Knowledge and Information Systems
https://doi.org/10.1007/s10115-024-02078-7

REGULAR PAPER

A Rényi-type quasimetric with random interference detection

Roy Cerqueti[1,2] · Mario Maggi[3,4]

Received: 22 June 2022 / Revised: 31 January 2024 / Accepted: 15 February 2024

Abstract
This paper introduces a new dissimilarity measure between two discrete and finite probability distributions. The followed approach is grounded jointly on mixtures of probability distributions and an optimization procedure. We discuss the clear interpretation of the constitutive elements of the measure under an information-theoretical perspective by also highlighting its connections with the Rényi divergence of infinite order. Moreover, we show how the measure describes the inefficiency in assuming that a given probability distribution coincides with a benchmark one by giving formal writing of the *random interference* between the considered probability distributions. We explore the properties of the considered tool, which are in line with those defining the concept of quasimetric—i.e. a divergence for which the triangular inequality is satisfied. As a possible usage of the introduced device, an application to rare events is illustrated. This application shows that our measure may be suitable in cases where the accuracy of the small probabilities is a relevant matter.

Keywords Inefficiency measurement · Mixture of distributions · Statistical measures · Random interference · Rare events

1 Introduction

Measuring the similarity between two probability distributions is one of the grounds of information theory, as clearly certified by the celebrated concept of entropy proposed by [27]. Shannon's entropy is a device used for stating whether a given probability distribution is closer to a uniform random variable or a degenerate random variable having mass on a

✉ Roy Cerqueti
roy.cerqueti@uniroma1.it

Mario Maggi
mario.maggi@unipv.it

1 Department of Social and Economic Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy

2 GRANEM, Université d'Angers, Angers, France

3 Department of Economics and Management, University of Pavia, 27100 Pavia, Italy

4 CAMRisk, University of Pavia, 27100 Pavia, Italy

Published online: 09 April 2024

unique point. The former case is associated with the maximum level of information of the considered random variable, while the latter is the minimum.

Nowadays, we are still witnessing the popularity of Shannon's contribution, with several information theorists following his route and working actively on concepts of similarity between probability distributions. Indeed, such instruments find applications in several information-theory contexts, like data clustering and compression, pattern recognition and signal restoring (see, e.g. [1–3, 16, 20, 30, 36]).

An important generalization of Shannon's entropy is Kullback–Leibler divergence (see [17])—also called *relative entropy*—which is an instrument able to measure the similarity between a probability distribution and a reference one. The property of assigning different roles to the considered probability distributions explains the versatility of the relative entropy when a random target is pursued. This is the case in several applied science contexts, ranging from risk-neutral measures for option pricing in finance (see, e.g. [29]), to the assessment of fluid turbulences in hydrodynamics (see, e.g. [37]), to information criteria for the selection of the states in the environment of Markov switching models (see, e.g. [28]), to the multi-agent collaboration mechanisms in the context of independent reinforce learning (see, e.g. [35]), to the problem of nonnegative matrix factorization for reducing the dimension of a dataset composed by nonnegative numbers (see, e.g. [13]) or to the challenging theme of fraudulent reviewers detection in the environment of the E-commerce platforms (see, e.g. [10]).

Interestingly, Kullback–Leibler divergence can be considered a special case of Rényi divergence (see [26]). Indeed, Rényi divergence is a family of similarity measures depending on a nonnegative parameter—the so-called *order*. It is easy to show that the Kullback–Leibler is the Rényi divergence of unitary order. The properties of the Rényi divergence are explored in depth by [34].

From a purely methodological perspective, any given similarity measure brings peculiar information on how the measured probability distributions differ. Thus, the same probability distributions can be very far when one specific measure is employed, or they can even coincide when changing the measurement device. More generally, one can say that two different metrics provide different orders of the same set of probability distributions—of course, once the law stating how the measure leads to the partial order is fixed. This explains the endless scientific debate on how to conceptualize theoretically a way to measure the difference between probability distributions (see, e.g. [5, 8, 19, 24] and, more recently, [32] and [7] we also address the reader to the excellent survey of distance/similarity measures proposed in [4]).

We notice that the divergences can be viewed as *weak* concepts of metric—or, in general, as statistical distances—since they do not satisfy the standard axiomatization of metrics. Indeed, a *divergence* is a statistical distance which is not symmetric—for a symmetric version of the Kulback-Leibler divergence, see [14]—and violates the triangular inequality. Thus, by following the arguments above, one can argue that statistical distances do not need to be metrics for playing a relevant role in applications (see the discussion on this in [9]).

This paper enters this debate. It proposes a novel concept of similarity measure between discrete and finite probability distributions. We consider a benchmark (target) probability distribution and a to-be-measured one. The measure is based jointly on an optimization procedure and a mixture decomposition of the considered probability distributions, with the intervention of an additional random quantity. In particular, such a measure is the optimized coefficient of a convex combination between the reference probability distribution and the newly introduced one, where the combination is imposed to be equal to the to-be-measured distribution.

As we will see, the proposed dissimilarity measure is not a metric in that it violates the symmetry property. Specifically, it is a *quasimetric*, i.e. a divergence for which the triangular inequality holds true. It is worth noticing that also the Rényi divergence or order infinity—which is the Rényi divergence as its order goes to infinity—fulfils the triangular inequality (e.g. see [22]), being then a quasimetric. In this respect, we show that our proposed measure is related to the Rényi divergence of infinite order in that it can be obtained through a monotone transformation of it. As a consequence, we here offer a new dissimilarity measure producing the same ordering of the Rényi divergence of infinite order.

This said the proposed dissimilarity measure brings some relevant innovations with respect to its Rényi-type counterpart. First, it is bounded. This property allows us to state how the distance between two distributions is close to its theoretical maximal or minimal level. In this respect, we notice that Rényi divergence explodes in some cases related to the presence of null probabilities, making it inadequate to explore situations with rare events. Second, it has a simple interpretation in terms of mixtures. This property provides a clear view of the discrepancy between the considered variables. Third—in line with the mixture-based interpretation presented above—the computation of the proposed measure leads to the identification of a third distribution representing the random gap between the considered probability distributions. Such a gap has an important informative content in that it might be seen as the random interference—the one with the minimum amount of disturbance—between the measured probability distributions. In so doing, we also contribute to the theme of identifying the efficiency gap in information-theoretical contexts (see, e.g. [33] and, more recently, [25]). This opportunity is not given in the case of Rényi divergence of infinite order.

Furthermore, we offer here an original proof that our dissimilarity measure satisfies the triangular inequality based on its geometric representation.

In line with the arguments above, the methodological proposal is tested over the paradigmatic case of rare events. In doing so, we provide a proper illustration of the constitutive terms compounding the dissimilarity measure, along with a description of the random interference. Noticeably, we show that the considered dissimilarity measure penalizes the underestimation of the probabilities of the rare events.

To further illustrate the features of the proposed measure, we show an application to the binomial and normal distributions. This application also highlights some possible numerical drawbacks produced, even in simple cases, by other divergences with unbounded values, such as the Kullback–Leibler.

Finally, we present an empirical application of the proposed methodological device. Specifically, we deal with financial data related to the Standard & Poor Index of the New York Stock Exchange. We model the distribution of the returns by using a normal and a generalized Student's t distribution. The outcomes of the empirical experiments certify that the considered dissimilarity measure is particularly appropriate for exploring real-world instances; moreover, its properties lead to an intuitive interpretation of the considered phenomenon.

In addition, we show a preliminary extension of the dissimilarity measure to the case of continuous probability density functions, providing an example for Gaussian distributions.

The rest of the paper is organized as follows. Section 2 contains the definition of the dissimilarity measure here introduced. Section 3 outlines and discusses the properties of the dissimilarity measure, along with the original proof of the triangular inequality. Section 4 is devoted to the application of the proposed methodology to the case of extreme events, to the binomial and normal distributions. Section 5 presents an introduction to the application of the MDM notion to continuous distributions and provides an example in the case of Gaussian distributions. The last section offers some conclusive remarks.

2 Definition of the dissimilarity measure

We consider the set of discrete probabilities on n possible outcomes and identify a given distribution with the corresponding probability vector $p = [p_1, p_2, \dots, p_n] \in \mathbb{R}^n$, with $p \geq [0]$ and $\sum_{i=1}^n p_i = 1$, being $[0]$ the null (n-dimensional) vector and the inequality symbol $\geq$ has to be intended in a component-wise sense, so that $p \geq [0]$ is equivalent to $p_i \geq 0$, for each $i = 1, \dots, n$. We collect such probability vectors in the set $\mathcal{P}_n$.

