation measurements in this remark. These separation measurements yield symmetric measure of statistical associations, that is $\xi^{\zeta}(Y, X) = \xi^{\zeta}(X, Y)$. Also, for every X and Y absolutely continuous, we have $\mathbb{M}^{\zeta}[Y] = \mathbb{M}^{\zeta}[X] = 1$. Because ζ is convex, $\xi^{\zeta}(Y, X)$ is maximal if $Y = g(X)$ (in this case, fixing X determines Y), so that $\xi^{\zeta}(Y, X) = 1$. By symmetry, we also have $\xi^{\zeta}(X, Y) = 1$. However, if g is not injective, fixing Y does not determine X. Thus, the measure $\xi^{\zeta}(X, Y)$ is maximal, but the dependence is not deterministic.

2.2.3 Chatterjee's new correlation coefficient

Consider as separation measurement the Cramér-von Mises distance for probability measures in $\mathrm{Y} = \mathrm{R}$, defined by

$$\mathrm{CvM}(\nu, \nu') = \int \left(F_Y(y) - F_{Y'}(y)\right)^2 \mathrm{d}F_Y(y), \qquad Y \sim \nu, \;\; Y' \sim \nu'. \tag{28}$$

Gamboa et al. (2018) discuss the probabilistic sensitivity measure based on (28) defining

$$\xi^{\mathrm{CvM}}(Y, X) = \mathbb{E}\left[\int \left(F_Y(y) - F_{Y|X}(y)\right)^2 \mathrm{d}F_Y(y)\right]. \tag{29}$$

The measure of association ξ^{CvM} in (29) coincides with the limit of Chatterjee's correlation coefficient (Chatterjee, 2021) (apart for a normalization factor).

Proposition 2.18 The measure of association ξ^{CvM} satisfies zero-independence, monotonicity, and max-functionality, with maximum $\mathbb{M}^{\mathrm{CvM}}[Y] = 1/6$.

2.2.4 Optimal transport-based separations

We next discuss three measures of statistical association shaped by the theory of optimal transport. Given a function $k : \mathrm{Y} \times \mathrm{Y} \to [0, +\infty)$ which is continuous and satisfies $k(y, y') = 0 \quad \Leftrightarrow \quad y = y'$, we consider:

(1) the Kantorovich optimal transport separation

$$\mathrm{OT}(\nu, \nu') := \inf_{\pi \in \Pi(\nu, \nu')} \int k(y, y')\, \mathrm{d}\pi(y, y') = \inf\{\mathbb{E}[k(Y, Y')] : Y \sim \nu.\ Y' \sim \nu'\}, \tag{30}$$

where $\Pi(\nu, \nu')$ denotes the set of all couplings in $\mathscr{P}(\mathrm{Y} \times \mathrm{Y})$ with marginals ν and ν', respectively;

(2) the entropic optimal transport separation

$$\mathrm{EK}_\varepsilon(\nu, \nu') := \inf_{\pi \in \mathscr{P}(\mathrm{Y} \times \mathrm{Y})} \int_{\mathrm{Y} \times \mathrm{Y}} k(y, y')\, \mathrm{d}\pi(y, y') + \varepsilon \mathrm{KL}(\pi, \nu \otimes \nu'), \tag{31}$$

(3) the family of logarithmic entropy-transport separations (Liero et al., 2018)

$$\mathrm{LET}(\nu, \nu') := \inf_{\pi \in \mathscr{M}_+(\mathrm{Y} \times \mathrm{Y})} \int_{\mathrm{Y} \times \mathrm{Y}} k(y, y')\, \mathrm{d}\pi(y, y') + \mathrm{KL}(\pi^1, \nu) + \mathrm{KL}(\pi^2, \nu'), \tag{32}$$

where the infimum is extended to all the finite nonnegative measures π in $\mathrm{Y} \times \mathrm{Y}$ with marginals π^1 and π^2, respectively.

With the classical optimal transport separation in (30), we obtain the family of optimal transport-based global sensitivity measures $\xi^{\mathrm{OT}}(Y, X) = \mathbb{E}[\mathrm{OT}(\mathbb{P}_Y, \mathbb{P}_{Y|X})]$ studied in Borgonovo et al. (2024). When $k(y, y') = |y - y'|^p$, we recover the pth power of the p-Wasserstein distance, $p \geq 1$, as a separation measurement $\mathrm{W}_p^p(\nu, \nu') = \inf_{\pi \in \Pi(\nu, \nu')} \int |y - y'|^p\, \mathrm{d}\pi(y, y')$, and the corresponding functional

$$\xi^{\mathrm{W}_p^p}(Y, X) = \mathbb{E}\big[\mathrm{W}_p^p(\mathbb{P}_Y, \mathbb{P}_{Y|X})\big], \tag{33}$$

which is the numerator of the Wasserstein correlation coefficient of Wiesel (2022).

The entropic separation in (31) is an important regularization of (30) (Peyré & Cuturi, 2019), whose corresponding measure of statistical association $\xi^{\mathrm{EK}_\varepsilon}(Y, X)$ has been introduced and studied in Borgonovo et al. (2024). The logarithmic entropy separation in (32) is another regularization of (30) which arises from unbalanced optimal transport. When $k(y, y') = |y - y'|^2$, (31) yields the Gaussian Hellinger–Kantorovich metric which metrizes weak convergence. We denote the corresponding measure of statistical association by $\xi^{\mathrm{LET}}(Y, X)$.

Proposition 2.19 The measures of statistical association $\xi^{\mathrm{OT}}(Y, X)$, $\xi^{\mathrm{W}_p^p}(Y, X)$, $\xi^{\mathrm{EK}_\varepsilon}(Y, X)$, and $\xi^{\mathrm{LET}}(Y, X)$ satisfy zero-independence, monotonicity, and max-functionality.

2.2.5 Kernel-based separations

We call kernel a function $K(\cdot\,, \,\cdot)$ on $\mathrm{Y} \times \mathrm{Y}$ which is nonnegative definite and symmetric. We denote by $\mathcal{H}_K$ the reproducing kernel Hilbert space induced by K, endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and the norm $\| \cdot \|_{\mathcal{H}_K}$. In addition, we require that Y is separable and that $K(\cdot\,, \,\cdot)$ is bounded and continuous, so that $\mathcal{H}_K$ is also separable (Steinwart & Christmann, 2008, Lemma 4.33). In the remainder, for every $\nu \in \mathscr{P}(\mathrm{Y})$, we denote by $m_K(\nu)$ the unique element of $\mathcal{H}_K$, defined as

$$\langle f, m_K(\nu) \rangle_{\mathcal{H}_K} = \int f(y)\, \mathrm{d}\nu(y), \tag{34}$$

where the integral w.r.t. ν is well defined as every function f of $\mathcal{H}_K$ is continuous and bounded (Steinwart & Christmann, 2008, Lemma 4.28). For every pair of distributions $\nu, \nu' \in \mathcal{P}(\mathrm{Y})$, we call the quantity

$$\mathrm{MMD}_K^2(\nu, \nu') = \| m_K(\nu) - m_K(\nu') \|_{\mathcal{H}_K}^2, \tag{35}$$

squared maximum mean discrepancy (MMD) between ν and ν'. The function MMD_K^2 induced by a bounded and continuous kernel K is a separation measurement. Following Deb et al. (2020) and Barr and Rabitz (2022), we can define the corresponding measures of statistical association as follows.

Definition 2.20 We call the quantity:

$$\xi^{\mathrm{MMD}}(Y, X) := \mathbb{E}[\mathrm{MMD}_K^2(\mathbb{P}_Y, \mathbb{P}_{Y|X})] \tag{36}$$

kernel-based measure of statistical association of Y with X.

The properties of $\xi^{\mathrm{MMD}}(Y, X)$ are determined by the associated kernel. We recall that a kernel K is *characteristic* if $m_{\nu_1} = m_{\nu_2}$ yields $\nu_1 = \nu_2$ for every $\nu_1, \nu_2 \in \mathscr{P}(\mathrm{Y})$. We also say that K separates the points if

$$K(y, y_1) = K(y, y_2) \quad \text{for every } y \in \mathrm{Y} \Rightarrow y_1 = y_2. \tag{37}$$

Theorem 2.21 Let K be a bounded continuous kernel on Y.

