International Statistical Review

International Statistical Review (2024) doi: 10.1111/insr.12592

Alternative Approaches for Estimating Highest-Density Regions

Nina Deliu[1,2] **and Brunero Liseo**[1]

[1]*MEMOTEF Department, Sapienza University of Rome, Rome, Italy*
[2]*MRC—Biostatistics Unit, University of Cambridge, Cambridge, UK*
Correspondence *Nina Deliu, MEMOTEF Department, Sapienza University of Rome, Rome, Italy*
Email: nina.deliu@uniroma1.it

Summary

Among the variety of statistical intervals, highest-density regions (HDRs) stand out for their ability to effectively summarise a distribution or sample, unveiling its distinctive and salient features. An HDR represents the minimum size set that satisfies a certain probability coverage, and current methods for their computation require knowledge or estimation of the underlying probability distribution or density *f*. In this work, we illustrate a broader framework for computing HDRs, which generalises the classical density quantile method. The framework is based on *neighbourhood* measures, that is, measures that preserve the order induced in the sample by *f*, and include the density *f* as a special case. We explore a number of suitable distance-based measures, such as the *k*-nearest neighbourhood distance, and some probabilistic variants based on *copula models*. An extensive comparison is provided, showing the advantages of the copula-based strategy, especially in those scenarios that exhibit complex structures (e.g. multimodalities or particular dependencies). Finally, we discuss the practical implications of our findings for estimating HDRs in real-world applications.

Key words: anomaly detection; copula models; density estimation; *k*-nearest neighbourhood; statistical intervals.

1 Introduction

A ubiquitous problem in statistics is to derive statistical intervals or regions—especially in the multivariate setting—for population parameters or other unknown quantities. Their role is to provide a way to quantify and describe the uncertainty about a quantity of interest, or simply a way to summarise the information contained in a distribution. Statistical regions may address different problems. For example, a *confidence interval* (CI) describes the uncertainty related to an estimate for an unknown *parameter*, while a *prediction interval* provides bounds for one or more future *observations*. Alternatively, a *tolerance interval* is the interval expected to contain a specified proportion of the sampled population. In a Bayesian setting, *highest posterior density* or *credible* regions provide, in a natural way, set estimates for a specific parameter (Box & Tiao, 1992; Turkkan & Pham-Gia, 1993). We refer to Meeker *et al.* (2017) and Krishnamoorthy & Mathew (2009) for an overview. Furthermore, even when the focus is on a specific type of interval, for example, a two-sided 95% prediction interval, questions on how this should be defined may still arise. Should we use the interval symmetric about the mean or the interval of shortest length, among others? Although each of these intervals has 95% coverage, they may all be different. Consider, for example, the case of nonsymmetric and/or multimodal distributions, with an illustration given in Figure 1. An additional layer of complexity arises in multivariate settings, where there are no unique agreed definitions, and generalisations include the concept of simultaneous CIs (Guilbaud, 2008), multivariate CIs (Korpela *et al.*, 2017), or different definitions for multivariate quantiles (Cai, 2010; Coblenz *et al.*, 2018; Figalli, 2018), among others.

The multivariate setting will be the target of this work, with a focus on bivariate distributions. In particular, our interest is devoted to statistical *regions* for summarising probability distributions in the form of *highest-density regions* (HDRs; Hyndman, 1996). Statistical regions other than HDRs are beyond the scope of the present work, and we refer to Meeker *et al.* (2017) and Krishnamoorthy & Mathew (2009) for a comprehensive survey on the broader topic. As the name suggests, an HDR specifies the set of points of highest density: the density for points inside the region must be higher than that for points outside it. More specifically, considering a d -dimensional continuous variable of interest $X \in \mathbb{R}^d$, $d \geq 1$, with probability density function f, the problem is to estimate minimum volume sets of the form $C(f_\alpha) = \{x : f(x) \geq f_\alpha\}$, such that $P(X \in C(f_\alpha)) \geq 1 - \alpha$, where $1 - \alpha$, with $\alpha \in (0, 1)$, represents a prespecified coverage probability. Although they share substantial similarities with *multivariate quantiles*, estimating an HDR differs from estimating level sets in that one is interested in specifying a probability content rather than the level directly. This complicates the problem, and we refer to Doss & Weng (2018) for more details.

The scope of an HDR can be wide and diverse; the following are possible applications.

Forecasting To obtain a prediction or forecast region for a set of observable variables in order to inform the most likely future realisations and convey in a simple way the accuracy of a forecast (for illustrative examples, see, e.g. Hyndman, 1996; Kim *et al.*, 2011).

Anomaly detection To detect abnormal observations from a sample: if a data point does not belong to a region of 'normal' data (the HDR), then it is regarded anomalous (see, e.g. Steinwart *et al.*, 2005, and references therein).

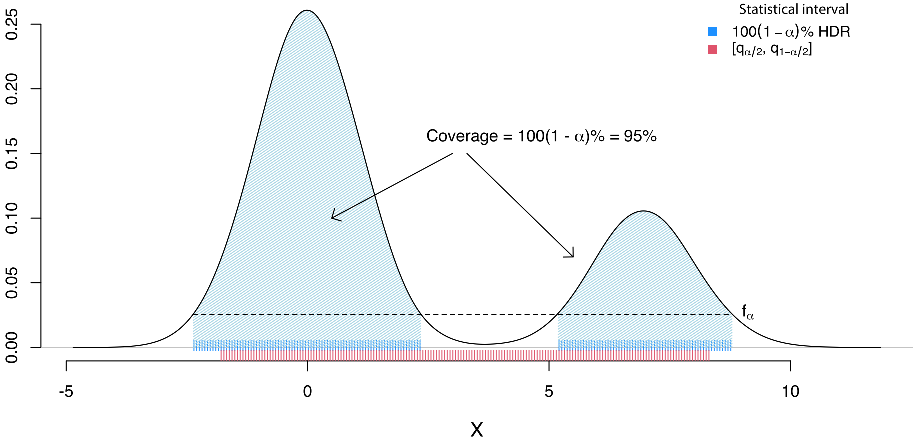

FIGURE 1. *Comparison between two* $100(1 - \alpha)\%$ *probability intervals for a normal mixture density: an HDR and an equal-tailed interval. The coverage parameter* α *is set to 0.05.*

Unsupervised or semi-supervised classification To identify areas or clusters with a relatively high concentration of a given phenomenon; see, for example, the work of Saavedra-Nieves (2022), aimed at finding areas of high incidence of coronavirus.

This work will primarily be driven by the problem of anomaly detection, characterising a broad spectrum of applied domains, going from astrophysics to diagnostics and sport analytics. More specifically, our interest is to develop a valid and efficient framework in support of the worldwide doping detection mission headed by the World Anti-Doping Agency (WADA, 2021). In practice, WADA's current analytical implementation is based on identifying reference ranges that discriminate well between normal and abnormal values for predefined biomarkers of interest (Sottas *et al.*, 2007). This is done following a univariate approach, with reference ranges, in the form of equal-tailed intervals, derived for each biomarker separately. Clearly, addressing this problem over increased dimensions presents significant challenges, including the presence of data with complex dependence structures, in addition to distributional multimodalities or skewness.

Due to their flexibility 'to convey both multimodality and asymmetry', HDRs are argued to be a more effective summary of the distribution (Hyndman, 1995). In the case of unimodal symmetric distributions, such as the normal distribution, an HDR coincides with the usual probability region symmetric about the mean, spanning the $\alpha/2$ and $1 - \alpha/2$ quantiles. However, in the case of a multimodal distribution, it may consist of several disjoint subregions, each containing a local mode. This provides useful information that could not be traced by other probability regions such as an equal-tailed interval (see Figure 1 for an illustrative example).

As the name suggests, estimating an HDR is based on knowing the density function f of the variable of interest X. However, as typically occurs in practice, this quantity is unknown and the estimation of an HDR requires estimating f first. The seminal paper of Hyndman (1996) discusses the *density-quantile approach* for computing HDRs in such settings. Although for unidimensional problems this task can be achieved very accurately using methods such as the kernel density estimator (KDE Parzen, 1962) or the local likelihood approach (Hjort & Jones, 1996), it may be inefficient for multidimensional problems (Liu *et al.*, 2007). In fact, over increased dimensions, KDE suffers from the difficulty of finding optimal kernel functions and the corresponding bandwidths (i.e. the smoothing parameters). In particular, bandwidth selection in KDE is recognised as the most crucial and difficult step (see, e.g. chapter 2 in Wand & Jones, 1994a), with no definite and widely accepted solution. Furthermore, high-dimensional data also pose challenges from an algorithmic/computational perspective when deriving the associated HDR.

In this work, we illustrate a broad framework for estimating HDRs, which generalises the current density-quantile approach implementable on the basis of a consistent estimator of the (multivariate) density (Hyndman, 1996). The proposed framework is based on *neighbourhood* measures (Munoz & Moguerza, 2006), that is, measures that preserve the order induced in the sample by the density function. Notably, it includes the widely-used density estimation procedure as a special case. We then elaborate on and evaluate a number of suitable probabilistic- and distance-based measures, including a variation of the measure adopted by the *k-nearest neighbours* algorithm. In particular, motivated by the ubiquitous role of *copula* modelling (Nelsen, 2006) in modern statistics, among probabilistic-based measures, we explore the use of copulae in an HDR estimation context. Interestingly, copulae introduce more flexibility to deal with multivariate random vectors, by separately estimating the marginals and their dependence structure, that is, the copula model. In addition, by placing a strong focus on the dependence model, a copula approach has the advantage of better capturing data specificities, especially when these exhibit complex relationships such as asymmetric and/or tail dependencies.