From the geometric point of view, p is a point on the unit $(n-1)$-dimensional simplex in $\mathbb{R}^n$. Consider a benchmark discrete probability distribution, with probability vector $\delta \in \mathcal{P}_n$, and another generic probability distribution X, with probability vector $\delta^X \in \mathcal{P}_n$. It is always possible to write δ^X as the mixture—which here means convex combination—between δ and another suitably defined probability vector δ^Y, as follows:

$$\delta^X = \alpha\delta^Y + (1-\alpha)\delta, \quad \alpha \in [0, 1]. \tag{1}$$

In writing (1), we implicitly assume that δ, δ^X and δ^Y are considered on the same n outcomes. This is not restrictive at all, in that we can consider for all of them the union of the sets of their possible outcomes, hence possibly obtaining some cases of null components in the probability vectors. As we will see below, it is not a problem in our framework.[1]

Definition 2.1 Let δ^X and δ be two discrete probability distributions in $\mathcal{P}_n$, being δ the benchmark distribution and δ^X the investigated distribution. The *Mixture Dissimilarity Measure* (MDM) between δ^X and δ is the smallest $\alpha \in [0, 1]$—namely, $\alpha^* = M(\delta^X, \delta)$—such that there exists a probability distribution $\delta^Y(\alpha^*)$ satisfying (1).

The distribution $\delta^Y(\alpha^*)$ can be defined as the *random interference* of the distribution δ^X with respect to the benchmark δ.

In the light of the mixture formulation (1), we think Definition 2.1 deserves a geometric interpretation in $\mathbb{R}^n$. In (1), α and δ^Y may be considered two indicators of how δ^X is similar to the benchmark δ: roughly speaking, α is the "size" of the difformity—so that, the difformity between δ^X and δ decreases as α decreases—and δ^Y its "shape" or "direction".

For every δ^X and $\delta > [0]$, there exist an infinite number of possible choices of α leading to a $\delta^Y(\alpha) \in \mathcal{P}_n$ such that condition (1) is true. Loosely speaking, the possible mixtures range from a large weight α—which implies the corresponding selection of a δ^Y whose components are rather similar to those of δ^X—to a small α – corresponding to a δ^Y far from δ^X. It is clear that the corner case $\alpha = 1$—and, consequently, $\delta^Y = \delta^X$—is not interesting at all. A way to make the decomposition (1) as informative as possible is to solve the following:

Problem 2.2 *Given $\delta, \delta^X \in \mathcal{P}_n$, find the distribution $\delta^Y \in \mathcal{P}_n$ with the largest Euclidean distance from δ^X and such that* (1) *is true.*

Problem 2.2, which is directly related to Definition 2.1, may shade more light on the meaning of MDM, and allows to find an easy way to compute $\alpha^* = MDM(\delta^X, \delta)$ and $\delta^Y(\alpha^*)$.

In the case $\delta = \delta^X$, Problem 2.2 has the obvious solution $\alpha^* = 0$, for any $\delta^Y \in \mathcal{P}_n$.

To solve Problem 2.2, for $\alpha > 0$ rewrite (1) in the form

$$\delta^Y = \frac{1}{\alpha}\left[\delta^X - (1-\alpha)\delta\right], \quad \alpha \in (0, 1], \tag{2}$$

[1] To simplify the notation, some dependencies can be omitted. In relation (1), α and δ^Y depend on δ and δ^X—such dependence is omitted in the employed notation; moreover, they are not independent quantities. In particular, the selection of α leads to the consequent identification of the δ^Y for which (1) is satisfied. In this respect, we will state the dependence of δ^Y on α by denoting $\delta^Y = \delta^Y(\alpha)$, when needed.

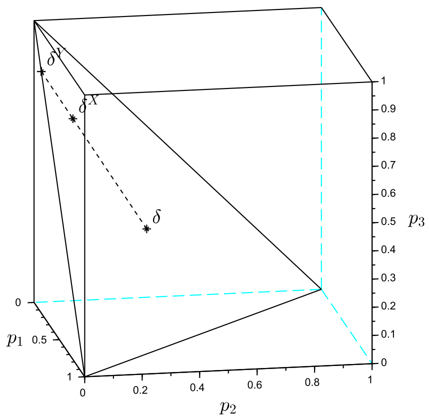

Fig. 1 Two examples of graphical representation of the mixture decomposition of δ^X; case $n = 3$. Left: the benchmark distribution is set to $\delta = [1/3, 1/3, 1/3]$, the other considered distribution is $\delta^X = [0.2, 0.1, 0.7]$; the value of α is set to 0.7 and, consequently, the δ^Y for which (2) is satisfied is $\delta^Y(0.7) \approx [0.1429, 0, 0.8571]$. Right: $\delta = [0.5, 0.3, 0.2]$, $\delta^X = [0.3, 0.6, 0.1]$, $\alpha = 0.5$, and the δ^Y leading to the validity of (2) is $\delta^Y(0.5) = [0.1, 0.9, 0]$

and notice that, from the geometric point of view in $\mathbb{R}^n$, δ^Y lies on the line through δ and δ^X, in the half line with origin in δ^X and non-containing δ. Figure 1 illustrates a graphical representation of the mixture decomposition by showing two examples in the case $n = 3$.

Formally, Problem 2.2 can be written as the constrained optimization problem

$$\begin{aligned} &\min_{\alpha} \ \alpha \\ &\text{s.t.} \ \ \tfrac{1}{\alpha}\left[\delta^X - (1-\alpha)\delta\right] \geq [0] \\ &\qquad \alpha \in [0, 1] \end{aligned} \tag{3}$$

whose solution is

$$\alpha^* = 1 - \min_{\delta_i > 0} \frac{\delta_i^x}{\delta_i}. \tag{4}$$

Clearly, the random interference $\delta^Y(\alpha^*)$— obtained by (2)—lays on the boundary of the unit simplex, so at least one of its components is null. See Fig. 1 for an illustration.

The optimized value α^* is the inefficiency measure of assuming that δ^X coincides with the benchmark δ. Its variation range is [0, 1], and inefficiency increases as α^* does. The corner cases $\alpha^* = 0$ and $\alpha^* = 1$ represent the situations where one has the minimum and maximum level of inefficiency, respectively. The interference $\delta^Y(\alpha^*)$ is a random adjustment which can be seen as the inefficiency gap in assuming that δ^X and δ are the same. Plugging α^* and $\delta^Y(\alpha^*)$ into (1), the probability distribution δ^X is obtained as a *contaminated* benchmark δ, where α^* describes the percentage of contamination while δ^Y is the random term which is responsible for such a contamination.

To illustrate the relations between the mixture distance $M(\delta^X, \delta)$ and the distributions, we provide a graphical analysis of the two examples considered in Fig. 1. Figure 2 shows the values of α^* for distributions δ^X which have the same Euclidean distance— equal to 0.1, in the represented cases—from δ on $\mathbb{R}^3$. Figure 3 shows the level sets of $M(\delta^X, \delta)$, projected on the plane of the first two components (probabilities). We remark how the mixture distance depends on the relative location of δ^X and δ with respect to the simplex boundaries, with a steeper increase along the boundary which is closest to δ. In other words, let i be the index

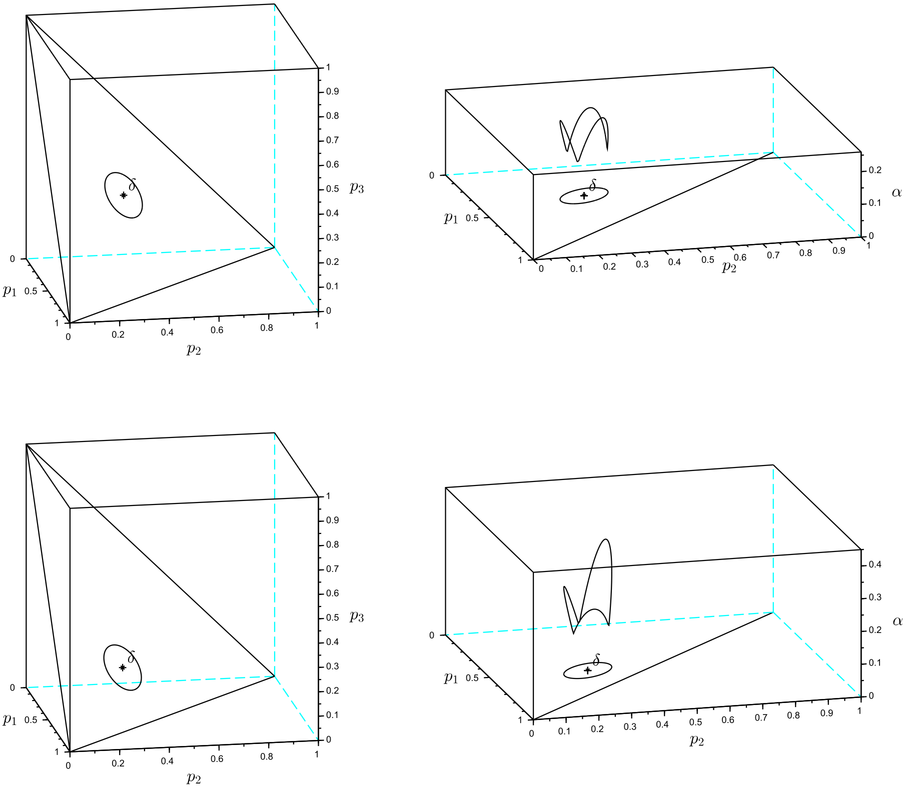

Fig. 2 mdm α^* for distributions δ^X which have Euclidean distance from δ equal to 0.1, with $\delta = [1/3, 1/3, 1/3]$ on the top, and $\delta = [0.5, 0.3, 0.2]$ on the bottom. On the left, is a graphical representation of the simplex; on the right, is the value for α^* relative to the projection of the circle on the horizontal plane

of the minimum component of δ, the steepest increase in $M(\delta^X, \delta)$ is produced by reducing δ_i^X below δ_i.