1. MMD_K^2 is a separation measurement which is jointly convex and lower semicontinuous in $\mathscr{P}(\mathrm{Y}) \times \mathscr{P}(\mathrm{Y})$.
2. The corresponding quantity $\mathbb{M}^{\mathrm{MMD}}[Y]$ defined by (12) reads as

$$\begin{aligned}\mathbb{M}^{\mathrm{MMD}}[Y] &= \int_{\mathrm{Y}} K(y, y)\,\mathrm{d}\nu(y) - \iint_{\mathrm{Y}\times\mathrm{Y}} K(y, y')\,\mathrm{d}(\nu \otimes \nu)\\ &= \mathbb{E}[K(Y, Y)] - \mathbb{E}[K(Y, Y')],\end{aligned} \tag{38}$$

where Y' is an independent copy of Y, whereas ξ^{MMD} can also be expressed by

$$\begin{aligned}\xi^{\mathrm{MMD}}(Y, \mathrm{X}) = &\int_{\mathrm{X}} \Big(\iint_{\mathrm{Y}\times\mathrm{Y}} K(y, y')\mathrm{d}(\nu_x \otimes \nu_x)\Big)\,\mathrm{d}\mu(x)\\ &- \iint_{\mathrm{Y}\times\mathrm{Y}} K(y, y')\mathrm{d}(\nu \otimes \nu).\end{aligned} \tag{39}$$

3. If K separates points, then MMD_K^2 is strictly convex on Dirac masses.
4. If K is characteristic, then MMD_K^2 is a strict separation measurement. Thus, ξ^{MMD} satisfies zero-independence, max-functionality, and information monotonicity.

Sejdinovic et al. (2013) show the equivalence between kernel-based measures of statistical associations and measures based on the energy distance (Székely & Rizzo, 2013). Let (Z, ρ) be a negative type metric space.[1] Suppose that $Z_1 \sim \mu_1 \in \mathscr{P}(\mathrm{Z})$ and $Z_2 \sim \mu_2 \in \mathscr{P}(\mathrm{Z})$ have finite first-order moment w.r.t. ρ, i.e. $\mathbb{E}[\rho(Z_1, \bar{z})] < \infty$ and $\mathbb{E}[Z_2, \bar{z}] < \infty$ for some $\bar{z} \in \mathrm{Z}$. The separation measurement ζ_ρ^{ED} associated with the ρ-energy distance between μ_1, μ_2 is defined by

$$\zeta_\rho^{\mathrm{ED}}(\mu_1, \mu_2) = 2\mathbb{E}\big[\rho(Z_1, Z_2)\big] - \mathbb{E}\big[\rho(Z_1, Z_1')\big] - \mathbb{E}\big[\rho(Z_2, Z_2')\big], \tag{40}$$

where Z_1' and Z_2' are i.i.d. copies of Z_1 and Z_2, respectively. For instance, when $\mathrm{Z} = \mathbb{R}$ and $\rho(x, y) = |x - y|$ for $x, y \in \mathbb{R}$, $\zeta^{\mathrm{ED}}(\nu, \mu)$ is a version of the Cramér-von Mises distance (see Sejdinovic et al., 2013, Székely & Rizzo, 2017, p. 455). Sejdinovic et al. (2013,

[1] A space (Z, ρ) is a negative type metric space if the metric ρ on Z satisfies $\sum_{i,j} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$ for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ with $\sum_i \alpha_i = 0$ and $z_1, \ldots, z_n \in \mathrm{Z}$.

Theorem 22) show that ζ_ρ^{ED} coincides with $2\mathrm{MMD}_K^2$ where ρ is the metric associated with the kernel $K(\cdot, \cdot)$, i.e. $\rho(z, z') = K(z, z) + K(z', z') - 2K(z, z')$, for $z, z' \in \mathrm{Z}$. Therefore, if ξ_ρ^{ED} is the sensitivity measure associated to the energy distance (40) according to Definition 2.2 and K is the kernel associated with ρ, ξ_ρ^{ED} coincides with ξ^{MMD} (36) and it can be simultaneously interpreted as an energy-distance-based or a kernel-based measure of association.

3 An application: functional vs. statistical dependence

Achieving a correct interpretation of the dependence between random variables is essential for statistical modelling. The task is difficult because the spectrum goes from a complete lack of relationship to a fully deterministic dependence, and pairwise statistical independence does not rule out that Y might still be a function of X, as we are to see. Without a claim of exhaustiveness, the examples of this section provide a conceptual illustration of the nuances involved and how the properties we illustrated help guide the inference.

We consider a random vector in $\mathbb{R}^3$, (Y, X_1, X_2), with $Y = g(X_1, X_2) = X_1 \cdot X_2$, and let X_1 be uniformly distributed on $[-1, 1]$ [we write $X_1 \sim U(-1, 1)$], while X_2 is discrete uniformly distributed with support $\{-1, 1\}$. We consider the following three cases for the joint probability of (X_1, X_2):

1. X_1 and X_2 are statistically independent;
2. X_2 is a function of X_1 given by $X_2 = \begin{cases} -1 & \text{if } |X_1| \in \left[\frac{1}{2}, 1\right], \\ 1 & \text{if } |X_1| \in \left[0, \frac{1}{2}\right); \end{cases}$
3. X_2 is a function of X_1 given by $X_2 = -\mathrm{sign}(X_1)$.

Note that in cases 2 and 3, one removes the statistical independence between X_1 and X_2.

Case 1. Analytical calculations show the Pearson's linear correlation coefficient, denoted as $\varrho(Y, X_i)$, and correlation ratio $\eta^2(Y, X_i)$ are null for both X_1 and X_2. However, these values do not allow us to conclude that Y does not depend on X_1 or on X_2, as neither $\varrho(Y, X_i)$ nor $\eta^2(Y, X_i)$ satisfies zero-independence. However, it is $Y \sim U(-1, 1)$, $Y \mid X_2 = 1 \sim U(-1, 1)$, and $Y \mid X_2 = -1 \sim U(-1, 1)$. Thus, Y and X_2 are statistically independent. This explains the null values of $\varrho(Y, X_2)$ and $\eta^2(Y, X_2)$. Moreover, we have $\xi^{\zeta}(Y, X_2) = 0$ for any $\xi^{\zeta}(Y, X)$ satisfying zero-independence. Suppose we select a measure of statistical association $\xi^{\mathrm{TV}}(Y, X_1)$ based on the separation measurement in (26). We find $\xi^{\mathrm{TV}}(Y, X_1) = 1$ and $\xi^{\mathrm{TV}}(Y, X_2) = 0$ by zero-independence. As unity is the maximum value for $\xi^{\mathrm{TV}}(Y, X_i)$, the fact that $\xi^{\mathrm{TV}}(Y, X_1) = 1$ may lead us to think that there is a deterministic dependence between Y and X_1. However, $\xi^{\mathrm{TV}}(Y, X_1)$ does not satisfy max-functionality. Instead, consider using a measure of association based on the Wasserstein-2 distance, which is strictly convex on Dirac masses. We obtain

$$\iota^{\mathrm{W}_2^2}(Y, X_1) = \frac{\mathbb{E}_{X_1}\left[\mathrm{W}_2^2\left(U(-1, 1), \left(\frac{1}{2}\delta_{X_1} + \frac{1}{2}\delta_{-X_1}\right)\right)\right]}{2\mathbb{V}[Y]} = \frac{\int_{-1}^{1}\left(x_1^2 - x_1 + \frac{1}{3}\right)\frac{1}{2}\mathrm{d}x_1}{2\mathbb{V}[Y]} = \frac{1}{4} \tag{41}$$

and $\iota^{\mathrm{W}_2^2}(Y, X_2) = 0$. The fact that $\iota^{\mathrm{W}_2^2}(Y, X_1)$ is not maximal indicates that the dependence between Y and X_1 is not deterministic. The case discussed here stems from Example 3.7 in Handoko (2022). Our analysis provides a solution to the issues raised there. In particular, by selecting a measure of statistical association that possesses the max-functionality property, we can make the correct inference about this input–output mapping: only X_1 concurs in determining the distribution of Y, however the dependence is not deterministic. Regarding X_2, we have the case of a model response which is statistically independent of it, although Y is a function of X_2. This shows us once more that the notions of functional and statistical dependence are distinct and they do not imply each other.