The remainder of this manuscript is organised as follows. In Section 2, we introduce the problem of interest and review existing methods for HDR estimation. The general neighbourhood-quantile framework is described in Section 2.3. In Section 3, we discuss and propose alternative measures in the context of HDRs, including some variants based on copulae. Section 4 provides a comprehensive comparison among the introduced measures and the standard kernel density estimator. Empirical studies are focused on bivariate scenarios that vary according to the complexity of the data (marginal and dependence structure, multimodality etc.) and the sample size. A distribution on a compact support, namely, the Dirichlet distribution, is also considered to support practical applications with compositional data. An application to the MAGIC data set, which classifies high-energy Gamma particles in the atmosphere (Bock *et al.*, 2004), is illustrated in Section 4.5. We conclude in Section 5 by summarising the main findings and discussing their implications for estimating HDRs in real-world problems.

2 A General Framework for HDR Estimation

Let $\{X_1, \ldots, X_n\}$ be a sample of n independent and identically distributed (iid) replications of a random variable X defined on $\mathbb{R}^d$, with $d \geq 1$. In our problem, we assume to have access to a sample $s_n = \{x_1, \ldots, x_n\} \in S_n$ of their actual realisations, with S_n the sample space. We then wish to use s_n for estimating an HDR, that is, a statistical region containing those sample values of relatively high density (see Definition 1, attributed to Hyndman, 1996). We denote by $X_{ij}\,(x_{ij})$ the j-th component of $X_i\,(x_i)$, for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. We restrict our discussion to continuous random variables X and, unless otherwise stated, we denote by f their probability density function (PDF) and by F their cumulative density function (CDF).

Definition *(Highest-density region; Hyndman, 1996) Denote by f the PDF of a continuous, possibly multivariate, random variable $X \in \mathbb{R}^d$; the $100(1-\alpha)\%$ HDR is defined as the subset $C(f_\alpha)$ of the sample space of X such that:*

$$C(f_\alpha) = \{x\colon f(x) \geq f_\alpha\},$$

where f_α is the largest constant such that $P(X \in C(f_\alpha)) \geq 1 - \alpha$, with $\alpha \in (0, 1)$.

One of the most distinctive properties of HDRs is that, among all regions of probability coverage $100(1-\alpha)\%$, the HDR has the smallest possible volume. The notion of 'smallest' is to be understood with respect to some measure such as the usual Lebesgue measure; in the continuous one-dimensional case that would lead to the shortest-length set, while in two dimensions that would be the smallest-area set. It also follows from the definition that the boundary of an HDR consists of those values of the sample space with equal density. Hence a plot of a bivariate HDR has as the boundary a contour plot.

2.1 Density Quantile Approach

The study of HDRs has been largely enhanced by Hyndman, who proposed the *density quantile approach* (outlined in Proposition 1) to estimate multivariate HDRs (Hyndman, 1996). Today, this still represents the typical strategy and involves estimating the density.

Proposition 1. *(Hyndman, 1996) Let $\{f(x_1), \ldots, f(x_m)\}$ be a sample of independent observations of size m of the random variable $Y = f(X)$, with f a bounded and continuous function in x. Consider the ordered sample $\{f_{(1)}, \ldots, f_{(m)}\}$ with $f_{(j)}$ the j-th largest among the $f(x_i)$'s so that $f_{(j)}$ is the (j/m) sample quantile of Y. Then, given a constant $\alpha \in [0, 1]$, and denoted with $\lfloor j \rfloor$ the greatest integer less than or equal to j, in probability,*

$$\hat{f}_{\alpha} \doteq f_{(\lfloor \alpha m \rfloor)} \rightarrow f_{\alpha} \text{ as} m \rightarrow \infty, \; C_m(\hat{f}_{\alpha}) \doteq \{x : f(x) \; > \; \hat{f}_{\alpha}\} \rightarrow C(f_{\alpha}) \text{ as} m \rightarrow \infty.$$

Basically, the HDR is derived based on the sample quantile of the density f. However, the density function itself is often unknown and one has to estimate it based on a set of available iid observations $s_n \doteq \{x_1, \; \dots, \; x_n\}$. In this case, the $100(1 \; - \; \alpha)\%$ HDR can be estimated as

$$\hat{C}_n\left(\hat{f}_{\alpha}\right) \doteq \{x : f_n(x) \; > \; f_{(\lfloor \alpha n \rfloor)}\}, \tag{1}$$

withf_n being a possibly consistent estimator of$f$. Note that for small n, it may not be possible to get a reasonable density estimate. Also, with few observations and no prior knowledge on the underlying density function, there seems to be little point in attempting to summarise the density.

2.2 One-Class Neighbour Machines (OCNM) Approach

An alternative approach to HDR estimation, inspired by the theory of support vector machines (Schölkopf *et al.*, 2001) has been introduced in the machine learning literature by Munoz & Moguerza (2006). The procedure is outlined in Proposition 2, and involves the notion of *neighbourhood measures* (see Definition 2, attributed to Munoz & Moguerza, 2006).

Definition *(Neighbourhood measure Munoz & Moguerza, 2006) Let X be a random variable with density function f defined on $\mathbb{R}^d$. Denoted by S_n the set of random iid samples $s_n = \{x_1, \; \dots, \; x_n\}$ of size n (drawn from f), the real-valued function $g : \mathbb{R}^d \times S_n \rightarrow \mathbb{R}$ is a neighbourhood measure if one of the following holds:*

$$\text{(a)} f(x) \; < \; f(y) \quad \Rightarrow \lim_{n \rightarrow \infty} \mathbb{P}(g(x, \; s_n) \; > \; g(y, \; s_n)) = 1, \;\; x, y \in s_n, \; \forall s_n \in S_n,$$

$$\text{(b)} f(x) \; < \; f(y) \quad \Rightarrow \lim_{n \rightarrow \infty} \mathbb{P}(g(x, \; s_n) \; < \; g(y, \; s_n)) = 1, \;\; x, y \in s_n, \; \forall s_n \in S_n.$$

The function g is called either a (a) sparsity *or a (b)* concentration *measure.*

Proposition 2. *(Munoz & Moguerza, 2006) Consider an iid sample $s_n = \{x_1, \; \dots, \; x_n\}$ and a sparsity measure g. Define $\rho^* = g(x_{(vn)}, \; s_n)$, with $x_{(vn)}$ being the (vn)-th sample in the order induced in s_n by g, provided that $vn \in \mathbb{N}$; otherwise, the least integer greater than vn, denoted by ⊠vn⊠, is taken. Then, the binary decision function $h(x) = \text{sign}(\rho^* \; - \; g(x, \; s_n))$ is such that:*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(x_i) \neq \; - \; 1) = v\right) \rightarrow 1 \text{ as} n \rightarrow \infty.$$

where $\mathbb{I}$ denotes the indicator function, and

$$C_n^{\text{OCNM}} \doteq \{x : h(x) \geq 0\} \rightarrow C(f_v) \doteq \{x : f(x) \geq f_v\} \text{ as} n \rightarrow \infty,$$

where $C(f_v)$ is the minimum-volume set such that $\mathbb{P}(C(f_v)) \geq v$, with $v \in [0, 1]$.

Notably, the OCNM algorithm relaxes the HDR estimation problem in the following sense: instead of estimating and evaluating the density f, a more general and potentially simpler measure g that asymptotically preserves the order induced by the density, can be considered. It is, however, important to remark that the quality of the estimation procedure heavily depends on using a neighbourhood measure. If the measure used is neither a concentration nor a sparsity measure, there is no reason why the method should work.

2.3 Neighbourhood-Quantile Approach

We now discuss a hybrid approach that can be viewed as (i) a generalisation of the density quantile approach and (ii) a restatement of the OCNM method directly in terms of its solution ρ^*. Compared with the former, it ensures a wider applicability allowing for a more general set of functions, including the density f as a special case; with respect to the latter, it offers a more direct, interpretable, and computationally efficient method.

Theorem 1. *Let X be a continuous random variable defined on* $\mathbb{R}^d$ *with density function f, and consider a set* $s_n = \{x_1, \ldots, x_n\} \in S_n$ *of size* n *(drawn from* f*). Assume* $g: \mathbb{R}^d \times S_n \rightarrow \mathbb{R}$ *is a neighbourhood measure. Then an estimate of the* $100(1 - \alpha)\%$ *HDR can be obtained as*

$C_n = \{x: g(x, s_n) \leq g(x_{(\boxtimes(1-\alpha)n\boxtimes)}, s_n)\}ifg$is a sparsity measure, $C_n = \{x: g(x, s_n) \geq g(x_{(\lfloor \alpha n \rfloor)}, s_n)\}$

if$g$is a concentration measure,

where $g(x_{(\lfloor \alpha n \rfloor)}, s_n)$ is the α-quantile of the sample $\{g(x_1, s_n), \ldots, g(x_n, s_n)\}$, and $\lfloor j \rfloor$ denotes the greatest integer less than or equal to j.