3 Properties of the dissimilarity measure

We now present the main properties of the MDM in Definition 2.1.

First of all, we state that the connection between the MDM and the Rényi divergence. The proof is immediate, thanks to formula (4).

Property 3.1 *Given $\delta^X, \delta \in \mathcal{P}_n$, then $M(\delta^X, \delta) = 1 - \exp\left\{-d_{+\infty}(\delta, \delta^X)\right\}$, where $d_{+\infty}$ is the Rényi divergence of infinite order, defined as*

$$d_{+\infty}(\delta, \delta^X) = \sup_{\delta_i > 0} \log \frac{\delta_i}{\delta_i^X},$$

with the conventional agreement that $\delta_i^X = 0$ is associated to $\log(+\infty) = +\infty$ and $\exp\{-\infty\} = 0$, while $\delta_i = 0$ gives $\log(0) = -\infty$.

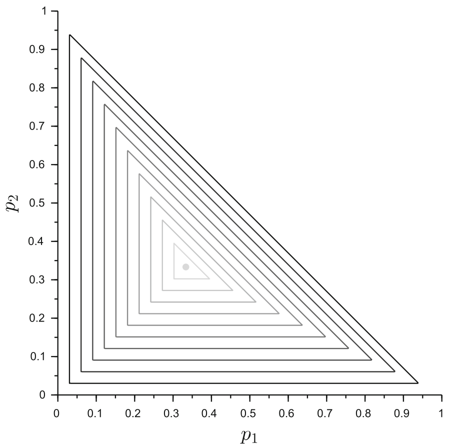

Fig. 3 Level sets, projected on the plane (p_2, p_1), of the MDM $\alpha^* = M(\delta^X, \delta)$ with respect to $\delta = [1/3, 1/3, 1/3]$ (left) and $\delta = [0.5, 0.3, 0.2]$ (right). The dot indicates the reference measure

Property 3.1 implies that the MDM and $d_{+\infty}$ induce the same ordering on the set $\mathcal{P}_n$, for any target distribution δ. Moreover, the properties of the Rényi divergence of infinite order can be easily addressed to the MDM. The following proposition lists some relevant properties of MDM.

Proposition 3.2 *The MDM is endowed with the following properties:*

1. $M(\delta^X, \delta) \in [0, 1]$, *for each* $\delta^X, \delta \in \mathcal{P}_n$.
2. *Consider* $\delta^X, \delta \in \mathcal{P}_n$. *Then* $M(\delta^X, \delta) = 0 \iff \delta^X = \delta$.
3. *There exist two distributions* δ^X *and* δ *in* $\mathcal{P}_n$, *such that* $M(\delta^X, \delta) \neq M(\delta, \delta^X)$.
4. *For every pair of distributions* $\delta, \delta^X \in \mathcal{P}_n$, *then* $M(\delta^X, \delta)$ *is unique.*
5. $M(\delta^X, \delta) = 1 \Rightarrow \delta^X \not> [0]$. *Moreover, if* $\delta > [0]$, *then:* $\delta^X \not> [0] \Rightarrow M(\delta^X, \delta) = 1$.

Proof The proof proceeds in pointwise form.

1. This result follows from Definition 2.1, and a fortiori from the notion of mixture distribution (e.g. see [21]).
2. In fact, thanks to (1), $M(\delta^X, \delta) = 0 \Rightarrow \delta^X = \delta$, and from Definition 2.1 it is easy to obtain $\delta^X = \delta \Rightarrow M(\delta^X, \delta) = 0$.
3. It is simple to find a pair of distributions for which the MDM is not symmetric. For instance, let us consider the particular case with δ the uniform distribution in $\mathcal{P}_n$ (i.e. $\delta_i = 1/n, i = 1, \dots n$), and δ^X, such that $\delta_1^X = \frac{1}{n} + \varepsilon, \delta_2^X = \frac{1}{n} - \varepsilon, \delta_i^X = \frac{1}{n}, i > 2$, with $\varepsilon \in \left(0, \frac{1}{n}\right)$. In this case, thanks to (4), $M(\delta^X, \delta) = n\varepsilon \neq M(\delta, \delta^X) = \frac{n\varepsilon}{1+n\varepsilon}$.
4. This property is evident from (4).
5. If $M(\delta^X, \delta) = 1$, then formula (1) implies $\delta^X = \delta^Y(1)$, and we know that $\delta^Y(1)$ has at least one null component; if $\delta^X \not> [0]$, i.e. $\delta_i^X = 0$ for some i, then the hypothesis $\delta > [0]$ implies that $\max_i \left\{1 - \frac{\delta_i^X}{\delta_i}\right\} = 1$, so that and the constraints of problem (3) boil down to $\alpha^* = 1$. Notice that the assumption that $\delta > [0]$ can be relaxed by stating that there exists $i = 1, \dots, n$ such that $\delta_i^X = 0$ and $\delta_i > 0$.

$\square$

Of course, the MDM also satisfies the triangular inequality, being an increasing concave transformation of $d_{+\infty}$.[2] However, we find it worth providing an original proof of such a property by following a geometric approach—on the ground of the formulation of the MDM in Definition 2.1. To this aim, some technical preliminaries are needed.

The first technical result follows immediately from a simple geometric argument. We enunciate it.

Lemma 3.3 *Consider $\delta, \delta^X \in \mathcal{P}_n$. Assume that $M(\delta^X, \delta) = \alpha^*$, with random interference $\delta^Y(\alpha^*)$. Then:*

$$\alpha^* = \frac{\|\delta^X - \delta\|}{\|\delta^Y(\alpha^*) - \delta\|}, \tag{5}$$

where $\|\cdot\|$ indicates the Euclidean norm.

Lemma 3.3 is useful for checking our second technical statement.

Lemma 3.4 *Let us consider $\delta^A, \delta^B, \delta^C \in \mathcal{P}_n$. If $\exists \beta \in [0, 1]$ such that*

$$\delta^B = \beta\delta^A + (1-\beta)\delta^C, \tag{6}$$

then

$$M(\delta^A, \delta^C) \leq M(\delta^A, \delta^B) + M(\delta^B, \delta^C), \tag{7}$$

and the three corresponding interference terms are equal: $\delta^{Y_{AC}} = \delta^{Y_{AB}} = \delta^{Y_{BC}}$

Proof Since the three vectors $\delta^A, \delta^B, \delta^C$ are aligned, and δ^B is between the other two vectors, a simple application of Definition 2.1 leads to $\delta^{Y_{AC}} = \delta^{Y_{AB}} = \delta^{Y_{BC}}$; let us indicate this common interference vector as δ^Y. Therefore, the four vectors $\delta^Y, \delta^A, \delta^B, \delta^C$ are aligned, following this order, on the same segment on the unit simplex. Consequently, thanks to (5) in Lemma 3.3, we obtain

$$M(\delta^A, \delta^C) = \frac{\|\delta^A - \delta^C\|}{\|\delta^Y - \delta^C\|},\ M(\delta^A, \delta^B) = \frac{\|\delta^A - \delta^B\|}{\|\delta^Y - \delta^B\|},\ M(\delta^B, \delta^C) = \frac{\|\delta^B - \delta^C\|}{\|\delta^Y - \delta^C\|}.$$

We remember also that, thanks to the alignment of the vectors

$$\|\delta^A - \delta^C\| = \|\delta^A - \delta^B\| + \|\delta^B - \delta^C\|, \quad \text{and} \quad \|\delta^Y - \delta^C\| \geq \|\delta^Y - \delta^C\|.$$

It is now straightforward to obtain the thesis, in fact

$$M(\delta^A, \delta^C) = \frac{\|\delta^A - \delta^C\|}{\|\delta^Y - \delta^C\|} = \frac{\|\delta^A - \delta^B\|}{\|\delta^Y - \delta^C\|} + \frac{\|\delta^B - \delta^C\|}{\|\delta^Y - \delta^C\|}$$

[2] From Property 3.1, $M(\delta^X, \delta)$ is an increasing and concave transformation of $d_{+\infty}(\delta, \delta^X)$. Therefore, given the probability distributions $\delta^A, \delta^B, \delta^C \in \mathcal{P}_n$, the following holds

$$d_{+\infty}(\delta^C, \delta^A) \leq d_{+\infty}(\delta^B, \delta^A) + d_{+\infty}(\delta^C, \delta^B).$$

Let $f : \mathbb{R} \to \mathbb{R}$ increasing and convex, then

$$\begin{aligned} f\left(d_{+\infty}(\delta^C, \delta^A)\right) &\leq f\left(d_{+\infty}(\delta^B, \delta^A) + d_{+\infty}(\delta^C, \delta^B)\right) \qquad (f \textit{ increasing}) \\ &\leq f\left(d_{+\infty}(\delta^B, \delta^A)\right) + f\left(d_{+\infty}(\delta^C, \delta^B)\right) \qquad (f \textit{ concave}). \end{aligned}$$

□

$$\leq \frac{\|\delta^A - \delta^B\|}{\|\delta^Y - \delta^B\|} + \frac{\|\delta^B - \delta^C\|}{\|\delta^Y - \delta^C\|} = M(\delta^A, \delta^B) + M(\delta^B, \delta^C).$$

□

We are now in the position of checking the triangular inequality for the MDM.