Case 2. It is $Y \sim U(-1, 1)$, $Y \mid X_2 = -1 \sim U([-1, -(1/2))p((1/2), 1])$, and $Y \mid X_2 = 1 \sim U([-(1/2), 1/2])$. Thus, Y and X_2 are statistically dependent. However, we find $\varrho(Y, X_2) = \eta^2(Y, X_2) = 0$. Using $\xi^{\mathrm{TV}}(Y, X_i)$ and $\iota^{W_2^2}(Y, X_i)$, we find $\xi^{\mathrm{TV}}(Y, X_2) = 1/2$ and $\iota^{W_2^2}(Y, X_2) = 1/8$, which correctly signals the statistical dependence between Y and X_2. For the association of Y with X_1, we have $\varrho(Y, X_1) = -(3/4)$ and $\eta^2(Y, X_1) = 1$, and we also find $\xi^{\mathrm{TV}}(Y, X_1) = 1$ and $\iota^{W_2^2}(Y, X_1) = 1$. This last value indicates that the dependence between Y and X_1 is deterministic. In fact, fixing X_1 fully determines Y in this case, because X_2 is itself a deterministic function of X_1.

Case 3. We have $Y \sim U([-1, 0])$. The mapping turns into $Y = -|X_1|$. We find $\eta^2(Y, X_1) = \xi^{\mathrm{TV}}(Y, X_1) = \iota^{W_2^2}(Y, X_1) = 1$ and $\eta^2(Y, X_2) = \xi^{\mathrm{TV}}(Y, X_2) = \iota^{W_2^2}(Y, X_2) = 0$. Here, again, conditioning on X_2 does not change the output distribution. So the pair (Y, X_2) is statistically independent, although X_2 and Y are both functions of X_1. Thus, removing the independence between X_1 and X_2 in this example yields a cornucopia of possibilities with subtle differences. Overall, the examples underscore the importance of carefully selecting the measure of statistical association to avoid misleading inferences.

4 Convexity and estimation

In most applications, measures of statistical association in (3) need to be computed numerically. Realizations of X and Y can be generated by a computer experiment (simulation setup) or collected from measurements of natural or social phenomena (data setup). A design that implements (3) directly rests on a double loop of Monte Carlo simulations and requires a computer model that links realizations of X to the corresponding values of Y (simulation setup). It is not applicable in the data setup. The numerical cost of this strategy is $C^{\text{Brute Force}} \approx n_X N^2$, with a linear dependence on n_X, the number of input variables or features X, and a quadratic dependence on the sample size N. However, the nearest-neighbour method and Pearson's partitioning allow us to estimate measures of association in (3) from an individual sample, with a notable theoretical reduction in computational burden. Sections 4.1 and 4.2 present general results for these two approaches. Section 4.3 performs illustrative numerical experiments for both methods.

4.1 Nearest neighbours

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. $\eta \in \mathscr{P}(\mathrm{X} \times \mathrm{Y})$ such that $X_1 \sim \mu \in \mathscr{P}(\mathrm{X})$ and $Y_1 \sim \nu \in \mathscr{P}(\mathrm{Y})$. Assume that X is endowed with metric ρ. We need to estimate $\mathbb{E}[\zeta(\nu, \nu_x^{\mathcal{F}})]$. We may estimate ν by the empirical distribution of $Y_1, \ldots, Y_n$ and the expectation w.r.t. μ via the empirical average over $X_1, \ldots, X_n$. The most challenging part is to estimate the conditional measure ν_x. We approximate ν_x for $x = X_i$, where $1 \le i \le n$, by taking the empirical distribution of the subset $\{Y_j : X_j$ is 'close' to $X_i\}$. The notion of being close to each other can be formalized as follows (see Bhattacharya, 2019; Deb et al., 2020). Let S be a finite subset of X. One says that $\mathcal{G}$ is a geometric graph functional on X if, for any S, it defines a graph $(S, \mathcal{E}(\mathcal{G}(S)))$, whose vertex set and edge set are S and $\mathcal{E}(\mathcal{G}(S))$, respectively.

Next, we let $\mathcal{G}_n = \mathcal{G}(X_1, \ldots, X_n)$ denote a graph on the n points $X_1, \ldots, X_n \in \mathrm{X}$. Let $\hat{\nu}_n = (1/n)\sum_{i=1}^{n} \delta_{\mathrm{Y}_i}$ be the empirical distribution of $Y_1, \ldots, Y_n$ and $\hat{\nu}_n(X_i; \mathcal{G}_n)$ the empirical distribution of $\{Y_j : (i, j) \in \mathcal{E}(\mathcal{G}_n)\}$, i.e.

$$\hat{\nu}_n(X_i; \mathcal{G}_n) = \frac{1}{d_i} \sum_{j\,:\,(i,j) \in \mathcal{E}(\mathcal{G}_n)} \delta_{\mathrm{Y}_j}. \tag{42}$$

In Equation (42), $(i, j) \in \mathcal{E}(\mathcal{G}_n)$ if and only if there is an edge $i \to j$ or $j \to i$ in $\mathcal{G}_n$. We call $d_i = \#\{j : (i, j) \in \mathcal{E}(\mathcal{G}_n)\}$ the degree of X_i.

Using (42), an estimate of the separation measure $\xi^{\zeta}(Y, X)$ can be defined as

$$\xi_n^{\zeta} := \frac{1}{n} \sum_{i=1}^{n} \zeta\big(\hat{\nu}_n, \hat{\nu}_n(X_i; \mathcal{G}_n)\big). \tag{43}$$

We then ask whether ξ_n^{ζ} consistently estimates ξ^{ζ} as $n \to \infty$. This question can be addressed under the theoretical framework developed in Deb et al. (2020). We state the following three assumptions (Deb et al., 2020, cf. conditions (A1)–(A3) and Bhattacharya, 2019, cf. conditions N1 and N2):

(A1) Small ball probability: The metric ρ on X is such that $\rho(X_i, X_{N(i)}) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, where $N(i)$ is uniformly sampled from the indices of the neighbours $\{j : (i, j) \in \mathcal{E}(\mathcal{G}_n)\}$ of X_i.

(A2) Local graph functional:

- (a) There exists sequence of natural numbers r_n (which may be unbounded, $r_n \to \infty$) such that almost surely (a.s.) $\min_{1\le i\le n} d_i \ge r_n$;
- (b) Let us denote $\mathcal{G}_{n,i}$ to be the perturbation of the graph $\mathcal{G}_n$ where X_i is replaced by an i.i.d. element X_i', i.e. $\mathcal{G}_{n,i} = \mathcal{G}(X_1, \ldots, X_i', \ldots, X_n)$. There exists a sequence $q_n > 0$, not necessarily bounded, such that we have $\max_{1\le i\le n} \max\{|\mathcal{E}(\mathcal{G}_n) \setminus \mathcal{E}(\mathcal{G}_{n,i})|, |\mathcal{E}(\mathcal{G}_{n,i}) \setminus \mathcal{E}(\mathcal{G}_n)|\} \le q_n$, a.s.;
- (c) One choice of q_n and r_n must satisfy

$$\limsup_{n\to\infty} \frac{q_n}{r_n} \le C \quad \text{with} \quad C > 0. \tag{44}$$

(A3) Asymptotic degrees of same order: There is a sequence t_n (which may be unbounded) such that, a.s., $\max_{1\le i\le n} d_i \le t_n$ and $\limsup_{n\to\infty} (t_n/r_n) \le D > 0$.

Theorem 4.1 Assume that $\mathcal{G}_n$ satisfies the assumptions (A1)–(A3). Furthermore, suppose that ζ is continuous in the weak topology and convex on Dirac masses and $\int_{\mathsf{X}} \zeta(\nu, \nu_x^{\mathcal{F}}) d\mu(x) < \infty$. Then $\xi_n^{\zeta} \xrightarrow{\mathbb{P}} \xi^{\zeta}(Y, X)$. Furthermore, if $\int_{\mathsf{X}} \zeta^2(\nu, \nu_x^{\mathcal{F}}) d\mu(x) < \infty$, then $\xi_n^{\zeta} \xrightarrow{a.s.} \xi^{\zeta}(Y, X)$.