The proof is straightforward when considering the relationship between the region C_n^{OCNM} as defined in Proposition 2 and the region C_n as defined in Theorem 1. In fact, without loss of generality, taking g to be a sparsity measure, and noticing that v plays the role of $1 - \alpha$, one can observe that:

$$\begin{aligned} C_n^{\text{OCNM}} &= \{x: h(x) = \text{sign}(\rho^* - g(x, s_n)) \geq 0\} \\ &= \{x: \text{sign}(g(x_{(\boxtimes vn\boxtimes)}, s_n) - g(x, s_n)) \geq 0\} \\ &= \{x: g(x, s_n) \leq g(x_{(\boxtimes(1-\alpha)n\boxtimes)}, s_n)\} \\ &= C_n. \end{aligned}$$

Remark. If g is chosen to be a concentration measure, then, to ensure a $100(1 - \alpha)\%$ coverage (notice that in this case $\mathbb{P}(C(f_v)) = 1 - v$), the decision value ρ^* induced by the concentration measure is given by $\rho^* = g(x_{(\lfloor \alpha n \rfloor)}, s_n)$.

Noticing that the density represents a concentration measure, the resemblance with the density quantile approach in Proposition 1 should now be clear. In fact, from the perspective of the density quantile approach, one can view Theorem 1 as a generalisation of Proposition 1, where$f$ is replaced by any function g that satisfies the criteria of neighbourhood measures. Intuitively, provided that the density function f is replaced by a function that preserves the order induced in the sample by$f$, the estimated HDR is asymptotically valid. Neighbourhood measures ensure this ranking. To see it, without loss of generality, consider a sparsity measure g and a sample $s_n = \{x_1, \ldots, x_{(1-\alpha)n}, \ldots, x_n\}$ ordered so that

$$f(x_1) < \ldots < f(x_{(1-\alpha)n}) < \ldots < f(x_n),$$

where, for simplicity, we suppose $(1 - \alpha)n \in \mathbb{N}andf(x_j) \neq f(x_j)$ for all $i \neq j$. From Definition 2 (a), for each pair (x_i, x_j), with $i < j$, it holds that $\mathbb{P}(g(x_i, s_n) > g(x_j, s_n)) \rightarrow 1$. Hence, given $\epsilon \in (0, 1)$, there exists $n_{ij} \in \mathbb{N}$ such that $\mathbb{P}(g(x_i, s_{n_{ij}}) > g(x_j, s_{n_{ij}})) > 1 - \epsilon$. Taking $n \geq \max\{n_{ij}\}$, it is guaranteed that $\mathbb{P}(g(x_1, s_n) > \ldots > g(x_{(1-\alpha)n}, s_n) > \ldots > g(x_n, s_n)) > 1 - \epsilon$. Therefore, as $n \rightarrow \infty$

$$\mathbb{P}\left(g(x_1, s_n) > \ldots > g(x_{(1-\alpha)n}, s_n) > \ldots > g(x_n, s_n)\right) \to 1.$$

Remark Proposition 1 is a special case of Theorem 1 with $g(x, s_n)$ being either: (a) a concentration measure $g(x, s_n) \propto \hat{f}(x, s_n)$ or; (b) a sparsity measure $g(x, s_n) \propto \frac{1}{\hat{f}(x, s_n)}$, where $\hat{f}$ can be any consistent density estimator. Among the plethora of density estimators, in this work we focus on KDE (Parzen, 1962), and use it as a benchmark measure for the proposed comparators in Section 3. Specifically, given a set of iid observations $s_n \doteq \{x_1, \ldots, x_n\}$ drawn from an unknown target density f, the KDE measure $M_0(x, s_n, h)$ at the location x is defined as

$$M_0(x, s_n, h) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right), \tag{2}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ denotes the kernel function, satisfying $\int K(x)dx = 1$ and $K(x) \geq 0, \forall x$, and $h > 0$ the bandwidth hyperparameter. Details on the chosen kernel and bandwidth value will be given in Section 4.

3 Alternative Measures for Estimating HDRs

We now propose a number of neighbourhood measures that could be used to estimate HDRs. We elaborate on some measures that have been successfully employed in areas such as classification or clustering and introduce some novel ideas, including a copula-based approach. Although some of the discussed distances are popular in existing statistical domains, for example, the k-nearest neighbours distance and its use in classification and regression, these have not been considered in the context of HDRs.

3.1 ***k*NN-Euclidean Distance**

This corresponds to the sum of all Euclidean distances of a given point x from its k-nearest neighbours. The concept of closeness ('nearest') is defined according to the Euclidean metric or L^2-norm denoted by $\| \cdot \|_2$; for $k = 1$, the nearest neighbour is the point x itself, and in this case, the distance is zero.

Definition *kNN-Euclidean distanceGiven a data point $x \in \mathbb{R}^d$ of a sample set s_n of size n, and an integer $k \in [1, n]$, we define the k-nearest neighbourhood Euclidean distance of point x from sample s_n as*

$$M_1(x, s_n, k) \doteq \sum_{i=1}^{k} \|x - x_{(i)}\|_2,$$

where $x_{(i)}$ denotes the i-th observation in the reordered data such that $0 = \|x - x_{(1)}\|_2 \leq \ldots \leq \|x - x_{(i)}\|_2 \leq \ldots \|x - x_{(n)}\|_2$.

$M_1(x, s_n, k)$ is entirely based on a metric distance, that is, the Euclidean metric, and represents a sparsity measure. The property follows from the convergence in probability of the k -nearest neighbour density estimator (Silverman, 1996). Similar distances have been used in the literature for nonparametric classification and regression starting from the seminal work of Fix & Hodges (1951) and Cover & Hart (1967), which led to the well-known k-nearest neighbours (kNN) algorithm. Notably, the distance to the nearest neighbours can also be seen as a local density estimate (Loftsgaarden & Quesenberry, 1965; Silverman, 1996) and as a special case of a variable-bandwidth KDE with a uniform kernel (Terrell & Scott, 1992). A large kNN distance indicates that the density is small and vice versa. Furthermore, by ranking each data point according to the distance from its k-nearest neighbours, this measure can also be used as an outlier score in anomaly detection (Ramaswamy *et al.*, 2000).

Compared with other methods, the kNN approach has several advantages such as (i) being purely nonparametric, hence able to flexibly adapt to any continuous distribution; (ii) having a reasonable time complexity; (iii) depending on a unique hyperparameter k, whose tuning is relatively simple. The choice of k should be made based on the sample data. Generally, higher values of k reduce the effect of noise; however, they underfit the model—making boundaries between classes less distinct—and are computationally more expensive, especially for large d. A general rule of thumb in classification is $k = \lfloor \sqrt{n} \rfloor$, with n the number of samples in the dataset. Further investigation of the role of k in different settings and how it relates to the missclassification error is given in Supplementary Material B and in Meeker *et al.* (2017).

3.2 *k*NN-CDF Distance

Although the Euclidean distance and other general notions of measures (including concepts like area or volume) can certainly be useful for capturing relevant insights on the topology of a given set, they may be equipped with a probability measure to better resemble the notion of density. Let $\mathbb{P}$ be a probability measure defined on $\mathbb{R}^d$ and denote by F its associated CDF and by F_i the CDF of the i-th marginal.

Given two data points $x_1 = (x_{11}, \dots, x_{1d}) \in \mathbb{R}^d$ and $x_2 = (x_{21}, \dots, x_{2d}) \in \mathbb{R}^d$ from the sample set s_n, a preliminary version of the CDF distance between the two data points, denoted by $d_{\mathbb{P}}(x_1, x_2)$, has been defined in Venturini (2015) in the context of clustering problems as

$$d_{\mathbb{P}}(x_1, x_2) = \sqrt{\sum_{j=1}^{d} \left(F_j(x_{1j}) - F_j(x_{2j})\right)^2}. \tag{3}$$

In the typical case of unknown marginal CDFs, their empirical counterpart $F_{n,j}(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(x_{ij} \le t)$, $j = 1, \dots, d$, could be considered.