Property 3.5 *(Triangular inequality) For every choice of three vectors of probabilities $\delta^A, \delta^B, \delta^C \in \mathcal{P}_n$, the following holds*

$$M(\delta^A, \delta^C) \leq M(\delta^A, \delta^B) + M(\delta^B, \delta^C). \tag{8}$$

Proof To prove this result, we need a premise for the graphical reasoning proposed below. If $M(\delta^B, \delta^C) \geq M(\delta^A, \delta^C)$, the inequality (8) is trivially verified. Otherwise, if the two random interferences $\delta^{Y^{AC}}$ and $\delta^{Y^{AB}}$ associated to $M(\delta^A, \delta^C)$ and $M(\delta^A, \delta^B)$ lie on the same facet of the unit simplex boundary, then the comparison between the three probabilities is equivalent to the comparison of $\delta^A, \delta^{B'}, \delta^C$, where $\delta^{B'}$ is the vector belonging to both *(i)* the level set $\{\delta \in \mathcal{P}_n : M(\delta, \delta^C) = M(\delta^B, \delta^C)\}$, and *(ii)* the segment from δ^C and the random interference $\delta^{Y^{AC}}$ associated to $M(\delta^A, \delta^C)$. The equivalence follows from the fact that the level sets of $M(\delta, \delta^B)$ and $M(\delta, \delta^C)$ are—in the relevant region of the simplex—given by portions of parallel hyperplanes. In this case, being $\delta^A, \delta^{B'}, \delta^C$ aligned, the triangular inequality (8) holds, thanks to Lemma 3.4.

For the graphical proof, remark that we can restrict the analysis to the region of the domain of the MDM defined as the intersection $\mathcal{T}$ between the probability simplex and the plane trough $\delta^A, \delta^B, \delta^C$. This is justified by the fact that the random interferences are vectors aligned with the two arguments of the MDM so that all the vectors $\delta^A, \delta^B, \delta^C, \delta^{Y_{AC}}, \delta^{Y_{BC}}, \delta^{Y_{AB}}$ lie on the same plane. We point out that the intersection $\mathcal{T}$ is a polygon whose sides lie on the boundaries of the simplex. Moreover, the level sets of the MDM intersections with $\mathcal{T}$ are similar polygons where each side is parallel to the corresponding boundary of $\mathcal{T}$.

Consider a probability δ^C and let δ^A be any probability vector on a given level set of $M(\delta, \delta^C)$. The intersection $\mathcal{L}$ of this level set with $\mathcal{T}$ is the edge of a polygon. Figure 4a represents a case where the intersection is a triangle, but the same arguments apply to any polygon shape. If δ^B belongs to the shaded region of Fig. 4a, we have $M(\delta^B, \delta^C) \geq M(\delta^A, \delta^C)$; therefore, the inequality (8) is verified.

Consider the case in which δ^B belongs to the interior of the polygon delimited by the considered level set. As indicated in Fig. 4b, let us consider the (only) level set of $M(\delta, \delta^B)$ that has an entire side in common with $\mathcal{L}$. Along this side (the thicker one in Fig. 4b), the triangular inequality (8) holds, thanks to the reasoning put forward in the premise at the beginning of this proof.

Therefore, if δ^A belongs to this (thicker) side

$$M(\delta^A, \delta^C) \leq M(\delta^A, \delta^B) + M(\delta^B, \delta^C). \tag{9}$$

Given δ^B and δ^C, for all $\delta^A \in \mathcal{L}$ both $M(\delta^A, \delta^C)$ and $M(\delta^B, \delta^C)$ are constant. Moreover $M(\delta^A, \delta^B)$ reaches its minimum in the thicker region, lying the remaining part of $\mathcal{L}$ on higher level sets of $M(\delta, \delta^B)$, increasing the right-hand side of (9), which continues to be verified. The arbitrariness of $\delta^A, \delta^B, \delta^C$ completes the proof. □

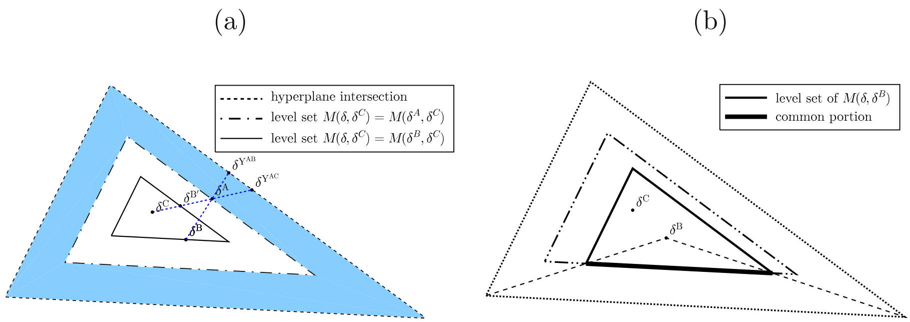

Fig. 4 Graphical proof of Property 3.5. The dashed-dotted line is the level set of the MDM with respect to δ^C such that $M(\delta, \delta^C) = M(\delta^A, \delta^C)$. The solid and thick lines on the right are the level set of the MDM with respect to δ^C; the thicker portion is common to the two-level sets shown

4 Applications

This section proposes some applications of the SMS to some simple cases: rare events and common distributions.

4.1 Application to rare events

Rare events, i.e. events with low probability (or which occur with a low frequency), are of great interest in many scientific areas. In fact, rare events are often also extreme in size. Therefore, their effects can be relevant. For instance, some areas of interest are seismology, epidemiology, economics and finance. In these fields, a measure of similarity may provide a tool to evaluate the accuracy loss derived from the use of a given probability δ^X, instead of the true one δ. The probability vector δ^X is commonly obtained by estimation. When rare events are a matter of concern, like earthquakes or financial crashes, the MDM may be suitable since it provides a measure that focuses on small probabilities: the MDM assigns a larger value when a small probability is approximated by a smaller probability than a larger one. Therefore, the underestimation of rare events probabilities is more severely penalized than the overestimation.

To illustrate this behaviour, we compare the non-conformity between distributions measured by the MDM, the mean absolute deviation (MAD: $\mathrm{MAD}(\delta, \delta^X) = \frac{1}{n}\sum_{i=1}^{n} |\delta - i - \delta_i^X|$) and the Kullback–Leibler divergence (KL: $\mathrm{KL}(\delta, \delta^X) = D_{KL}\left(\delta \| \delta^X\right) = \sum_{i=1}^{n} \delta_i \log_2 \frac{\delta_i}{\delta_i^X}$). We consider, as reference probability δ, the probabilities assigned by a Student's t distribution with $\nu = 7$ degrees of freedom to 30 bins: 28 equally spaced between -15 and 15, and two collecting the remaining probability: $(-\infty, -15]$ and $(15, +\infty)$. In this case, the rare events are those in the tails of the distribution. As the degrees of freedom parameter ν decreases, the probabilities of tail events increase (see Fig. 5 for an example with $\nu = 3, 7, 11$). The distribution δ with seven degrees of freedom is then compared with the distribution δ^X with ν degrees of freedom. Therefore, when $\nu < 7$ the distribution δ^X overestimates the rare events probabilities, whereas when $\nu > 7$, δ^X underestimates them. The choice of $\nu = 7$ as reference case allows to compare δ with both over- and underestimating rare events probabilities, by setting $\nu = 3, 4 \ldots, 11$ for δ^X.