In the following result, we provide a sub-Gaussian concentration bound on ξ_n^{ζ} in the spirit of Deb et al. (2020).

Proposition 4.2 Continuing with the same assumptions as in 4.1 and provided $\sup_{x\in\mathsf{X}} \zeta(\nu, \nu_x) \le M$ for some $M > 0$, there exists $C > 0$ (independent of n and t), such that for any $t > 0$

$$\mathbb{P}\big[|\xi_n^{\zeta} - \mathbb{E}[\xi_n^{\zeta}]| \ge t\big] \le 2\exp(-Cnt^2), \tag{45}$$

and consequently, $\sqrt{n}(\xi_n^{\zeta} - \mathbb{E}[\xi_n^{\zeta}]) = \mathcal{O}_{\mathbb{P}}(1)$.

This last equality reassures us that the variance of the estimator is bounded. One of the most prominent questions about the estimation of association measures is whether there is a CLT under independence between Y and X. This is particularly useful for creating a test for independence. We can exploit (Deb et al., 2020, Theorem 4.1) to obtain such result.

In the remainder, let $\Phi(\cdot)$ denote the cumulative distribution function of the Gaussian distribution. With $\mathcal{G}_n$, $(d_1, \ldots d_n)$, q_n, t_n, and r_n as previously defined, let K be a characteristic kernel. Recall that if $Y_1 \sim \nu$ and $Y_2 \sim \mu$, then,

$$\zeta^{\mathrm{ED}}(\nu, \mu) = 2\mathbb{E}[K(Y_1, Y_1')] + 2\mathbb{E}[K(Y_2, Y_2')] - 4\mathbb{E}[K(Y_1, Y_2)], \tag{46}$$

where Y_1' and Y_2' are independent replicas of Y_1 and Y_2. From the above expression, the proposed estimator of $\xi^{\mathrm{ED}}(Y, X)$ can be written as

$$\xi_n^{\mathrm{ED}} := \frac{4}{n(n-1)} \sum_{i=1}^{n} \sum_{j\ne i} K(Y_i, Y_j) - \frac{4}{n} \sum_{i=1}^{n} d_i^{-1} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n)} K(Y_i, Y_j). \tag{47}$$

Proposition 4.3 Under (A1)–(A3), suppose that $\mathbb{E}[K^2(Y_1, Y_2)] < \infty$. Then, $\mathbb{V}(\sqrt{n}\hat{\xi}_n^{\mathrm{ED}}) = \mathcal{O}(1)$. In addition, let $\theta = (M, D, \gamma, \epsilon) \in (0, \infty)^3 \times [0, 1/3)$. Let us pay attention to graph functionals and measures on $\mathrm{X} \times \mathrm{Y}$ of the type:

$$\mathcal{J}_\theta := \left\{ (\tilde{\mathcal{G}}, \tilde{\mu}) : \mathbb{E}[K^4(Y_1, Y_2)] \le M,\ \limsup_{n\to\infty} \max_{1\le i\le n} \frac{\tilde{d}_i}{(\log n)^\gamma} \le D \right.$$
$$\left. w.p.1,\ r_n^{-1}(q_n + t_n) \le D,\ n^\epsilon \mathrm{Var}(N_n) \ge 1 \ \textit{for all}\ n \ge M \right\},$$

where $(\tilde{d}_1, \ldots, \tilde{d}_n)$ is the sequence of degrees of $\tilde{\mathcal{G}}_n := \tilde{\mathcal{G}}(X_1, \ldots, X_n)$, with $(X_i, Y_i) \overset{i.i.d.}{\sim} \tilde{\mu}$, $i = 1, \ldots, n$, and $\tilde{\mu} = \tilde{\mu}_X \otimes \tilde{\mu}_Y$. For every $\theta \in (0, \infty)^3 \times [0, 1/3)$, we have

$$\lim_{n\to\infty} \sup_{(\tilde{\mathcal{G}}, \tilde{\mu}) \in \mathcal{J}_\theta} \sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\hat{\xi}_n^{\mathrm{ED}}}{\tilde{S}_n} \le z \right) - \Phi(z) \right| = 0 \quad \textit{where} \tag{48}$$

$$\tilde{S}_n^2 := \tilde{a}\left(\tilde{g}_1 + \tilde{g}_3 - \frac{2}{n-1} \right) + \tilde{b}\left(\tilde{g}_2 - 2\tilde{g}_1 - 2\tilde{g}_3 - 1 + \frac{4}{n-1} \right)$$
$$+ \tilde{c}\left(\tilde{g}_1 + \tilde{g}_3 - \tilde{g}_2 + \frac{n-3}{n-1} \right),$$

with $\tilde{a} := (1/(n(n-1))) \sum_{(i,j)\ \text{distinct}} K^2(Y_i, Y_j)$, $\tilde{b} := (1/(n(n-1)(n-2))) \sum_{(i,j,l)\ \text{distinct}} K(Y_i, Y_j)K(Y_i, Y_l)$, $\tilde{c} := (1/(n(n-1)(n-2)(n-3))) \sum_{(i,j,l,m)\ \text{distinct}} K(Y_i, Y_j)K(Y_l, Y_m)$, and

$$\tilde{g}_1 := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tilde{d}_i}, \quad \tilde{g}_2 := \frac{1}{n}\sum_{i,j} \frac{T^{\tilde{\mathcal{G}}_n}(i, j)}{\tilde{d}_i \tilde{d}_j}, \quad \tilde{g}_3 := \frac{1}{n} \sum_{(i,j)\in\mathcal{E}(\mathcal{G}_n)} \frac{1}{\tilde{d}_i \tilde{d}_j},$$

with $T^{\tilde{\mathcal{G}}_n}(i, j) := \sum_k \mathbf{1}((i, k) \in \mathcal{E}(\tilde{\mathcal{G}}_n))\mathbf{1}((k, j) \in \mathcal{E}(\tilde{\mathcal{G}}_n))$ counting the number of neighbours shared by X_i and X_j.

Proposition 4.3 shows the $\sqrt{n}$-consistency of $\hat{\xi}_n^{\mathrm{ED}}$ and establishes a CLT for the same under the independence between Y and X for a broad family of graph functionals. By Theorem 4.3, the limiting distribution of $\sqrt{n}\hat{\xi}_n^{\mathrm{ED}}/\tilde{S}_n$ is standard normal. We can then build a test for

$$\mathrm{H}_0 : \eta = \mu \otimes \nu \quad \text{versus} \quad \mathrm{H}_1 : \eta \neq \mu \otimes \nu. \tag{49}$$

Let z_α be the upper quantile of the standard Gaussian distribution. If $\sqrt{n}\hat{\xi}_n^{\mathrm{ED}} \ge z_\alpha \tilde{S}_n$, we reject H_0.

By Theorem 4.1, the asymptotic level of this test is α. Many other testing procedures such as the ones based on distance-covariance (Székely et al., 2007) and or the Hilbert-Schmidt independence criterion (Gretton et al., 2005) lack such a simple limiting distribution theory. Despite some recent works on finite sample guarantees of those tests (e.g. Albert et al., 2022), the associated permutation-based methods to estimate the rejection threshold make them computationally expensive.

4.2 Pearson estimation

To fix ideas, let X be continuous. The intuition of Pearson (1905) is to partition the support of X (X), in H classes X_h, $h = 1, 2, \ldots, H$, and replacing the point condition $X = x$ with the class condition $X \in \mathrm{X}_h$. A Pearson estimate of $\xi^\zeta(Y, X)$ in (3) is then given by

$$\widehat{\xi^\zeta}_H(Y, X) = \sum_{h=1}^{H} \zeta(\hat{\nu}, \widehat{\nu}_{X\in \mathrm{X}_h})\frac{n_h}{N}, \tag{50}$$

where N is the sample size, $\widehat{(\cdot)}$ denotes an empirical estimate, $\nu_{X\in\mathrm{X}_h}$ is the conditional probability of Y given that X is in X_h, and n_h is the number of realizations of X that fall in class X_h. For the moment assume that the sample size is sufficient to allow for statistically accurate estimates, so that we can write $\zeta(\hat{\nu}, \widehat{\nu}_{X\in\mathrm{X}_h})\zeta(\nu, \nu_{X\in\mathrm{X}_h})$. Then, consider a sequence of refining partitions, in which the cardinality increases and the next partition is finer than the previous one. Intuitively, as the cardinality H increases, the class condition becomes progressively closer to a point condition.