The distance $d_{\mathbb{P}}$ can be interpreted as the composition of the (nonlinear) transformation $F : \mathbb{R}^d \rightarrow [0, 1]$ and the computation of the ordinary Euclidean distance. It fulfils all the properties of being a proper metric. Furthermore, it has the nice property that the distance between two

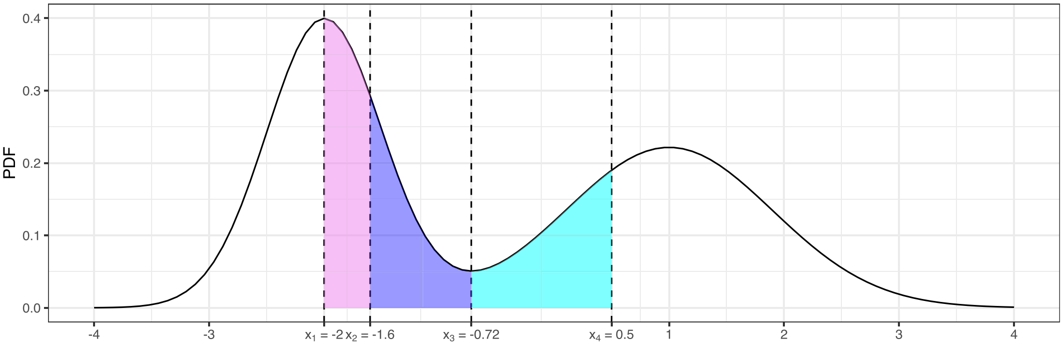

FIGURE 2. *Illustration of the CDF-distances computed on the set of points $x_1 = -2$, $x_2 = -1.6$, $x_3 = -0.72$, $x_4 = 0.5$ based on a Gaussian mixture model.*

points is proportional to the probability contained between the two points. However, as defined in Equation (3), it presents some key limitations for inferring the underlying density and estimating HDRs. Consider for simplicity the univariate Gaussian mixture model illustrated in Figure 2 and the set of points $x_1 = -2,\ x_2 = -1.6,\ x_3 = -0.72,\ x_4 = 0.5$. When evaluating the CDF distance between x_2 and all other points, we have that $d_{\mathbb{P}}(x_2, x_1) \approx d_{\mathbb{P}}(x_2, x_3)$, but there is little to say about the density around x_1 and x_3, apart from realising that they are clearly different. One could gain more information if one were to compute the CDF distance from the global mode x_1, assuming one has this information, noticing that the higher $d_{\mathbb{P}}(x_1, x_i)$, for all i, the lower the density of x_i. However, this is no longer verified for multimodal densities. In fact, we have $d_{\mathbb{P}}(x_1, x_4) > d_{\mathbb{P}}(x_1, x_3)$, but $f(x_4) > f(x_3)$, contrasting with the definition of a neighbourhood measure.

Motivated by these limitations and inspired by the idea of the kNN distance in Definition 3, we propose a new variant consisting of the sum of CDF-distances between a point x and its k -nearest neighbours. Here, 'nearest' should again be understood according to the Euclidean distance. The rationale is simple: the higher the CDF distances between a point and its neighbours, the higher one expects the density to be at that point. Because the k neighbours of each data point will be at different (Euclidean) distances compared with the k neighbours of the other data points, one needs to properly scale or weight the CDF distances according to this information. The proposed measure is reported in Definition 4.

Definition ***k**NN-CDF distanceGiven F_i, the i-th marginal of a CDF F, for $i = 1, \ldots, d$, and an integer $k \in [1, n]$, we define the k-nearest neighbourhood CDF distance as*

$$M_2(x, s_n, k) = \begin{cases} \sum_{i=2}^{k} \frac{d_{\mathbb{P}}(x, x_{(i)})}{\|x - x_{(i)}\|_2} = \sum_{i=2}^{k} \frac{\sqrt{\sum_{j=1}^{d} \left(F_j(x_j) - F_j(x_{(i)j})\right)^2}}{\sqrt{\sum_{j=1}^{d} \left(x_j - x_{(i)j}\right)^2}} k > 10 k = 1, \end{cases}$$

where $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, and $x_{(i)} = (x_{(i)1}, \ldots, x_{(i)d})$ is the i-th observation in the sample s_n such that $\|x - x_{(1)}\|_2 \leq \ldots \leq \|x - x_{(i)}\|_2 \leq \ldots \|x - x_{(n)}\|_2$.

Property *SemimetricConsider the M_2 measure as defined in Definition 4. It can be easily verified that, for any $k \in [1, n]$, if one restricts the CDF distance to x and its k-th neighbour $x_{(k)}$, then $M_2(x, x_{(k)}, k) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a semimetric. In fact, for any set $x, x_{(k)} \in \mathbb{R}^d$, it follows that:*

- If $x = x_{(k)},\ M_2(x, x_{(k)}, 1) = 0$ by definition.
- If $x \neq x_{(k)},\ M_2(x, x_{(k)}, k) > 0$. In fact, $(x_j - x_{(k)j})^2 > 0$ and $(F_j(x_j) - F_j(x_{(k)j}))^2 > 0$ for all j and all k, when $x \neq x_{(k)}$ and X is continuous.
- $M_2(x, x_{(k)}, k) = M_2(x_{(k)}, x, k)$, as $(x_j - x_{(k)j})^2 = (x_{(k)j} - x_j)^2$ and $(F_j(x_j) - F_j(x_{(k)j}))^2 = (F_j(x_{(k)j}) - F_j(x_j))^2$ for all j and all k.

  The triangular inequality, stating that $M_2(x, x_{(k)}, k) \leq M_2(x, x_{(k')}, k) + M_2(x_{(k')}, x_{(k)}, k)$, for all $x, x_{(k)}, x_{(k')} \in \mathbb{R}^d$, does not hold. It is enough to consider a simple counterexample such as $d = 1$, F the CDF of a standard normal distribution, and the three points $x = 1 < x_{(k)} = 2 < x_{(k')} = 3$.

3.3 *ϵ-Neighbourhood Multivariate CDF Distance*

The measure $M_2(x, s_n, k)$ introduced in Section 3.2 has several advantages; for example: (i) it takes into account the probabilistic information in the data, (ii) it is based on marginal CDFs which may be easy to estimate, and (iii) it has convenient computational times. However, using only the marginal CDFs may compromise their validity whenever the marginal components of a multivariate random variable $X \in \mathbb{R}^d$ show a significant dependence structure. In that case, the use of a multivariate CDF may be more appropriate. If the multivariate CDF is unknown, one could use its empirical counterpart. Notice, however, that this estimation part may compromise computational efficiency, especially in high dimensions and for large values of k. We therefore focus on an ϵ-neighbourhood version, which reduces the number of operations from k to 1, once ϵ is determined.

Definition *ϵ-neighbourhood multivariate CDF distanceGiven the CDF F of a multivariate variable $X \in \mathbb{R}^d$, we define the ϵ-neighbourhood multivariate CDF distance of a point $x = (x_1, \dots, x_d)$ to be the probability of that point belonging to a hyperrectangle of dimension d defined on vertexes $[x_1 - \epsilon_1, x_1 + \epsilon_1] \times \dots \times [x_d - \epsilon_d, x_d + \epsilon_d]$ scaled by hyperrectangle's d-volume:*

$$M_3(x, \epsilon) = \frac{\mathbb{P}(X \in [x - \epsilon, x + \epsilon])}{\prod_{j=1}^{d} 2\epsilon_j} = \frac{\sum_{v \in \mathcal{V}} (-1)^{n(v)} F(v)}{\prod_{j=1}^{d} 2\epsilon_j} \propto \sum_{v \in \mathcal{V}} (-1)^{n(v)} F(v), \tag{4}$$

where $v = (v_1, \dots, v_d)$, with $v_j \in \{x_j - \epsilon_j, x_j + \epsilon_j\}$, for $j \in 1, \dots, d$, and $n(v) = \sum_{j=1}^{d} \mathbb{I}(v_j = x_j - \epsilon_j)$. The sum is computed over the 2^d vectors of the set $\mathcal{V}$.

As in the case of the density f itself, when the CDF F is unknown, a consistent estimator, say F_n, may be used leading to

$$M_3(x, s_n, \epsilon) = \frac{\sum_{v \in \mathcal{V}} (-1)^{n(v)} F_n(v)}{\prod_{j=1}^{d} 2\epsilon_j} \propto \sum_{v \in \mathcal{V}} (-1)^{n(v)} F_n(v).$$

Relationship in Equation (4) can be derived by recursion starting from small d values. For $d = 2$, for example, it is easy to verify that

$$\begin{aligned}
M_3(x, \epsilon) &\propto \mathbb{P}(X \in [x - \epsilon, x + \epsilon]) = \mathbb{P}(X_1 \in [x_1 - \epsilon_1, x_1 + \epsilon_1], X_2 \in [x_2 - \epsilon_2, x_2 + \epsilon_2]) \\
&= \mathbb{P}(X_1 \in [x_1 - \epsilon_1, x_1 + \epsilon_1], X_2 \le x_2 + \epsilon_2) - \mathbb{P}(X_1 \in [x_1 - \epsilon_1, x_1 + \epsilon_1], X_2 < x_2 - \epsilon_2) \\
&= \mathbb{P}(X_1 \le x_1 + \epsilon_1, X_2 \le x_2 + \epsilon_2) - \mathbb{P}(X_1 < x_1 - \epsilon_1, X_2 \le x_2 + \epsilon_2) - \mathbb{P}(X_1 \le x_1 + \epsilon_1, X_2 \le x_2 - \epsilon_2) + \mathbb{P}(X_1 < x_1 - \epsilon_1, X_2 < x_2 - \epsilon_2) \\
&= F(x_1 + \epsilon_1, x_2 + \epsilon_2) - F(x_1 - \epsilon_1, x_2 + \epsilon_2) - F(x_1 + \epsilon_1, x_2 - \epsilon_2) + F(x_1 - \epsilon_1, x_2 - \epsilon_2) \\
&= \sum_{v \in \mathcal{V}} (-1)^{n(v)} F(v),
\end{aligned}$$

with $\mathcal{V} = \{(x_1 + \epsilon_1, x_2 + \epsilon_2), (x_1 - \epsilon_1, x_2 + \epsilon_2), (x_1 + \epsilon_1, x_2 - \epsilon_2), (x_1 - \epsilon_1, x_2 - \epsilon_2)\}$.