From Fig. 6, it is clear that all the measures increase as ν gets far from 7, but the behaviour is different for the three measures. First of all, we notice that all measures display asymmetric

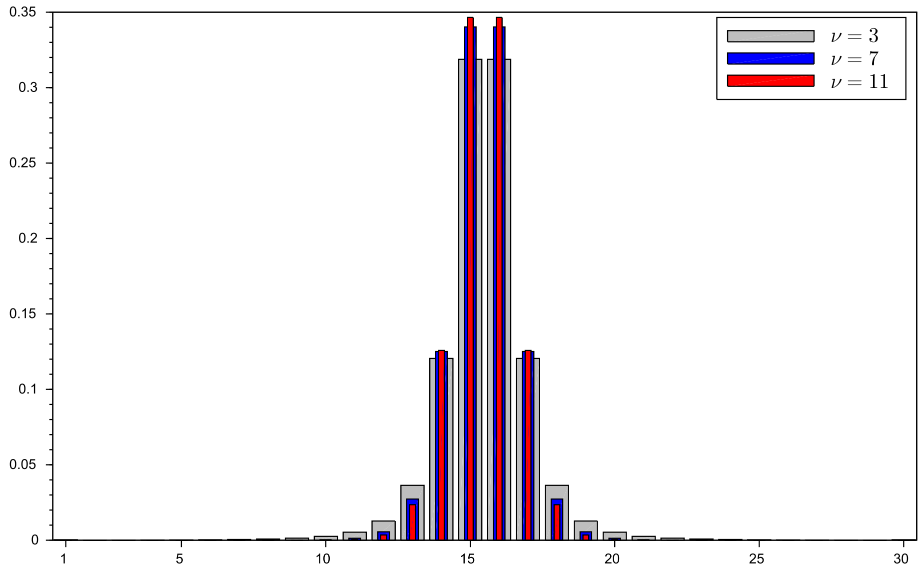

Fig. 5 Histogram of Student's t histograms for different degrees of freedom. The described discretization is applied. The horizontal scale simply indicates the order of the bin, rather than the value of the random variable

behaviours. KL is smooth, and the MDM and MAD are not. Considering the effect of rare event probabilities, the main difference between the measures is that MAD and KL penalize the overestimation of the rare event probabilities, whereas the MDM behaves oppositely: underestimating small probabilities rapidly drives the MDM close to its maximum value. This feature may be desirable in the study of rare, extreme and catastrophic events.

In addition, Fig. 7 shows the resulting interference random term δ^Y related to the MDM in the various cases. A possible interpretation is that when $\nu < 7$, i.e. the rare event probabilities are overestimated, the sample distribution δ^X is obtained by sampling a fraction α of times from a distribution assigning larger probabilities on the "sides" of the distributions. In this case, the fraction α is small; in fact, from Fig. 6 we can see that $\alpha < 0.1$. Instead, for $\nu > 7$, i.e. in case of underestimation of small probabilities, δ^X is obtained sampling most of the times—the fraction α is larger than 0.5 and gets quickly close to 1—from a distribution concentrated around the mean. It is, therefore, possible to conclude that in the latter case, our sample is highly contaminated, and so the MDM assigns a large dissimilarity measure.

4.2 Binomial and Gaussian examples

In this section, we show the results of the application of the MDM to two common distributions: the binomial and the Gaussian.

Consider the binomial distribution with parameters $\pi \in [0, 1]$ and $m \in \mathbb{N}^+$. Let $\delta \in \mathcal{P}_{m+1}$ be the reference distribution of the binomial with parameters $\pi = \frac{1}{2}$, $m = 10$. Consider the binomial distributions δ^X with parameters $\pi = 0.1, 0.2, \ldots 0.9$ and $m = 10$. The MDM rapidly increases, being around 0.89 for $\pi = 0.4$ and $\pi = 0.6$, and further increasing for π farther from 0.5. We remark that when π is close to 0 or 1, some binomial probabilities are very close to 0, producing numerical concerns for some unbounded divergences, such as the Kullback–Leibler, that become extremely large or cause numerical overflows.

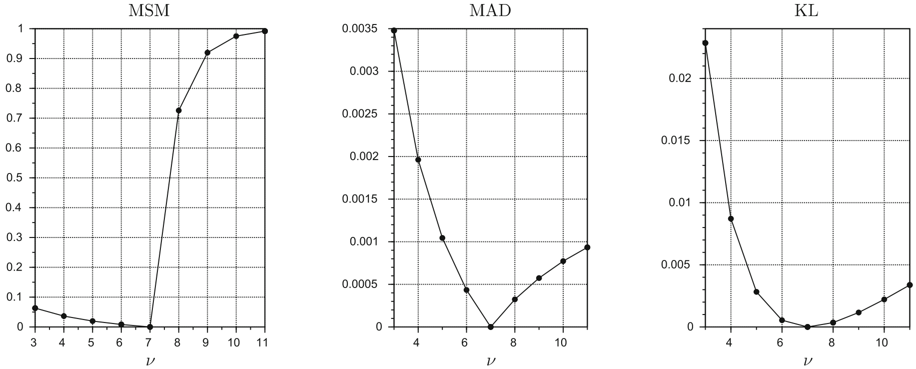

Fig. 6 Values of the statistical distances MDM, MAD and KL between a probability δ^X obtained discretizing a Student's t with ν degrees of freedom, and the reference δ, which is a Student's t with $\nu = 7$

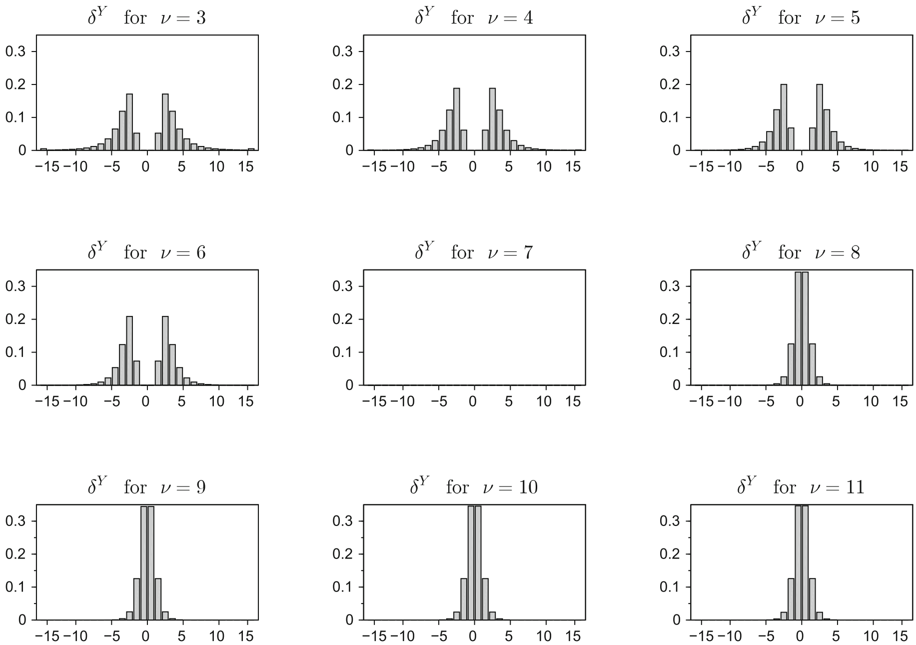

Fig. 7 Histograms of the interference distribution δ^Y in the case of comparison between Student's t distributions, with different degrees of freedom for δ^X

The interference term δ^Y is shown in Fig. 8. This term essentially follows and compensates for the asymmetry produced by the parameter π.

Consider the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Let $\delta \in \mathcal{P}_n$ be the reference distribution obtained discretizing the standard normal distributions (i.e. $\mu = 0, \sigma = 1$) in n bins, symmetric around 0. Consider δ^X, the discretization of the normal distribution with parameters $\mu = 0$ and $\sigma \in \{0.2, 0.4, 0.75, 0.9, 1, 1.5, 2, 3, 4\}$. We notice that the MDM rapidly increases towards 1 for $\sigma < 1$, reproducing the same phenomenon discussed in Sect. 4.1 concerning the rare events. We also highlight, in this case, the possible numerical overflows produced by unbounded divergences in the case of small σ. In fact, with

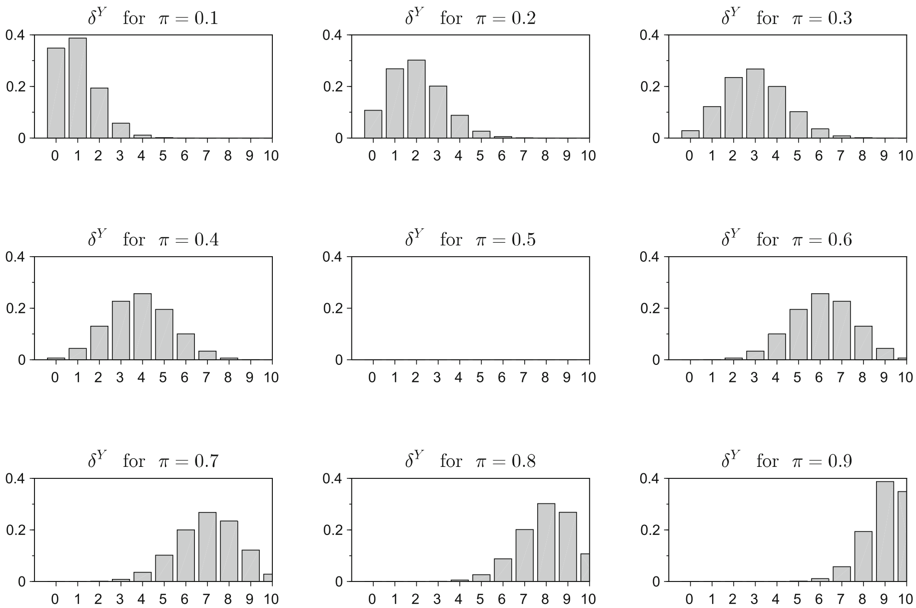

Fig. 8 Histograms of the interference distribution δ^Y in the case of comparison between binomial (π, m) distributions, with $m = 10$, and different values of π

a small σ the extreme bins have infinitesimal probabilities. The interference term δ^Y is shown in Fig. 9, with a behaviour similar to the one shown in Fig. 7, with a reversed order. Also in this case when small tail probabilities are underestimated, in addition to a large MDM, the random gap term δ^Y is concentrated around 0.