This intuition is at the basis of the estimation of Pearson's correlation ratio in (21) (Pearson, 1905). However, the approach has been extended to estimate information value and distribution-based indices (Borgonovo et al., 2014; Strong & Oakley, 2013). Nonetheless, we are not aware of a unifying result about the asymptotic consistency of the estimates. Theorem 2.13 provides a formal basis to obtain such a result.

Consider a sequence of random variables (X^N, Y^N), $N \in \mathbb{N}$, defined in some (standard Borel) measure space $(\Omega^N, \mathcal{F}^N, \mathbb{P}^N)$ with values in $\mathrm{X} \times \mathrm{Y}$ such that the joint law $\pi^N = (X^N, Y^N)_\sharp \mathbb{P}^N$ is weakly converging to $\pi = (X, Y)_\sharp \mathbb{P}$. We will also suppose that a continuous function $\Upsilon : \mathrm{Y} \to [0, +\infty)$ is given such that

$$\lim_{N\to\infty} \mathbb{E}^N[\Upsilon(Y^N)] = \mathbb{E}[\Upsilon(Y)] < \infty. \tag{51}$$

A typical example is given by $\Omega^N := \{1, 2, \ldots, N\}$ with the uniform measure and $X^N(n) := X_n(\omega)$, $Y^N(n) := Y_n(\omega)$, $n = 1, \ldots, N$, are obtained by evaluating a sequence $(X_n, Y_n)_{n\in\mathbb{N}}$ of mutually independent random variables sharing the same joint law of (X, Y).

Here, we can distinguish the simpler case when X takes values in a finite set (or, equivalently, $\mathcal{F}$ is finite) from the general one. In the finite case, we will assume that $\mathrm{X} = \{x^1, x^2, \ldots, x^H\}$ is a finite set, $\mathcal{F} = 2^{\mathrm{X}}$, and we consider the quantity

$$\xi^\zeta_N := \xi^\zeta(Y^N, X^N). \tag{52}$$

In the general case, we introduce a countable collection of measurable partitions given by $\mathscr{X}^M = \{\mathrm{X}^M_h\}_{h=1,\ldots,H(M)}$, $M \in \mathbb{N}$, which generate a corresponding family of σ-algebras $\mathcal{F}^M = \sigma(\mathscr{X}^M)$ satisfying

$$\mathcal{F}^M \subset \mathcal{F}^{M+1}, \qquad \bigvee_{M\in\mathbb{N}} \mathcal{F}^M = \mathcal{F}, \qquad \mu(\partial \mathrm{X}^M_h) = 0 \quad \text{for every } M \in \mathbb{N},\ 1 \le h \le H(M). \tag{53}$$

We then consider the quantities

$$\xi^\zeta_{M,N} := \xi^\zeta(Y^N, (X^N, \mathcal{F}^M)). \tag{54}$$

Definition 4.4 Let $\Upsilon : \mathrm{Y} \to [0, \infty)$ be a continuous function with finite expectation as per (51). We say that ζ is Υ-weakly continuous if for every pair of sequences $(\nu_n)_{n\in\mathbb{N}}$ and $(\nu'_n)_{n\in\mathbb{N}}$ in $\mathscr{P}(\mathrm{Y})$ weakly converging to ν and ν', respectively, as $n \to \infty$ and additionally satisfying

$$\nu'_n \le C\nu_n \quad \textit{for a suitable constant } C > 0 \textit{ and for every } n \in \mathbb{N}, \tag{55}$$

$$\lim_{n\to\infty} \int_{\mathrm{Y}} \Upsilon \, \mathrm{d}\nu_n = \int_{\mathrm{Y}} \Upsilon \, \mathrm{d}\nu < \infty, \tag{56}$$

it holds $\zeta(\nu_n, \nu'_n) \to \zeta(\nu, \nu')$ as $n \to \infty$.

Theorem 4.5 Suppose that $\zeta(\cdot\,,\ \cdot)$ is convex, lower semicontinuous in its second argument, and conditionally continuous according to Definition 4.4.

1. If X is finite, then $\lim_{N\to\infty}\xi_N^{\zeta} = \xi^{\zeta}(Y, X)$.
2. In the general case, if (53) holds true,

$$\lim_{M\to\infty}\lim_{N\to\infty}\xi^{\zeta}_{M,N} = \xi^{\zeta}(Y, X). \tag{57}$$

Theorem 4.5 implies that for N sufficiently large, calculations of $\xi^{\zeta}_{M,N}$ should approximate the true value $\xi^{\zeta}(Y, X)$ from below. Note that in (57), the two limits are nested: We are letting first $N \to \infty$, and then $M \to \infty$. This fact has a practical consequence that we discuss in the next subsection devoted to numerical experiments.

We provide a CLT for Pearson's type estimators. More precisely, we show such a result when ζ_ρ is a kernel-based distance measure and the corresponding measure of association $\widehat{\xi^{\zeta}}_{H_n}(Y, X)$ is defined via (50). Following (46), we may rewrite $\widehat{\xi^{\zeta}}_{H}(Y, X)$ as

$$\widehat{\xi^{\zeta}}_{H_n}(Y, X) := \frac{4}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}K(Y_i, Y_j) - \frac{4}{n}\sum_{i=1}^{n}\frac{1}{\tilde{d}_i - 1}\sum_{j\,:\,X_j\in\mathsf{X}_h(i), j\neq i}K(Y_i, Y_j), \tag{58}$$

where $\mathsf{X}_h(i)$ denotes the class in X to which X_i belongs and $\tilde{d}_i$ denotes the total number of X_j belonging to $\mathsf{X}_h(i)$. We need the following assumptions on the classes $\{\mathsf{X}_h : 1 \le h \le H_n\}$ which are similar to conditions (A1)–(A3) of the solution:

(B1) Small ball neighbourhood: Let $\tilde{N}(1), \ldots, \tilde{N}(n)$ be a sequence of independent and uniformly sampled indices from the class $\mathsf{X}_h(i)$, with $\rho(X_i, X_{\tilde{N}(i)}) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$, for all i.

(B2) Populated neighbourhood: Assume that the number of X_j in each class X_h for $h \in \mathbb{Z}_{[1,H_n]}$ are uniformly lower bounded, i.e. there exists a deterministic positive sequence $\tilde{r}_n \to \infty$, such that $\min_{1\le i\le n}\tilde{d}_i := |\{j : X_j \in \mathsf{X}_h\}| \ge \tilde{r}_n$ w.p. 1.

(B3) Asymptotic degrees of same order: the total numbers of X_j in each X_h should be asymptotically of the same order, i.e. we have a deterministic sequence $\tilde{t}_n$ (bounded or not) such that: $\max_{1\le i\le n}\tilde{d}_i \le \tilde{t}_n$ w.p. 1, and $\limsup_{n\to\infty}(\tilde{t}_n/\tilde{r}_n) \le C$, for some real number $C > 0$.