We now state the following result for this measure.

Property *Density equivalenceConsider the M_3 measure as defined in Definition 5 and assume that F is differentiable at any point x of its support. Then, as $\epsilon \to 0$, $M_3(x, \epsilon) \to f(x)$, for any $x \in \mathbb{R}^d$.*

The proof follows from basic probability and calculus theory.

3.4 Copula-Based Measures

A d-dimensional copula $C: [0, 1]^d \to [0, 1]$ is a CDF with uniform marginal distribution functions (Nelsen, 2006). If we consider a random vector $X = (X_1, \ldots, X_d)$ with joint CDF F and marginals $F_1, \ldots, F_d$, then the copula of X is represented by the joint distribution of $F_1(X_1), \ldots, F_d(X_d)$ and is derived as

$$\begin{aligned} C(u_1, \ldots, u_d) &= \mathbb{P}(F_1(X_1) \leq u_1, \ldots, F_d(X_d) \leq u_d) \\ &= \mathbb{P}(X_1 \leq F_1^{-1}(u_1), \ldots, X_d \leq F_d^{-1}(u_d)) \\ &= F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)). \end{aligned}$$

Letting $u_j \doteq F_j(x_j)$, this yields the following well-known result due to Sklar (1959):

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \text{ for all} x = (x_1, \ldots, x_d) \in \mathbb{R}^d. \tag{5}$$

Furthermore, when the random vector X is continuous with density f, we also have that

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times f_1(x_1) \times \ldots \times f_d(x_d), \tag{6}$$

where c is the density of the random vector $(F_1(X_1), \ldots, F_d(X_d)) \in [0, 1]^d$.

In summary, one can decompose every d-dimensional CDF F or PDF f into a composition of their marginal distribution functions and a d-copula. This allows us to redefine both M_0 in Equation (2) and M_3 in Definition 5 in terms of their copula representation. For example, in the case of the M_3 measure, we may construct its copula-based alternative given in Definition 6.

Definition *ϵ-neighbourhood copula-based CDF distanceGiven the CDF F of a random vector $X = (X_1, \ldots, X_d)$, we define the ϵ-neighbourhood copula-based CDF distance of point $x = (x_1, \ldots, x_d)$ to be*

$$M_3^{\text{Cop}}(x, \epsilon) = \frac{\sum_{v \in \mathcal{V}} (-1)^{n(v)} C(F_1(v_1), \ldots, F_d(v_d))}{\prod_{j=1}^d 2\epsilon_j} \propto \sum_{v \in \mathcal{V}} (-1)^{n(v)} C(F_1(v_1), \ldots, F_d(v_d)), \tag{7}$$

where $v = (v_1, \ldots, v_d)$, with $v_j \in \{x_j - \epsilon_j, x_j + \epsilon_j\}$, for $j \in 1, \ldots, d$, and $n(v) = \sum_{j=1}^d \mathbb{I}(v_j = x_j - \epsilon_j)$.

For $d = 2$, it is immediate to verify that

$$\begin{aligned} M_3^{\text{Cop}}(x, \epsilon) &\propto C(F_1(x_1 + \epsilon_1), F_2(x_2 + \epsilon_2)) - C(F_1(x_1 - \epsilon_1), F_2(x_2 + \epsilon_2)) \\ &\quad - C(F_1(x_1 + \epsilon_1), F_2(x_2 - \epsilon_2)) + C(F_1(x_1 - \epsilon_1), F_2(x_2 - \epsilon_2)). \end{aligned}$$

The main advantage of this representation over the one involving the joint CDF is that the estimation of the multivariate distribution, when unknown, is performed through the estimation of the (univariate) marginals, evading thus the curse of dimensionality (see, e.g. Nagler & Czado, 2016). Furthermore, copulae offer a flexible framework that captures complex dependence structures while offering direct control over the marginals.

4 Empirical Evaluation

4.1 Simulation Setup

We now evaluate and compare the proposed measures in an extensive number of simulated bivariate settings that vary according to the complexity of the data. In particular, we consider scenarios with different dependence structures induced by the copula model (e.g. tail or asymmetric dependencies), different marginal distributions (e.g. heavy-tailed or multimodal distributions), as well as different sample sizes.

(i) We consider different copula models for the dependence structure among marginals, all sharing the same degree of dependence, expressed via Kendall's τ, and set equal to 0.5. The list of different copulae follows.

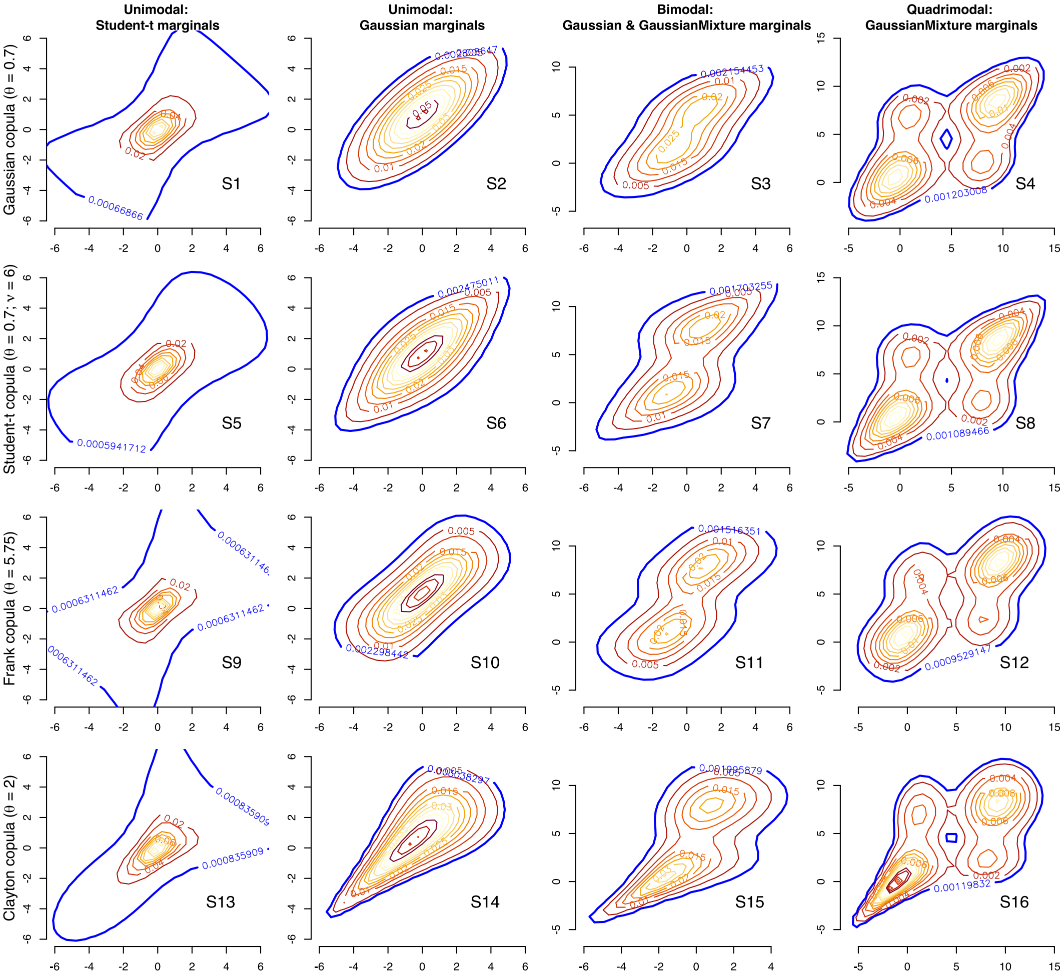

FIGURE 3. *Evaluated scenarios with their contour plots at different levels, included the one delimiting the 95% HDR (blue colour).*

$C^{\text{Gauss}}_{\theta=0.7}$	Gaussian family (Elliptical class) with zero tail dependence and radial symmetry.
$C^{\text{t}}_{\theta=0.7,\,\nu=6}$	Student's t family (Elliptical class) with radial symmetry.
$C^{\text{Frank}}_{\theta=5.75}$	Frank family (Archimedean class) with radial symmetry.
$C^{\text{Clay}}_{\theta=2}$	Clayton family (Archimedean class) not restricted to radial symmetry.