4.3 Empirical application

To show a possible way to benefit from the behaviour of MDM in case of rare events, we propose a financial application. It is a stylized fact that financial return distributions display heavy tails, i.e. the frequency of extreme returns is larger than what a normal distribution accounts for. This fact is reflected by the common use of risk measures that explicitly consider the probability of extreme events. Among the risk measures, the Value at Risk (VaR) can be considered a benchmark (see [15]). The VaR is the negative of the percentile of the return distribution, computed for a low probability level, usually 5%, or 1%. In other words, the $\text{VaR}_{5\%}$ is the level of loss that can be exceeded with a probability equal to 5%. As such, it is a risk measure that focuses on the possible losses of a financial asset.

Our application considers the Standard & Poor Index (SPX) of the New York Stock Exchange. Following common practices in this field, we model the return distribution by a normal or a generalized Student's t distribution (e.g. see [6, 12, 23]).[3] Then, we estimate the distribution parameters through maximum likelihood (ML), minimum KL divergence

[3] Other models can be selected to take into account the skewness and other features of the returns distribution. For instance, [11, 18, 31] consider the generalized non-central skew-t distribution as a more flexible model to describe financial variables. However, the considered probabilistic assumptions are in agreement with the mentioned relevant literature contributions and allow a clear and convincing empirical representation of the methodological proposal.

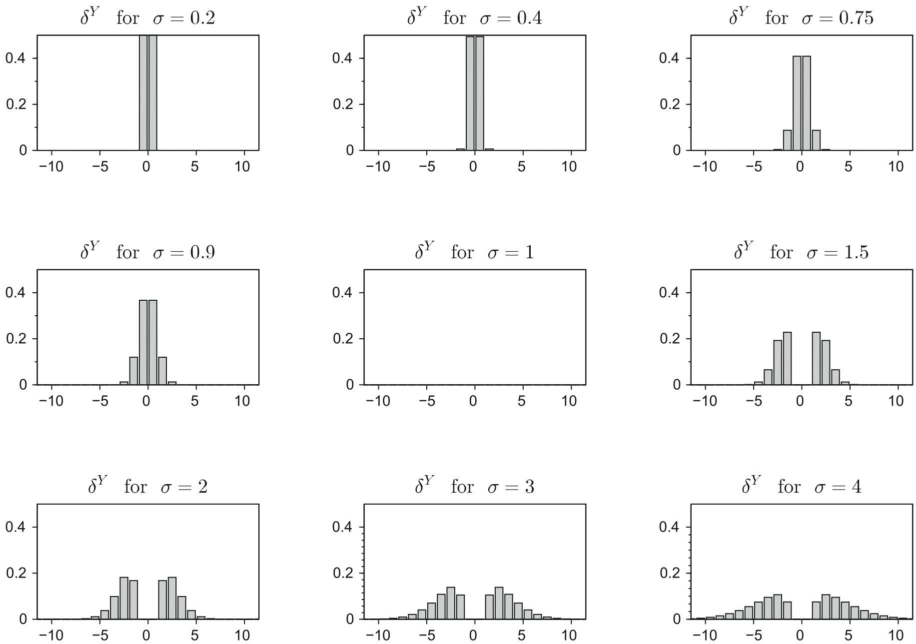

Fig. 9 Histograms of the interference distribution δ^Y in the case of comparison between normal $(0, \sigma)$ distributions, with different values of σ

and min MDM.[4] Finally, we compare the historical VaR at the 5% level of the empirical return distribution with the ones implied by the estimated distribution. We consider SPX daily returns from April 2002 to April 2022 to take into account different market conditions, such as expansion, stability and crisis. In this way, the return distribution displays the usual stylized features: small mean (0.0002721), relatively high standard deviation (0.0122819), negative skewness (−0.4480573) and high kurtosis (15.332965). See Fig. 10 for a graphical representation of the empirical distribution. To make a more robust analysis of the features of the application of the indicated estimation methods, we resample 1000 simulated samples of the same size as the data (5035 observations), from the empirical distribution of the returns. On each simulated sample, we estimate the normal and the Student's t models with the three methods: ML, min KL and min MDM. First of all, we highlight that the min KL method numerically stucks 295 (i.e. about 30%) times. Figure 11 shows the parameter estimate distributions. The ones relative to the KL case are computed only on the successful (about 70%) cases; the consequence is that the area under the displayed densities is 0.7, instead of 1, leaving the out-of-scale values out of the plots. From the plots, we may conclude that all the methods roughly agree on the mean of the normal (μ), and on the degrees of freedom of the Student's t (ν). Instead, the MDM overestimates the normal standard deviation (σ) and the t scale parameter (s), probably to better account for the extreme event frequency. Besides, we compute the VaR implied by the estimated distribution and compare them with the sample realization. This allows computing the errors whose distributions are shown in Fig. 11. In general, it is possible to conclude that the MDM minimization produces less

[4] For KL and MDM, the distribution has been discretized in 25 equally spaced bins from the first and the 99th percentile, plus two bins for the extreme percentiles.
The optimization is performed by a genetic algorithm to reduce the multiple local maxima issue.

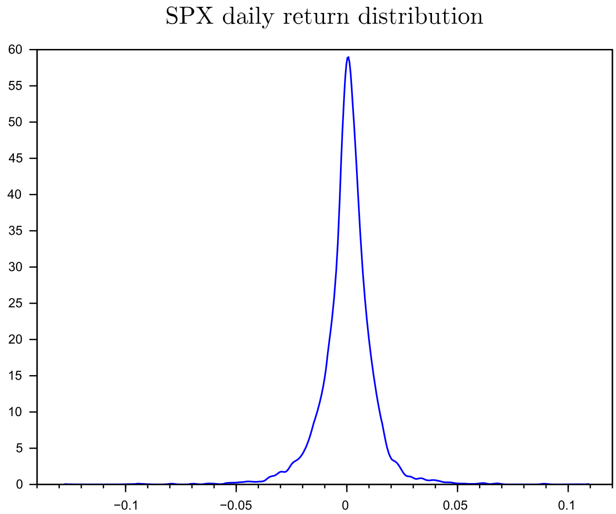

Fig. 10 Gaussian kernel smoothing of the distribution of the SPX daily returns, from April 2002 to April 2022

precise estimates because the variances are larger than the other methods. However, it is worth noting that numerical convergence problems never occur. The MDM method produces the best results in two out of four cases: normal distribution, $\text{VaR}_{5\%}$; Student's t distribution, $\text{VaR}_{1\%}$. In the normal distribution, $\text{VaR}_{1\%}$ case, the MDM is in between the others, whereas in the Student's t distribution, $\text{VaR}_{5\%}$ it is the worst.

The results obtained by this non-exhaustive application may lead to the conclusion that the MDM can be a suitable tool for studying real datasets because its performances align with other methods. The main advantage with respect to KL divergence is the numerical stability which follows from the bounded range of the MDM. Moreover, the MDM produces a second outcome: the interference distribution that may help interpret the phenomenon at hand.

5 Some remarks on the continuous case

This paper deals with discrete and finite probability distributions. An interesting avenue of future research could concern the extension of the results to numerable or continuous probability spaces. In this respect, we present here some initial remarks concerning the case where $\mathcal{P}$ is the set of probability distributions endowed with density. Specifically, consider a benchmark probability density function $\delta(s)$, and another probability density function $\delta^X(s)$, with $s \in \mathbb{R}$. Following the same reasoning put forward in Sect. 2, it is always possible to write $\delta^X(s)$ as the mixture between $\delta(s)$ and another suitably defined probability density function $\delta^Y(s)$, as follows:

$$\delta^X(s) = \alpha\delta^Y(s) + (1-\alpha)\delta(s), \quad \alpha \in [0, 1], \; \forall s \in \mathbb{R}. \tag{10}$$

Definition 5.1 Let $\delta^X(s)$ and $\delta(s)$ be two probability density functions, being $\delta(s)$ the benchmark distribution and $\delta^X(s)$ the investigated distribution. The *Mixture Dissimilarity Measure* (MDM) between $\delta^X(s)$ and $\delta(s)$ is the smallest $\alpha \in [0, 1]$—namely, $\alpha^* = M(\delta^X(s), \delta(s))$—such that there exists a probability distribution $\delta^Y(s, \alpha^*)$ satisfying (10).