Theorem 4.6 Let Y be independent of X. Assume that (B1)–(B3) hold and $\mathbb{E}[K^2(Y_1, Y_2)] < \infty$. Further assume that $\tilde{d}_i/(\log n)^{\gamma} \le D$ for all $i \in \mathbb{Z}_{[1,n]}$ and for some $\gamma > 0$. Then the following result holds for every fixed $\gamma \in (0, \infty)$:

$$\lim_{n\to\infty}\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\sqrt{n}\widehat{\xi^{\zeta}}_{H_n}(Y, X)}{\mathfrak{S}_n} \le z\right) - \Phi(z)\right| = 0, \tag{59}$$

where $\tilde{a}$, $\tilde{b}$, $\tilde{c}$ are same as in Theorem 4.3 and

$$\mathfrak{S}_n^2 := \tilde{a}\left(2\tilde{p}_1 - \frac{2}{n-1}\right) + \tilde{b}\left(\tilde{p}_2 - 4\tilde{p}_1 - 1 + \frac{4}{n-1}\right) + \tilde{c}\left(2\tilde{p}_1 - \tilde{p}_2 + \frac{n-3}{n-1}\right),$$

with $\tilde{p}_1 := (1/n)\sum_{i=1}^{n}(1/\tilde{d}_i)$ and $\tilde{p}_2 := (1/n)\sum_{i,j}(\tilde{\Delta}(i, j)/\tilde{d}_i\tilde{d}_j)$. Here $\tilde{\Delta}(i, j) := \sum_k \mathbf{1}(X_i, X_j, X_k \in \mathsf{X}_h$ for some $h)$ is the cardinality of the set of neighbours shared by X_i and X_j.

Thus, Proposition 4.3 and Theorem 4.6 show that not only the nearest-neighbour-type estimates but also the Pearson-type estimates of $\xi^{\zeta}(Y, X)$ are asymptotically normal when Y and X are independent.

4.3 Numerical experiments

In this section, we report the results of numerical experiments aimed at testing several of the theoretical aspects discussed in the previous part of the work, as well as the asymptotic unbiasedness and CLT for Pearson-type estimators. We consider a multivariate output problem, derive analytical results needed to construct the analytical benchmarks, and perform the numerical experiments.

We select an optimal transport-based measure of association,

$$\xi^{\mathrm{W}_2^2}(Y, X_i) = \mathbb{E}\left[\zeta^{\mathrm{W}_2^2}(\mathbb{P}_Y, \mathbb{P}_{Y\,|\,X_i})\right], \tag{60}$$

and a kernel-based one

$$\xi^{\mathrm{ED}}(Y, X_i) = \mathbb{E}[\zeta^{\mathrm{ED}}(\mathbb{P}_Y, \mathbb{P}_{Y\,|\,X_i})]. \tag{61}$$

$\xi^{\mathrm{W}_2^2}(Y, X_i)$ is the unnormalized version of $\iota^{\mathrm{W}_2^2}(Y, X_i)$ in (41). In $\xi^{\mathrm{ED}}(Y, X_i)$, we use the Euclidean norm as a kernel.

To construct the analytical benchmarks, we proceed as follows. For $\xi^{\mathrm{W}_2^2}(Y, X_i)$ we exploit the values in Borgonovo et al. (2024), for $\xi^{\mathrm{ED}}(Y, X_i)$ we build them anew. We recall that a random variable Z with mean μ_Z and variance–covariance metric Σ_Z is elliptically distributed if its characteristic function can be written as $F_Z(z; \mu_Z, \Sigma_Z^*) = e^{iz^T\mu_Z^*}G(z^T\Sigma_Z^* z)$, where $G : \mathbb{R}^+ \to \mathbb{R}^+$ is called the generator (see Cambanis et al., 1981, Theorem 2 for technical conditions), and μ_Z^* and Σ_Z^* are the distribution parameters. Synthetically, we write $Z \sim \mathcal{EC}(\mu_Z^*, \Sigma_Z^*, G)$ as in Cambanis et al. (1981). If the first moment and second moments are finite, then $\mu_Z = \mu_Z^*$ and $\Sigma_Z^* = \Sigma_Z$ (see Cambanis et al., 1981; Landsman & Valdez, 2003).

Proposition 4.7 Let $X = (X_1, X_2, \ldots, X_{n_X}) \sim \mathcal{EC}(m_X, \Sigma_X, G)$, with $m_X = (m_1, m_2, \ldots, m_{n_X})$, with finite second moment. Given $A \in \mathbb{R}^{n_Y \times n_X}$ and $b \in \mathbb{R}^{n_Y}$, if $Y = AX + b$ then $\xi^{\mathrm{ED}}(Y, X_i)$ is given by

$$\xi^{\mathrm{ED}}(Y,X_i) = \frac{1}{c_{n_Y}}\mathbb{E}\left[\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\frac{\|F_Y(t;m_Y,\Sigma_Y) - F_{Y|X_i}(t;m_{Y|X_i},\Sigma_{Y|X_i})\|^2}{\|t\|^{n_Y+1}}\mathrm{d}t\right], \tag{62}$$

where $c_{n_Y} = (\pi^{(n_Y+1)/2})/(\Gamma((n_Y+1)/2))$, $m_Y = Am_X^T + b$, $\Sigma_Y = A\Sigma_X A^T$, $\Sigma_{Y|X_i} = A\Sigma_i^c A^T$,

$$\Sigma_i^c = (\sigma_{t,j}^i)_{t,j=1,2,\ldots,n_X}, \quad \sigma_{t,j}^i = \sigma_{t,j} - \frac{\sigma_{t,i}\cdot\sigma_{i,j}}{\sqrt{\sigma_{i,i}}}, \tag{63}$$

and $m_{Y_k|X_i} = \sum_{j=1}^{n_X} a_{k,j}(m_j + (X_i - \mu_i)(\sigma_{i,j}^i/\sigma_{i,i}^i))$, for $k = 1, 2, \ldots, n_Y$.

Once $F_Y(t;\, m_Y, \Sigma_Y)$ is specified, Equation (62) can be implemented in a computer algebra system (here, Mathcad is used by the authors[2]) to obtain the corresponding numerical values. To set numerical benchmarks, we consider the following parameterization. We let $Y \in \mathbb{R}^2$ and $X \in \mathbb{R}^4$,

[2] All subroutines are available at https://github.com/emanueleborgonovo/ConvexityandMeasuresofStatisticalAssociation/tree/main.

Table 2. Analytical values of selected measures of association for our numerical implementation

	X_1	X_2	X_3	X_4	$Z = \sin(X_1)$, **estimated**
$\xi^{\mathrm{ED}}(Y, X_i)$	3.47	3.63	0.27	0	1.69
$\xi^{\mathrm{W}_2^2}(Y, X_i)$	6.47	6.52	2.86	0	3.50

$Y = AX$, with $A = \begin{bmatrix} 4 & -2 & 1 & 0 \\ 2 & 5 & -1 & 0 \end{bmatrix}$ with X normally distributed, with mean $\mu_X = (1, 1, 1, 1)$, and variance–covariance matrix $\Sigma_X = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0.5 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. Correspondingly, $\mu_Y = \begin{bmatrix} 3 & 6 \end{bmatrix}$ and $\Sigma_Y = \begin{bmatrix} 15 & 7.5 \\ 7.5 & 33 \end{bmatrix}$. The analytical values of $\xi^{\mathrm{ED}}(Y, X_i)$ and $\xi^{\mathrm{W}_2^2}(Y, X_i)$ are reported in Table 2. The variables X_1 and X_2 are the most important, X_3 is less influential. For X_4, we have $\xi^{\mathrm{ED}}(Y, X_4) = \xi^{\mathrm{W}_2^2}(Y, X_4) = 0$ by zero-independence.

Numerically, we generate a sample of size N from the distribution of X and calculate the corresponding values of Y. We then implement the Pearson estimators (50) of $\xi^{\mathrm{ED}}(Y, X_i)$ and $\xi^{\mathrm{W}_2^2}(Y, X_i)$. Here, the partition cardinality H plays a relevant role. We use equipopulated partition sets, so that the subsets X_b^M contain approximately the same number of realizations of X and link H to the sample size as to avoid scarcely populated partition sets that would compromise statistical accuracy.

We begin with experiments that test asymptotic convergence. We increase the sample size from $N = 100$ to $N = 2^{15}$ and select the partition cardinality starting with $H = 5$ when $N = 64$ and ending with $H = 100$ when $N = 2^{18}$. Results show that the estimates converge asymptotically to the corresponding analytical values and convergence is from below, in accordance with Theorems 2.13 and 4.5. We visualize results for experiments with fixed sample size and increasing partition cardinality. We set $N = 2^{15}$ and let H increase up to $H = 100$, with the other settings identical to the previous experiment.