(ii) Each of the above copulae is then combined with different simulation schemes for the marginals, according to the following details. Fixing $\sigma^2 = 2$ and $w_1 = 1 - w_2 = 0.5$, and taking $\mu_{11} = 0$, $\mu_{12} = 9$, $\mu_{21} = 1$, $\mu_{22} = 8$, we consider the following:

Unimodal—heavy tails	Student's t model $X_i \sim t_{\nu=2},\ \ i = 1, 2.$
Unimodal	Gaussian model $X_i \sim \mathcal{N}(\mu_{i1}, \sigma^2),\ \ i = 1, 2.$
Bimodal	Gaussian $X_1 \sim \mathcal{N}(\mu_{11}, \sigma_1^2)$ & Gaussian mixture $X_2 \sim \sum_{k=1}^{2} w_k \mathcal{N}(\mu_{2k} + 5\mathbb{I}(k = 2), \sigma^2).$
Quadrimodal	Gaussian mixture $X_i \sim \sum_{k=1}^{2} w_k \mathcal{N}(\mu_{ik}, \sigma^2),\ \ i = 1, 2.$

(iii) We also consider a distribution on a compact supports, that is, the Dirichlet distribution defined on the $(K - 1)$-simplex, with $K = 3$ and with parameters $\boldsymbol{\alpha} = (1, 1, 2)$. In this case, the dependence structure is uniquely determined and its closed-form copula expression is given in Section 4.4.1.

(iv) We finally cover different sample sizes, with $n \in \{50, 100, 500, 1000\}$.

In total, 17 different scenarios are considered, named S1 to S17. Their graphical representation, in the form of contour plots, is provided in Figure 3 (S1–S16) and Figure 4 (S17).

By exploring such an extended number of simulation setups, we hope to cover a broad spectrum of potential scenarios that may occur in practice. In addition to offering a comprehensive understanding of the performances of the different measures, it may provide guidance to applied scientists who may need to choose the most ideal measure for their specific application. For

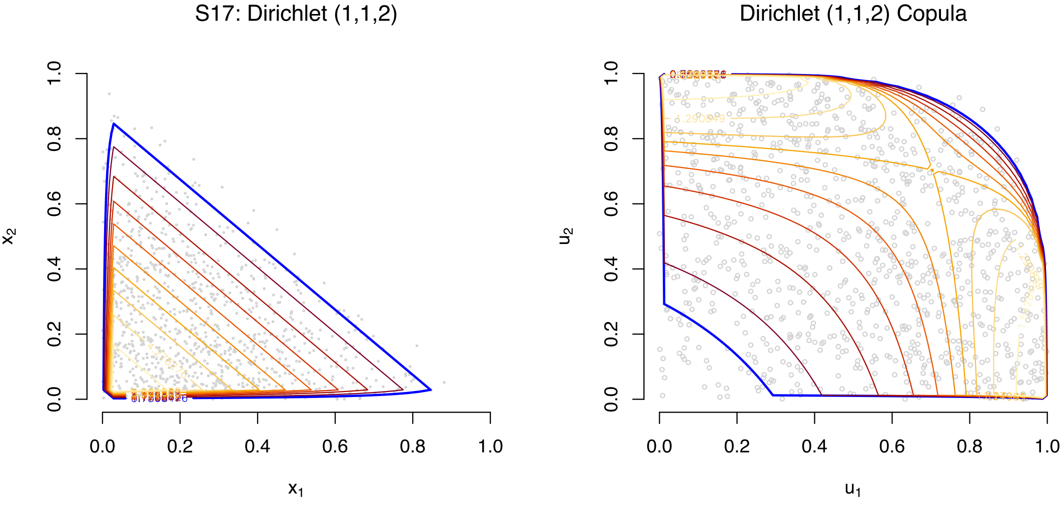

FIGURE 4. *Contours of a Dirichlet(1,1,2) density (left) and of its induced copula density (right). The 95% HDR is represented by the blue contour.*

example, the problem of doping detection when relying on the two primary biomarkers of the haematological module could well be related to scenario S15 (Gaussian and Gaussian mixture marginals and a Clayton copula; see Deliu & Liseo, 2024, for an illustrative example).

4.2 Evaluated Measures

In all settings, we employ the proposed neighbourhood-quantile method (see Section 2.3) on data samples of different sizes generated according to the aforementioned scenarios. For each scenario, we evaluate the following eight methods, with full details on their hyperparameter tuning reported in Supplementary Material B.

M_0:KDE Direct estimation of the bivariate density. We use KDE, with Gaussian kernel and bandwidth selection based on the asympotically optimal solution proposed in Chacón *et al.* (2011), where its adequacy is shown in general settings, including Gaussian mixture models.

M_0^{NPCop}:DE Nonparametric indirect density estimation with copula. We use standard KDE with the same optimal bandwidth of Chacón *et al.* (2011) for the univariate marginals, and KDE with the transformation local likelihood estimator and nearest-neighbour bandwidth for the copula density. We refer to Nagler & Czado (2016) for details.

$M_0^{\mathbf{PCop}}$**:DE** Parametric indirect density estimation with copula. We adopt a fully parametric approach (with maximum likelihood fitting) to estimate both marginals and the copula. For the copula model, we select the best model using the AIC criterion; no misspecification is introduced for the marginals.

$M_1:k$ **NN-Eucl** Cumulative Euclidean distances from the kNNs. We sum the Euclidean distances between each point and its k neighbours defined according to the Euclidean metric. The choice of the hyperparameter k is based on an extensive cross-validation procedure, leading to a general rule of thumb aligned with the existing literature: $k = \left[\sqrt{n/2}\right]$, with $[x]$ the integer closest to x.

$M_2:k$ **NN-CDF** Cumulative CDF distances from the kNNs. We sum the CDF distances between each point and its k neighbours, which are defined according to the Euclidean metric. The measure follows a univariate approach as detailed in Section 3.2. The choice of the hyperparameter k is based on an extensive cross-validation procedure, suggesting a uniform choice across different scenarios and sample sizes: $k = 30$.

M_3:ϵ**-CDF** This represents the ϵ-neighbourhood multivariate CDF distance introduced in Section 3.3. The empirical CDF is used as the CDF estimator. The optimal choice of the hyperparameter ϵ is based on an extensive cross-validation procedure, leading to the following heuristic for S1-S16 as a result of an exponential decay model fit with respect to the sample size: $\epsilon = \exp(2.13 - 0.3\log n)$.

$M_3^{\mathbf{NPCop}}$:ϵ**-CDF** Fully nonparametric indirect estimation of the ϵ-CDF measure with copula. We use the empirical CDF for estimating the univariate marginals, and KDE with the transformation local likelihood estimator and nearest-neighbour bandwidth for the copula density. The optimal choice of the hyperparameter ϵ follows the same strategy as the previous measure, leading to the following heuristic for S1-S16: $\epsilon = \exp(1.74 - 0.26\log n)$.

$M_3^{\mathbf{PCop}}:\epsilon$ **-CDF** Parametric indirect CDF estimation with copula. We adopt a fully parametric approach (with maximum likelihood fitting) to estimate both marginals and the copula. For the copula model, we select the best model using the AIC criterion; no misspecification is introduced for the marginals. The optimal choice of the hyperparameter ϵ follows the heuristic $\epsilon = \exp(1.60 - 0.41 \log n)$ for S1-S16.

Exception made for the measures depending on ϵ (with considerations deferred to Section 4.4.1), the same hyperparameter choices are adopted in scenario S17.

4.3 Performance Metrics

The measurement of the performance of an HDR estimator can be closely related to the specific problem of interest. In a context where the interest is in detecting abnormal values, for example, estimating an HDR would allow to understand which points fall outside the normal-points region. This motivates certain metrics of common use in one-class classification problems, which we also employ in this work. However, we emphasise that the derivation of an HDR can in principle have a wider scope compared with a classification goal. Indeed, it would define the region of highest-density points (e.g. normal values), regardless of whether these points have been observed or not. Thus, it would not only allow us to classify an *observed* point as normal or abnormal, but it would also provide the entire region of normal values, useful, for example, in a prediction setting *prior to observing* a point.

Let FP, TP, FN, and TN be the number of *false positive*, *true positive*, *false negative*, and *true negative* points, respectively, where *positive* refers to those points that should be outside the true $(1 - \alpha)\%$ HDR and *negative* the others:

$$TN = \sum_{i \in s_n} \mathbb{I}(x_i \in C(f_\alpha)), \quad FN = \sum_{i \in s_n} \mathbb{I}(x_i \in \hat{C}_n(\hat{f}_\alpha) | x_i \notin C(f_\alpha)), TP = \sum_{i \in s_n} \mathbb{I}(x_i \notin C(f_\alpha)), \quad FP = \sum_{i \in s_n} \mathbb{I}(x_i \notin \hat{C}_n(\hat{f}_\alpha) | x_i \in C(f_\alpha)).$$

Well-established measures of inefficiency are false negative/positive rates (FNR/FPR), and the total error rate (ERR), that is, the one-complement of accuracy:

$$\text{FNR} = \frac{FN}{FN + TP}, \ \text{FPR} = \frac{FP}{FP + TN}, \ \text{ERR} = \frac{FN + FP}{FN + FP + TN + TP} = 1 - \text{Accuracy}.$$

To account for the potentially high imbalance between positives and negatives, we also evaluate the two-sided F1 score, and the Matthews correlation coefficient (MCC; Matthews, 1975) alternatively known in statistics as the ϕ-coefficient (see p. 282 in Cramér, 1946):

$$\text{F1} = \frac{2TP}{2TP + FP + FN} + \frac{2TN}{2TN + FP + FN}, \ \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

In particular, MCC has been shown to produce good scores only if the classification is adequate in all four elements of interest (true positives, false negatives, true negatives, and false positives), overcoming the overoptimistic inflated results, especially on imbalanced datasets, of other popular classification measures (Chicco & Jurman, 2020). All evaluations are based on $\alpha = 0.05$, that is, a coverage probability of 95%, or, alternatively stated, 5% and 95% of positives and negatives, respectively.