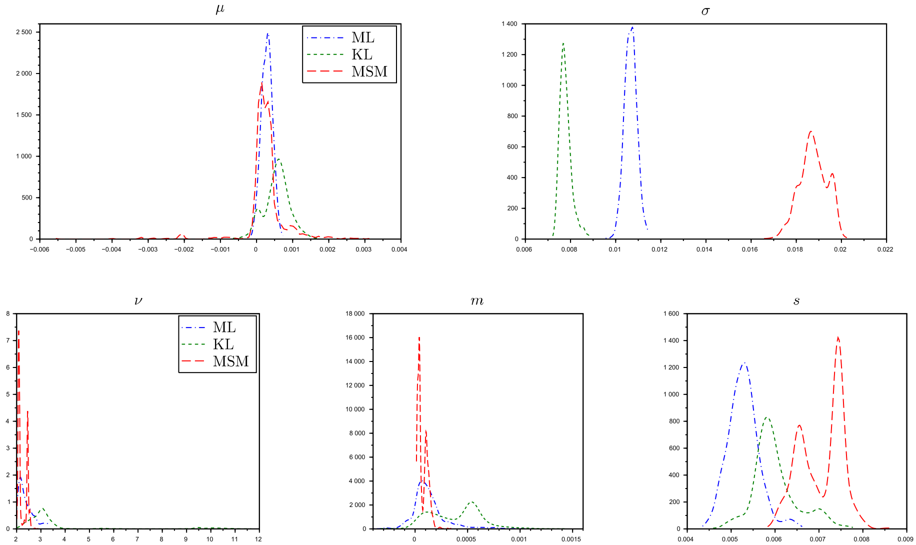

Fig. 11 Distribution of the estimate parameters obtained through 1000 resampling from the empirical distribution. First line: normal distribution $N(\mu, \sigma)$, mean μ, standard deviation σ. Second line: generalized Student's t, degrees of freedom ν, location m, scale s. For the KL case, the plots are drawn considering only the simulations where numerical convergence is obtained (about 70%)

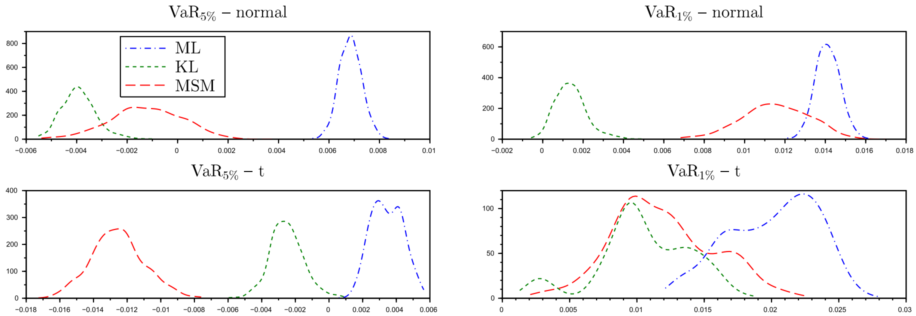

Fig. 12 VaR estimation errors. Differences between the VaR obtained by the estimated distributions and the actual VaR of the samples. For the KL case, the plots are drawn considering only the simulations where numerical convergence is obtained (about 70%)

The distribution $\delta^Y(s, \alpha^*)$ can be defined as the *random interference* of the distribution $\delta^X(s)$ with respect to the benchmark $\delta(s)$.

The geometric interpretation in $\mathbb{R}^n$ cannot be provided in the continuous case. However, Definition 5.1 can be related to the following

Problem 5.2 *Given the densities* $\delta(s), \delta^X(s),\ s \in \mathbb{R}$, *find the density* δ^Y *with the largest distance, in the sense of the norm* $\|f, g\| = \sqrt{\int_{-\infty}^{+\infty} (f(s) - g(s))^2 \, \mathrm{d}s}$, *from* $\delta^X(s)$ *and such that* (10) *is true.*

In the case $\delta(s) = \delta^X(s)$, Problem 5.2 has the obvious solution $\alpha^* = 0$, for any $\delta^Y(s) \in \mathcal{P}$.

To solve Problem 5.2, for $\alpha > 0$ rewrite (10) in the form

$$\delta^Y(s) = \frac{1}{\alpha}\left[\delta^X(s) - (1-\alpha)\delta(s)\right], \quad \alpha \in (0, 1]. \tag{11}$$

and rewrite it as the constrained optimization problem

$$\begin{array}{ll} \min_{\alpha} & \alpha \\ \text{s.t.} & \frac{1}{\alpha}\left[\delta^X(s) - (1-\alpha)\delta(s)\right] \geq [0] \\ & \alpha \in [0, 1] \end{array} \tag{12}$$

whose solution is

$$\alpha^* = 1 - \inf_{s\in\{s\in\mathbb{R}|\delta(s)>0\}} \frac{\delta^X(s)}{\delta(s)}. \tag{13}$$

5.1 Example (normal density)

In the case of parametric distribution functions, the application of the results above can be affordable. In particular, for Gaussian distributions, the matter can be straightforward. In fact, let the reference distribution $\delta(s)$ be a normal density with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Consider another normal density $\delta^X(s)$, with parameters $\mu^X \in \mathbb{R}$ and $\sigma^X > 0$. In this case,

$$\operatorname*{arginf}_{s\in\{s\in\mathbb{R}\}} \left\{\frac{\delta^x(s)}{\delta(s)}\right\} = \operatorname*{argsup}_{s\in\{s\in\mathbb{R}\}} \left\{s^2\left(\sigma^2 - (\sigma^X)^2\right) + s\,2\left(\mu(\sigma^X)^2 - \mu^X\sigma^2\right)\right\}$$

therefore, from simple computations and (13), we obtain

$$\begin{array}{lll} \text{if } \sigma > \sigma^X, & s^* = \pm\infty, & \alpha^* = 1 \\ \text{if } \sigma = \sigma^X, & s^* = \operatorname{sign}(\mu - \mu^X)\infty, & \alpha^* = 1 \\ \text{if } \sigma < \sigma^X, & s^* = \frac{\mu^X\sigma^2 - \mu(\sigma^X)^2}{\sigma^2 - (\sigma^X)^2}, & \alpha^* = 1 - \frac{\sigma}{\sigma^X}\exp\left\{-\frac{(\sigma\sigma^X(\mu-\mu^X)^2}{2(\sigma^2-(\sigma^X)^2)}\right\} \end{array}$$

To illustrate this case, let $\delta(s)$ be the standard normal density (i.e. $\mu = 0, \sigma = 1$), and consider the normal densities $\delta^X(s)$, with parameters $\mu = 0$ and $\sigma \in \{0.2, 0.4, 0.75, 0.9, 1, 1.5, 2, 3, 4\}$. Figure 13 presents the interference densities obtained in this example. It deserves a remark the fact that the shape of the interferences is different with respect to the discretized case (see Fig. 9), although the dispersions of the δ^Ys are comparable; moreover, the values of MDM follow a similar behaviour:

σ^X	0.2	0.4	0.75	0.9	1	1.5	2	3	4
MDM discr.	1	1	1	0.9867	0	0.2749	0.4391	0.6175	0.7108
MDM cont.	1	1	1	1	0	0.3333	0.5	0.6667	0.75

6 Conclusions

This paper enters the well-established debate in information theory on the measurement of the dissimilarity between two probability distributions associated with two random quantities. Specifically, we introduce the MDM—a novel dissimilarity measure—coming out from a joint application of a mixture distributions approach and an optimization model. Such a measure is shown to be equivalent to the Rényi divergence of infinite order, being an increasing transformation of it. The proposed measure also provides the identification of the random

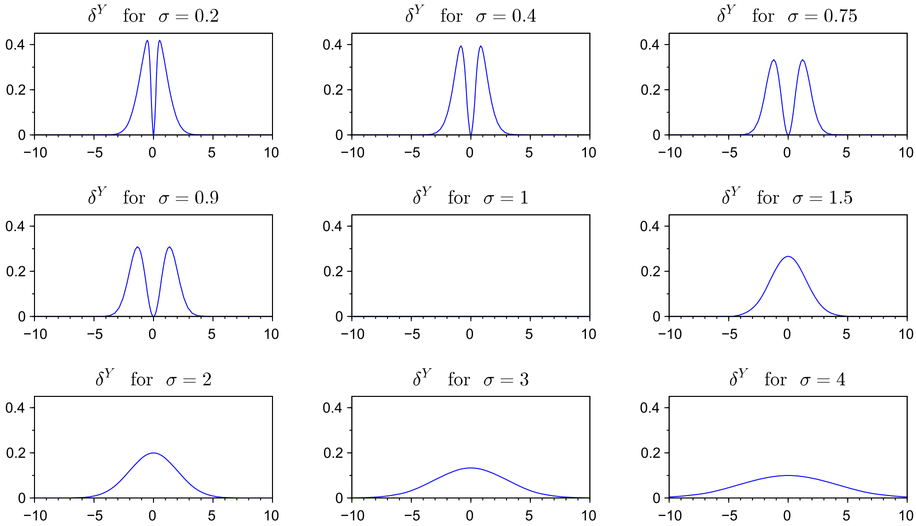

Fig. 13 Interference distributions δ^Y in the case of comparison between normal $(0, \sigma)$ distributions, with different values of σ

interference appearing when comparing a given probability distribution with a benchmark one.