The two panels in Figure 2 show that the estimates converge to the analytical values from below as H increases, again in accordance with Theorems 2.13 and 4.5. In fact, at large sample sizes, refining the partition can be interpreted as receiving increasingly precise information on X_i and therefore as obtaining an algebra which is getting closer to the algebra generated by X_i. We also observe that the estimates are almost insensitive to choices of the partition size for values of between 30 and 100. This *plateau* effect is in line with results in Strong and Oakley (2013), who document this effect for the first time. Next, we consider an experiment in which instead of receiving information about X_1, we receive the distorted information $Z_2 = g(X_1) = sin(X_1)$. Analytical values are now out of reach, but numerical calculations show that the values of $\xi^{\mathrm{ED}}(Y, X_1)$ and $\xi^{\mathrm{W}_2^2}(Y, X_1)$ decrease, respectively, from 3.47 down to 1.69 and from 6.47 down to 3.50 (see the last column of Table 2). This result is in accordance with Theorem 2.5.

We then report results for simulations towards testing the central limit result in Theorem 4.6. We consider four experiments with sample sizes $N = 10{,}000$ and $N = 20{,}000$ and partition size fixed at $H = 32$. To test the asymptotic normality of the estimates, we replicate the experiments $R = 300$ times at each sample size generating a new input dataset using a crude Monte Carlo generator.

The left graph in Figure 3 displays the histogram of $\hat{\xi}^{\mathrm{ED}}(Y, X_4)$ over the 400 replicates at $N = 10,000$. The mean and standard deviation of $\hat{\xi}^{\mathrm{ED}}(Y, X_4)$ equal $1.20 \cdot 10^{-4}$ and $2.30 \cdot 10^{-3}$, respectively. Thus, the estimated mean is close to zero as per the analytical value. The right graph in Figure 3 displays the corresponding probability–probability plot that compares the empirical cumulative distribution function of $\widehat{\xi^{\mathrm{ED}}}(Y, X_4)$ against the corresponding theoretical values of a normal distribution with zero mean and standard deviation equal to the empirical standard deviation. The two dotted lines represent 5% confidence bounds. Because all the estimates are within the bounds, the normality test is passed at the 5% level in this experiment.

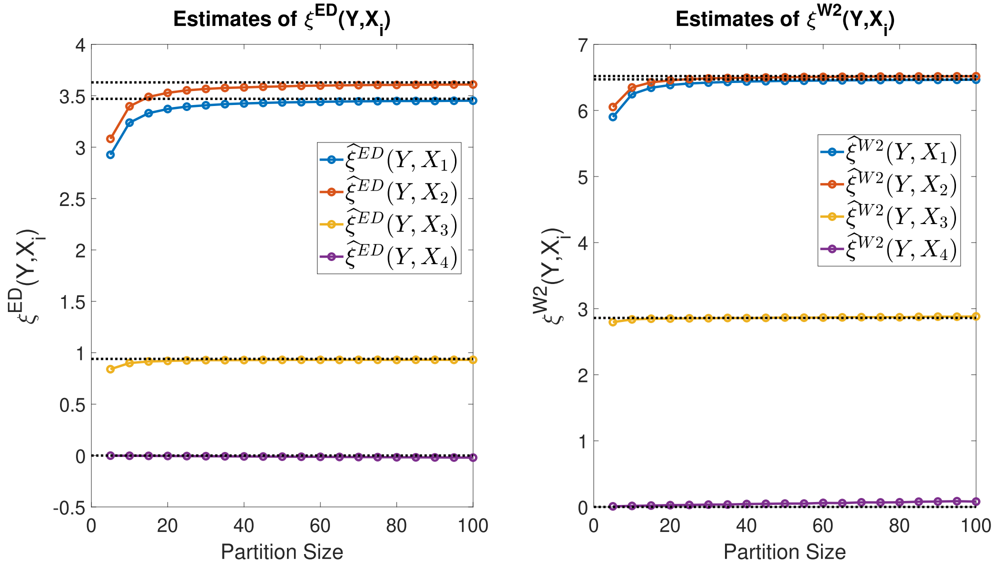

Figure 2. Vertical axis: estimates of $\xi^{ED}(Y, X_i)$ (left graph) and of $\xi^{W_2^2}(Y, X_i)$ (right graph); circles ° indicate estimates, dotted lines indicate analytical values. Horizontal axis: partition size increasing from $H = 10$ to $H = 100$. Sample size: $N = 2^{15}$.

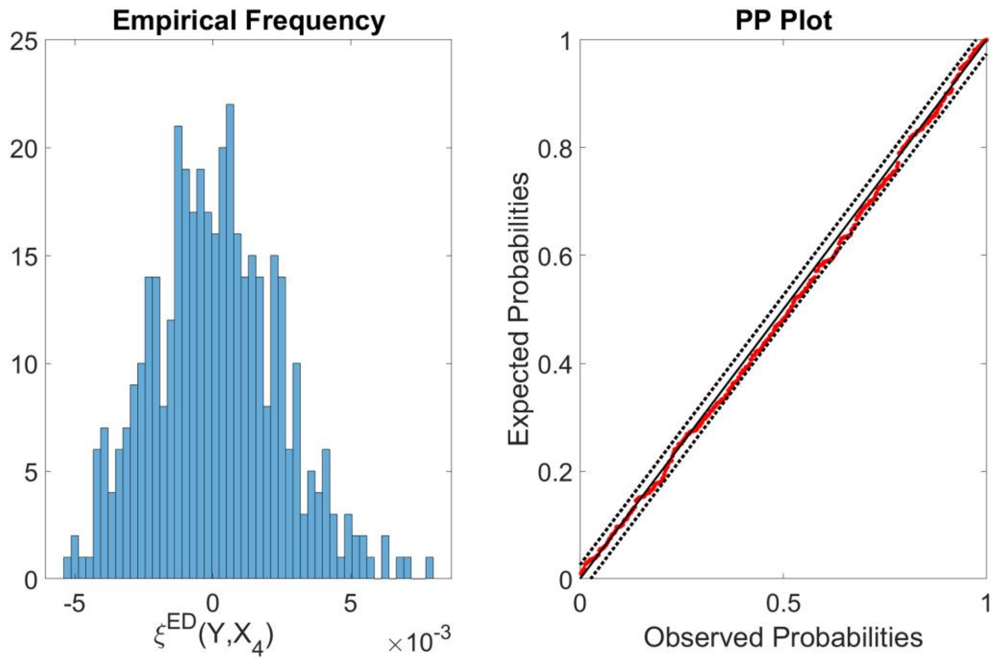

Figure 3. Histogram of estimates of $\xi^{ED}(Y, X_4)$. $N = 10{,}000$, $H = 20$, and $R = 400$.

5 Conclusions

Measures of statistical association may or may not possess properties such as zero-independence, max-functionality, and monotonicity. We have carried out an investigation at the root of the geometric features that make a measure of statistical association possess one or more of these properties. For non-constant random variables Y and X, we have focused on measures of statistical association based on a separation measurement between the marginal law of Y and the conditional laws of Y given X. We have assumed lower semicontinuity as a basic property of the separation measurement. We have then seen that convexity plays a special role (see Figure 1), as it ensures minimality, nonnegativity, monotonicity, and makes the measure of association maximal when $Y = g(X)$. To turn minimality and nonnegativity into zero-independence (i.e. to obtain necessary and sufficient conditions), we have shown that the crucial additional property on ζ is strictness. Notably, this property is not sufficient to make the measure of statistical association satisfy max-functionality. We have seen that max-functionality can be equivalently characterized by *strict* convexity on Dirac-δ masses. This result explains why some well-known indicators are maximal when there is a deterministic dependence between Y and X, and also when the dependence is noisy. These results have allowed us to analyse the properties of a sample of representative measures of statistical association, comprising Csiszár divergences, Chatterjee's new correlation coefficient, optimal transport-based and kernel-based measures of statistical association. Findings show that Csiszár divergences possess zero-independence and monotonicity, but do not satisfy max-functionality. For kernel-based measures of statistical association, we link their properties to relevant features of the corresponding kernels.