Table 1. *Performance results of the compared methods in selected scenarios for sample size n = 500. Data are summarised in terms of mean (standard deviation) across 1,000 independent Monte Carlo (MC) replicates.*

	Measures w/o copula				*Measures w copula*			
Metrics	M_0:*KDE*	M_1:*kNN-Eucl*	M_2:*kNN-CDF*	M_3:ϵ-*CDF*	M_0^{NPCop}:*DE*	M_0^{PCop}:*DE*	M_3^{NPCop}:ϵ-*CDF*	M_3^{PCop}:ϵ-*CDF*
Scenario S1 (Unimodal): Gaussian copula—Student's *t* marginals								
ERR	0.015 (0.006)	**0.013 (0.006)**	0.018 (0.007)	0.017 (0.006)	0.015 (0.007)	**0.009 (0.006)**	0.014 (0.006)	**0.009 (0.006)**
FPR	0.008 (0.005)	**0.007 (0.006)**	0.009 (0.007)	0.010 (0.006)	0.008 (0.006)	**0.004 (0.005)**	0.007 (0.006)	**0.005 (0.006)**
FNR	0.135 (0.092)	**0.117 (0.079)**	0.173 (0.075)	0.132 (0.089)	0.139 (0.088)	**0.073 (0.087)**	0.125 (0.086)	**0.075 (0.086)**
Accuracy	0.985 (0.006)	**0.987 (0.006)**	0.982 (0.007)	0.983 (0.006)	0.985 (0.007)	**0.991 (0.006)**	0.986 (0.006)	**0.991 (0.006)**
F1	1.842 (0.061)	**1.863 (0.063)**	1.807 (0.080)	1.826 (0.066)	1.840 (0.075)	**1.908 (0.061)**	1.854 (0.062)	**1.906 (0.061)**
MCC	0.846 (0.058)	**0.867 (0.059)**	0.811 (0.077)	0.830 (0.063)	0.844 (0.072)	**0.912 (0.055)**	0.858 (0.058)	**0.910 (0.055)**
Scenario S6 (Unimodal): Student's *t* copula—Gaussian marginals								
ERR	0.015 (0.006)	0.017 (0.006)	0.060 (0.009)	0.019 (0.006)	**0.014 (0.006)**	**0.011 (0.005)**	0.017 (0.006)	**0.011 (0.005)**
FPR	0.008 (0.005)	0.009 (0.005)	0.032 (0.006)	0.010 (0.006)	**0.007 (0.005)**	**0.006 (0.005)**	0.009 (0.005)	**0.006 (0.005)**
FNR	0.140 (0.089)	0.161 (0.094)	0.601 (0.090)	0.165 (0.092)	**0.128 (0.090)**	**0.096 (0.085)**	0.157 (0.093)	**0.095 (0.085)**
Accuracy	0.985 (0.006)	0.983 (0.006)	0.940 (0.009)	0.981 (0.006)	**0.986 (0.006)**	**0.989 (0.005)**	0.983 (0.006)	**0.989 (0.005)**
F1	1.838 (0.062)	1.816 (0.065)	1.363 (0.096)	1.804 (0.068)	**1.851 (0.059)**	**1.884 (0.058)**	1.820 (0.061)	**1.886 (0.058)**
MCC	0.843 (0.059)	0.820 (0.062)	0.365 (0.096)	0.808 (0.066)	**0.855 (0.055)**	**0.888 (0.053)**	0.825 (0.059)	**0.890 (0.053)**
Scenario S11 (Bimodal): Frank copula—Gaussian & Gaussian mixture marginals								
ERR	0.019 (0.007)	0.016 (0.006)	0.061 (0.009)	0.018 (0.006)	0.015 (0.006)	**0.011 (0.006)**	**0.015 (0.006)**	**0.011 (0.006)**
FPR	0.010 (0.005)	0.008 (0.005)	0.032 (0.006)	0.011 (0.006)	0.008 (0.005)	**0.006 (0.005)**	**0.008 (0.005)**	**0.006 (0.005)**
FNR	0.175 (0.095)	0.144 (0.093)	0.608 (0.092)	0.150 (0.093)	0.139 (0.093)	**0.096 (0.089)**	**0.136 (0.093)**	**0.096 (0.089)**
Accuracy	0.981 (0.007)	0.984 (0.006)	0.939 (0.009)	0.982 (0.006)	0.985 (0.006)	**0.989 (0.006)**	**0.985 (0.006)**	**0.989 (0.006)**
F1	1.800 (0.068)	1.832 (0.064)	1.355 (0.097)	1.810 (0.065)	1.837 (0.061)	**1.882 (0.059)**	**1.841 (0.061)**	**1.883 (0.059)**
MCC	0.804 (0.066)	0.836 (0.061)	0.357 (0.097)	0.815 (0.063)	0.842 (0.058)	**0.887 (0.055)**	**0.845 (0.058)**	**0.887 (0.054)**
Scenario S16 (Quadrimodal): Clayton copula—Gaussian mixture marginals								
ERR	0.029 (0.007)	0.027 (0.008)	0.072 (0.010)	0.032 (0.008)	**0.021 (0.007)**	**0.014 (0.006)**	0.024 (0.007)	**0.014 (0.006)**
FPR	0.015 (0.006)	0.014 (0.006)	0.038 (0.005)	0.019 (0.007)	**0.011 (0.006)**	**0.007 (0.006)**	0.013 (0.006)	**0.008 (0.006)**
FNR	0.283 (0.092)	0.256 (0.093)	0.720 (0.088)	0.276 (0.095)	**0.197 (0.089)**	**0.126 (0.089)**	0.226 (0.088)	**0.129 (0.089)**
Accuracy	0.971 (0.007)	0.973 (0.008)	0.928 (0.010)	0.968 (0.008)	**0.979 (0.007)**	**0.986 (0.006)**	0.976 (0.007)	**0.986 (0.006)**
F1	1.689 (0.080)	1.716 (0.081)	1.238 (0.089)	1.670 (0.083)	**1.777 (0.074)**	**1.850 (0.067)**	1.748 (0.077)	**1.848 (0.069)**
MCC	0.693 (0.079)	0.720 (0.080)	0.239 (0.090)	0.675 (0.082)	**0.782 (0.072)**	**0.855 (0.064)**	0.752 (0.075)	**0.852 (0.066)**

4.4 Simulation Results

Comparative performance results are reported in Table 1 in terms of their mean (standard deviation) across a number of 1,000 independent Monte Carlo (MC) replicates. We primarily focus on discussing four scenarios that are representative of the different copula-induced dependencies and multimodalities, specifically those on the diagonal of Figure 3 (S1, S6, S11 and S16), and the sample size $n = 500$; all the other scenarios and sample sizes are deferred to Supplementary Material C. The Dirichlet case is covered in Section 4.4.1.

In general, copula-based approaches result in the most performing measures, with classification errors (ERR, FPR and FNR) uniformly smaller than those of the other measures and at no cost in terms of variability. Their advantage is particularly interesting in more complex scenarios, such as the quadrimodal case, where the parametric copula approach (both M_0^{PCop}:DE and $M_3^{\textbf{PCop}}$:$\boldsymbol{\epsilon}$**-CDF**) practically halves the total error rate (ERR) of a standard M_0:KDE. Although the difference may be considered relatively negligible when looking at the ERR (maximum difference of 0.015 in S16) and the FPR (maximum distance of 0.008 in S16), it plays a significant role for the FNR, with an error difference of 0.157. In practice, the probability of a *true positive* being missed by the 'test' or measure decreases from around 28% (M_0**:KDE**; S16) to around 13% (M_0^{PCop}**:DE**, $M_3^{\textbf{PCop}}$:$\boldsymbol{\epsilon}$**-CDF**; S16). This aspect is well-captured by the alternative performance metrics of F1 and MCC, which offer a more reliable global measure of efficiency compared with the ERR or the Accuracy. As shown in Table 1—S16, when comparing M_0**:KDE** to $M_0^{\textbf{PCop}}$**:DE** or $M_3^{\textbf{PCop}}$:$\boldsymbol{\epsilon}$**-CDF**, the MCC, for example, increases from 0.693 to more than 0.85. Note that MCC varies from -1 (worst value) to 1 (best value).

When comparing the non-copula based measures, no substantial difference is noticed between M_0**:KDE**, M_1:k**NN-Eucl** and M_3:$\boldsymbol{\epsilon}$**-CDF**, with a slightly improved performance of the second. In particular, M_1:k**NN-Eucl** provides the only exception to the uniform advantage of copula-based approaches. This occurs in S1 for $n = 50$, where M_1:k**NN-Eucl** shows an enhancement, although negligible (results are given in Supplementary Material C—Table 2 and Figure 31). Interestingly, S1 represents a scenario with heavy-tailed marginals, that is, Student's t distribution with $v = 2$, which translates into a substantially wider 95% HDR (see Figure 3). Relatively good results are also achieved in the other heavy-tailed cases (S5, S9 and S13), and only in small samples (in particular $n = 50$; see Figure 31 in the supporting information). Outside these scenarios, compared with the classical M_0**:KDE**, M_1:k**NN-Eucl** has typically superior performances in more complex cases, in particular, in all quadrimodal distributions (S4, S8, S12 and S16), and for smaller sample sizes ($n = 50$ and $n = 100$). As the sample size increases ($n > 100$), M_0**:KDE** shows some improvements, especially in simpler settings with Gaussian marginals (result are given in Supplementary Material C).