We explore the main properties of the MDM. In doing so, we come to say that it is a quasimetric, i.e. a divergence with the addition of the validity of the triangular inequality. Furthermore—and differently from the Rényi divergence—the mixture nature of the MDM leads to a clear geometric interpretation for such a measure. In this respect, we offer a novel proof of the triangular property by following a geometric approach.

The proposed measure is tested in the context of rare events, showing a high level of usefulness. Indeed, it reasonably penalizes the deviations from the rare event probabilities more when such probabilities are underestimated than for overestimation. Moreover, as some simple examples in the cases of binomial and normal distributions showed, when some probabilities of the compared distribution δ^X are close to 0, the MDM yields a usable (large) value, whereas other unbounded divergences—such as the Kullback–Leibler—can incur in numerical overflows.

Importantly, the versatility of the proposed quasimetric allows its applicability over a wide set of real-world contexts, like pattern recognition and forecasting algorithms for finance and engineering studies.

As an avenue of future research, the MDM can be extended to the uncountable and continuous cases. We present some preliminary results in the case of continuous probability density functions, and an example with Gaussian distributions (see Sect. 5). We may observe that the comparison between the normal instance in the discrete case and in the continuous one suggests some remarkable differences in the behaviour of δ^Y. This evidence gives that the continuous case does not lead in general to the same results as the discrete one— hence, supporting the effects of the discretization on the interference. From a different perspective, one does not have substantial deviations in terms of α^*.

We also notice that the continuous normal distribution is associated with affordable computations, while other cases present formulas that cannot be easily simplified. In conclusion, we admit that the generalization of the MDM to the continuous case is a challenging opportunity for going on with further research.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

References

1. Antani S, Kasturi R, Jain R (2002) A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recogn 35(4):945–965
2. Balakrishnan S, Kolar M, Rinaldo A, Singh A (2017) Recovering block-structured activations using compressive measurements. Electron J Stat 11(1):2647–2678
3. Cerqueti R, Falbo P, Pelizzari C (2017) Relevant states and memory in Markov chain bootstrapping and simulation. Eur J Oper Res 256(1):163–177
4. Cha SH (2007) Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Methods Appl Sci 1(4):300–307
5. Chung JK, Kannappan PL, Ng CT, Sahoo PK (1989) Measures of distance between probability distributions. J Math Anal Appl 138(1):280–292
6. De Domenico F, Livan G, Montagna G, Nicrosini O (2023) Modeling and simulation of financial returns under non-Gaussian distributions. Physica A 622:128886
7. Dubey P, Müller HG (2022) Modeling time-varying random objects and dynamic networks. J Am Stat Assoc 117(540):2252–2267
8. Endres DM, Schindelin JE (2003) A new metric for probability distributions. IEEE Trans Inf Theory 49(7):1858–1860
9. Goldenberg I, Webb GI (2019) Survey of distance measures for quantifying concept drift and shift in numeric data. Knowl Inf Syst 60(2):591–615
10. Granero-Belinchon C, Roux SG, Garnier NB (2018) Kullback–Leibler divergence measure of intermittency: application to turbulence. Phys Rev E 97(1):013107
11. Hansen BE (1994) Autoregressive conditional density estimation. Int Econ Rev 35(3):705–730
12. Heikkinen VP, Kanto A (2002) Value-at-risk estimation using non-integer degrees of freedom of Student's distribution. J Risk 4(4):77–84
13. Hien LTK, Gillis N (2021) Algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. J Sci Comput 87(3):1–32
14. Johnson D, Sinanovic S (2001) Symmetrizing the Kullback–Leibler distance. IEEE Trans Inf Theory 1(1):1–10
15. Jorion P (2007) Value at risk—the new benchmark for managing financial risk, 3rd edn. McGraw-Hill
16. Kittler J, Zor C, Kaloskampis I, Hicks Y, Wang W (2018) Error sensitivity analysis of Delta divergence-a novel measure for classifier incongruence detection. Pattern Recogn 77:30–44
17. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
18. Li R, Nadarajah S (2020) A review of Students't distribution and its generalizations. Empir Econ 58:1461–1490
19. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37(1):145–151
20. Mandros P, Boley M, Vreeken J (2020) Discovering dependencies with reliable mutual information. Knowl Inf Syst 62(11):4223–4253
21. McLachlan GJ, Lee SX, Rathnayake SI (2019) Finite mixture models. Ann Rev Stat Its Appl 6:355–378
22. Mironov I (2017) Rényi differential privacy. In: 2017 IEEE 30th computer security foundations symposium (CSF). IEEE, pp 263–275
23. Platen E, Rendek R (2008) Empirical evidence on student-t log-returns of diversified world stock indices. J Stat Theory Pract 2(2):233–251

24. Rauber TW, Braun T, Berns K (2008) Probabilistic distance measures of the Dirichlet and Beta distributions. Pattern Recogn 41(2):637–645
25. Rasouli M, Chen Y, Basu A, Kukreja SL, Thakor NV (2018) An extreme learning machine-based neuromorphic tactile sensing system for texture recognition. IEEE Trans Biomed Circuits Syst 12(2):313–325
26. Rényi A (1961) On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Contributions to the theory of statistics. The Regents of the University of California, vol 1, pp 547–561
27. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
28. Smith A, Naik PA, Tsai CL (2006) Markov-switching model selection using Kullback–Leibler divergence. J Economet 134(2):553–577
29. Su F, Chan KS (2016) Option pricing with threshold diffusion processes. N Am Actuar J 20(2):133–141
30. Teoh HK, Quinn KN, Kent-Dobias J, Clement CB, Xu Q, Sethna JP (2020) Visualizing probabilistic models in Minkowski space with intensive symmetrized Kullback-Leibler embedding. Phys Rev Res 2(3):033221
31. Theodossiou P (1998) Financial data and the Skewed generalized T distribution. Manag Sci 44(12–part–1):1650–1661
32. Tran TH, Nguyen NT (2021) A model for building probabilistic knowledge-based systems using divergence distances. Expert Syst Appl 174:114494
33. Tulino AM, Li L, Verdú S (2005) Spectral efficiency of multicarrier CDMA. IEEE Trans Inf Theory 51(2):479–505
34. Van Erven T, Harremos P (2014) Rènyi divergence and Kullback-Leibler divergence. IEEE Trans Inf Theory 60(7):3797–3820
35. Xu X, Li R, Zhao Z, Zhang H (2022) Trustable policy collaboration scheme for multi-agent stigmergic reinforcement learning. IEEE Commun Lett 26(4):823–827
36. Yang R, Jiang Y, Mathews S, Housworth EA, Hahn MW, Radivojac P (2019) A new class of metrics for learning on real-valued and structured data. Data Min Knowl Disc 33(4):995–1016
37. Zhang W, Xie R, Wang Q, Yang Y, Li J (2022) A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback–Leibler divergence. Decis Support Syst 157:113765

Roy Cerqueti is full professor of Statistics at Sapienza University of Rome, Italy, and professor at the London South Bank University, UK. He is also affiliated to the GRANEM research centre at the University of Angers, France. He has a PhD in Mathematics for Economics and Finance at Sapienza University of Rome. His research activity includes quantitative methods for business and financial modelling, complex systems, information theory and stochastic calculus. He is author of more than one hundred scientific publications, most of them in prestigious international journals.

Mario Maggi is associate professor of Mathematical Finance at the University of Pavia, Italy. He also taught at other universities, including the University of Piemonte Orientale (Alessandria), Bologna (Rimini) and Bocconi (Milan). Since his PhD in Mathematical Finance (University of Brescia, Italy), his interests span from mathematical finance, behavioural decision making, to empirical analysis and numerical methods. He is the author of textbooks and many research papers published on international reviews, among others, Journal of Economic Psychology, Econometrics and Statistics, Annals of Operations Research, Annals of Finance.