We have then focused on two classes of estimators for these measures of statistical association. The theoretical premises have allowed us to prove the asymptotic unbiasedness of estimates based on nearest neighbours and, for the first time, for Pearson's 1905 partition design. We have also proven CLTs that allow analysts to obtain information about the level at which we can confidently screen out a variable as irrelevant when Y and X are independent, and illustrated them via numerical experiments.

Our findings apply to measures of statistical association defined as expectations between separation measurements [Equation (3)]. This class does not encompass the entire spectrum of measures of association. For instance, Rényi's maximal correlation coefficient (Rényi, 1959) or the optimal transport dependency of Nies et al. (2023) do not belong to this family. In fact, a dependency can be constructed as $\tau(Y, X) = \zeta(\mathbb{P}_{YX}, \mathbb{P}_Y \otimes \mathbb{P}_X)$. Then, investigating the geometric features of $\zeta(\cdot\,,\ \cdot)$ that ensure desirable properties to measures of statistical association built on alternative rationales and how these features play a role in the different rationales is a future research avenue.

Acknowledgments

The authors thank the editor, Professor Daniela Witten, the associate editor, and three anonymous reviewers for several perceptive and insightful comments that have greatly helped us improve the manuscript. A.F. thanks the support of the Lagrange Mathematics and Computation Research Center.

Conflicts of interest: None declared.

Funding

E.P. is funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) under grant #02E21112B.

Data availability

The data and code underlying the article are available at: https://github.com/emanueleborgonovo/ConvexityandMeasuresofStatisticalAssociation.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Albert M., Laurent B., Marrel A., & Meynaoui A. (2022). Adaptive test of independence based on HSIC measures. *Annals of Statistics*, *50*(2), 858–879. https://doi.org/10.1214/21-AOS2129

Barr J., & Rabitz H. (2022). A generalized kernel method for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(1), 27–54. https://doi.org/10.1137/20M1354829

Bhattacharya B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(3), 575–602. https://doi.org/10.1111/rssb.12319

Borgonovo E., Figalli A., Plischke E., & Savaré G. (2024). Global sensitivity analysis via optimal transport. *Management Science*, *0*(0). https://doi.org/10.1287/mnsc.2023.01796

Borgonovo E., Tarantola S., Plischke E., & Morris M. (2014). Transformation and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society Series B*, *76*(5), 925–947. https://doi.org/10.1111/rssb.12052

Cambanis S., Huang S., & Simons G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, *11*(3), 368–385. https://doi.org/10.1016/0047-259X(81)90082-8

Chatterjee S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, *116*(536), 2009–2022. https://doi.org/10.1080/01621459.2020.1758115

Da Veiga S. Gamboa F., Iooss B. & Prieur C. (2021). *Basics and trends in sensitivity analysis: Theory and practice in R*. SIAM.

Deb N., Ghosal P., & Sen B. (2020). 'Measuring association on topological spaces using kernels and geometric graphs', arXiv, arXiv:2010.01768, https://doi.org/10.48550/arXiv.2010.01768, 2010:1–66, preprint: not peer reviewed.

Dette H., Siburg K. F., & Stoimenov P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics*, *40*(1), 21–41. https://doi.org/10.1111/j.1467-9469.2011.00767.x

Fissler T., & Pesenti S. (2023). Sensitivity measures based on scoring functions. *European Journal of Operational Research*, *307*(3), 1408–1423. https://doi.org/10.1016/j.ejor.2022.10.002

Gamboa F., Klein T., & Lagnoux A. (2018). Sensitivity analysis based on Cramér von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, *6*(2), 522–548. https://doi.org/10.1137/15M1025621

Geenens G., & de Micheaux P. L. (2022). The Hellinger correlation. *Journal of the American Statistical Association*, *117*(538), 639–653. https://doi.org/10.1080/01621459.2020.1791132

Glick N. (1975). Measurements of separation among probability densities or random variables. *Canadian Journal of Statistics*, *3*(2), 267–276. https://doi.org/10.2307/3315284

Gretton A., Bousquet O., Smola A., & Schölkopf B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, & E. Tomita (Eds.), *Algorithmic learning theory. 16th international conference, ALT 2005* (pp. 63–77). Springer Verlag.

Handoko B. (2022). *Sensitivity analysis and its role in expert judgment* [Ph.D. Thesis]. University of Sheffield. https://etheses.whiterose.ac.uk/31329/.

Hirschfeld H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 520–524. https://doi.org/10.1017/S0305004100013517

Hotelling H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377. https://doi.org/10.2307/2333955

Landsman Z. M., & Valdez E. A. (2003). Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, *7*(4), 55–71. https://doi.org/10.1080/10920277.2003.10596118

Liero M., Mielke A., & Savaré G. (2018). Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones Mathematicae*, *211*(3), 969–1117. https://doi.org/10.1007/s00222-017-0759-8

Mézard M., & Montanari A. (2009). *Information, physics, and computation*. Oxford University Press.

Mori T., & Szekely G. (2019). Four simple axioms of dependence measures. *Metrika*, *82*(1), 1–16. https://doi.org/10.1007/s00184-018-0670-3

Nies T. G., Staudt T., & Munk A. (2023). 'Transport dependency: Optimal transport based dependency measures', arXiv, arXiv:2105.02073, https://doi.org/10.48550/arXiv.2105.02073, v3:1–79, May 2023, preprint: not peer reviewed.

Pan W., Wang X., Zhang H., Zhu H., & Zhu J. (2020). Ball covariance: A generic measure of dependence in Banach space. *Journal of the American Statistical Association*, *115*(529), 307–317. https://doi.org/10.1080/01621459.2018.1543600

Pearson K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Pearson K. (1905). *On the general theory of skew correlation and non-linear regression, volume XIV of mathematical contributions to the theory of evolution, Drapers' company research memoirs*. Dulau & Co.

Peyré G., & Cuturi M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, *11*(5–6), 355–607. https://doi.org/10.1561/2200000073

Rényi A. (1959). On measures of statistical dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, *10*(3–4), 441–451. https://doi.org/10.1007/BF02024507

Reshef D. N., Reshef Y. A., Finucane H. K., Grossman S. R., McVean G., Turnbaugh P. J., Lander E. S., Mitzenmacher M., & Sabeti P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518–1524. https://doi.org/10.1126/science.1205438

Saltelli A., & Tarantola S. (2002). On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of the American Statistical Association*, *97*(459), 702–709. https://doi.org/10.1198/016214502388618447

Schwartz L. (1973). Surmartingales régulières à valeurs mesures et désintégrations régulières d'une mesure. *Journal d'analyse Mathématique*, *26*(1), 1–168. https://doi.org/10.1007/BF02790426

Sejdinovic D., Sriperumbudur B., Gretton A., & Fukumizu K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, *41*(5), 2263–2291. https://doi.org/10.1214/13-AOS1140

Shen C., Priebe C. E., & Vogelstein J. T. (2020). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*, *115*(529), 280–291. https://doi.org/10.1080/01621459.2018.1543125

Soofi E. S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association*, *89*(428), 1243–1254. https://doi.org/10.1080/01621459.1994.10476865

Steinwart I., & Christmann A. (2008). *Support vector machines*. Springer.

Strong M., & Oakley J. (2013). An efficient method for computing partial expected value of perfect information for correlated inputs. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *33*(6), 755–766. https://doi.org/10.1177/0272989X12465123

Strong M., Oakley J. E., & Chilcott J. (2012). Managing structural uncertainty in health economic decision models: A discrepancy approach. *Journal of the Royal Statistical Society, Series C*, *61*(1), 25–45. https://doi.org/10.1111/j.1467-9876.2011.01014.x

Székely G. J., & Rizzo M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, *143*(8), 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018

Székely G. J., & Rizzo M. L. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, *42*(6), 2382–2412. https://doi.org/10.1214/14-AOS1255

Székely G. J., & Rizzo M. L. (2017). The energy of data. *Annual Reviews*, *4*(1), 447–479. https://doi.org/10.1146/annurev-statistics-060116-054026

Székely G. J., Rizzo M. L., & Bakirov N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Wiesel J. C. W. (2022). Measuring association with Wasserstein distances. *Bernoulli*, *28*(4), 2816–2832. https://doi.org/10.3150/21-BEJ1438