The worst behaviour is shown by M_2:k**NN-CDF** (with FNR being as high as 72% in S16). Note that this was expected as M_2:k**NN-CDF** follows a univariate rationale, motivating therefore the multivariate M_3:$\boldsymbol{\epsilon}$**-CDF** proposal (see Section 3).

4.4.1 Simplex scenario: Dirichlet

A particular interest is dedicated to the simplex scenario, which plays an important role as the sample space of compositional data. Compositional data quantitatively describe parts of some whole and consist of vectors of positive components subject to a unit-sum constraint (Aitchison, 1982). Measurements involving proportions, probabilities, or percentages can all be thought of as compositional data. These commonly arise in many disciplines; for example, in demography, cause-specific mortality rates can be studied by considering them as compositions, where no single rate is free to vary separately from the rest of the rate composition(see, e.g. Stefanucci & Mazzuco, 2022). This induces a particular and unique dependence structure (represented, e.g. by the correspondent copula), which is fully entangled in the whole system.

The Dirichlet distribution represents a natural candidate for analysing compositional data, as its support is the simplex. Let $X \sim \mathcal{D}(\alpha_1, \alpha_2, \alpha_3)$ denote a Dirichlet random vector defined on the 2-dimensional simplex, where $\alpha_j > 0$, for $j = 1, 2, 3$; we refer to chapter XI in Devroye (1986) for a detailed exposition. Up to a normalising constant, the density of X is given by

$$f(x_1, x_2; \alpha_1, \alpha_2, \alpha_3) \propto x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{\alpha_3 - 1}, \quad x_1, x_2 \in [0, 1]; x_1 + x_2 \leq 1.$$

In the 2-dimensional simplex, when $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = a$, the associated copula density has the following expression up to a normalising constant:

$$c(u_1, u_2; \alpha_1 = 1, \alpha_2 = 1, \alpha_3 = a) \propto \frac{\left[(1 - u_1)^{\frac{1}{a+1}} + (1 - u_2)^{\frac{1}{a+1}} - 1\right]^{a-1}}{[(1 - u_1)(1 - u_2)]^{\frac{a}{a+1}}},$$

with u_1, u_2 such that $u_1, u_2 \in [0, 1]$ and $(1 - u_1)^{\frac{1}{a+1}} + (1 - u_2)^{\frac{1}{a+1}} \geq 1$. The analytical

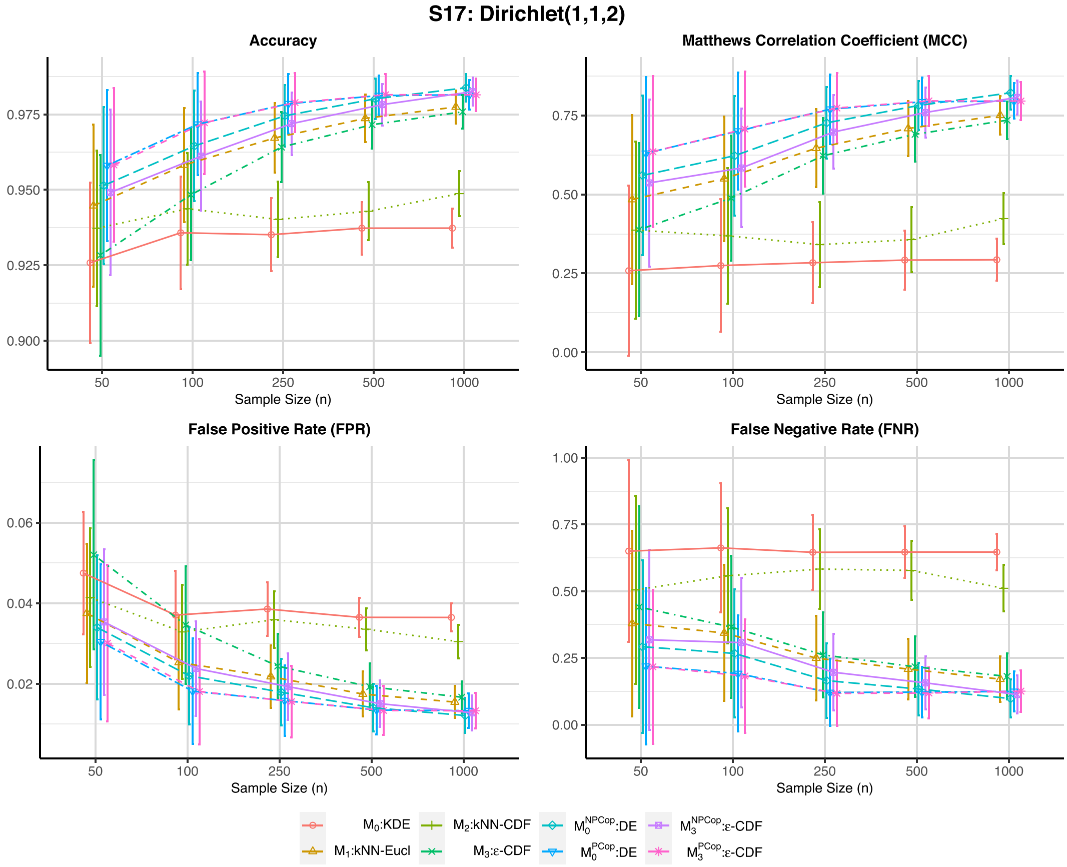

FIGURE 5. *Performance results of the compared methods in the Dirichlet scenario for varying sample sizes. Data are summarised in terms of mean and error bounds across 1,000 independent Monte Carlo (MC) replicates.*

derivation is given in Supplementary Material A, while its graphical representation—in terms of a set of random draws and different level sets—is illustrated in Figure 4.

To evaluate the proposed measures in this specific scenario, one first needs to perform an accurate tuning of their hyperparameters. As shown in Supplementary Material B, the optimal choice of k—which represents the optimal number of neighbours to take into account for computing both M_1:k**NN-Eucl** and M_2:k**NN-CDF**—remains the same as for the other scenarios: $k = \left[\sqrt{n/2}\right]$. However, for measures depending on ϵ—which defines the length, area, or volume of the neighbourhood—the optimal choice depends on the support of the underlying variable. Being defined on a simplex, in a Dirichlet scenario, the optimal values of ϵ have a smaller magnitude compared with noncompact or less restrictive scenarios such as S1-S16. In this case, the following choices are made for ϵ, guided by simulation studies reported in Supplementary Material B:

M_3:ϵ**-CDF** Performances are robust to sample size, with the empirical optimal $\epsilon = 0.10$.

M_3^{NPCop}:ϵ**-CDF** Performances depend on the sample size, according to a heuristic given by: $\epsilon = \exp(-1.22 - 0.23\log n)$. This relationship is obtained by fitting a nonlinear regression (exponential decay model), with the empirical optimal ϵ and the sample size n as dependent and independent variables, respectively.

M_3^{PCop}:ϵ**-CDF** Performances are robust to sample size, with the empirical optimal $\epsilon = 0.02$.

In terms of results, as suggested in Table 1, the more complex the scenario, the more difficult one should expect it to be to identify the highest-density points (or *true negatives*) versus the *true positives*. In this particular case, as shown in Figure 5, the discrepancy between the different measures is remarkable, with persistently high FNR values, even when the sample size increases, for M_0**:KDE** and M_2:k**NN-CDF**. All other measures improve with the sample size, achieving results comparable to the other scenarios.

4.5 MAGIC Data

We now apply the proposed measures to derive an HDR for the joint distribution of two selected variables from the MAGIC dataset (Bock *et al.*, 2004). These data simulate the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope and have been studied in classification problems (Dvořák & Savický, 2007), as well as to analyse the dependence structure of some of the characterising variables(see, e.g. Grazian *et al.*, 2022; Nagler & Czado, 2016).

In this evaluation, we focus on gamma-ray observations (overall $n = 12,332$) and consider the two variables ‘fConc1’ and ‘fM3Long’, after scaling them. We refer to Bock *et al.* (2004) for a full description of the dataset. In this case (as deduced from the complex structure of the data; see Figure 6), the parametric approach is inappropriate for both the estimation of the marginal distribution and, more importantly, the copula model. Thus the 95% HDR is estimated using the nonparametric measures only. Although in the absence of the underlying truth it is not possible to perform a reliable evaluation, it seems that the two nonparametric copula-based approaches ($M_0^{\textbf{NPCop}}$**:DE** and $M_3^{\textbf{NPCop}}$:ϵ**-CDF**), as well as the distance-based M_1:k**NN-Eucl** and M_3:ϵ**-CDF**, more sensibly exclude tail data points (which may be expected to have a lower density) from the HDR.

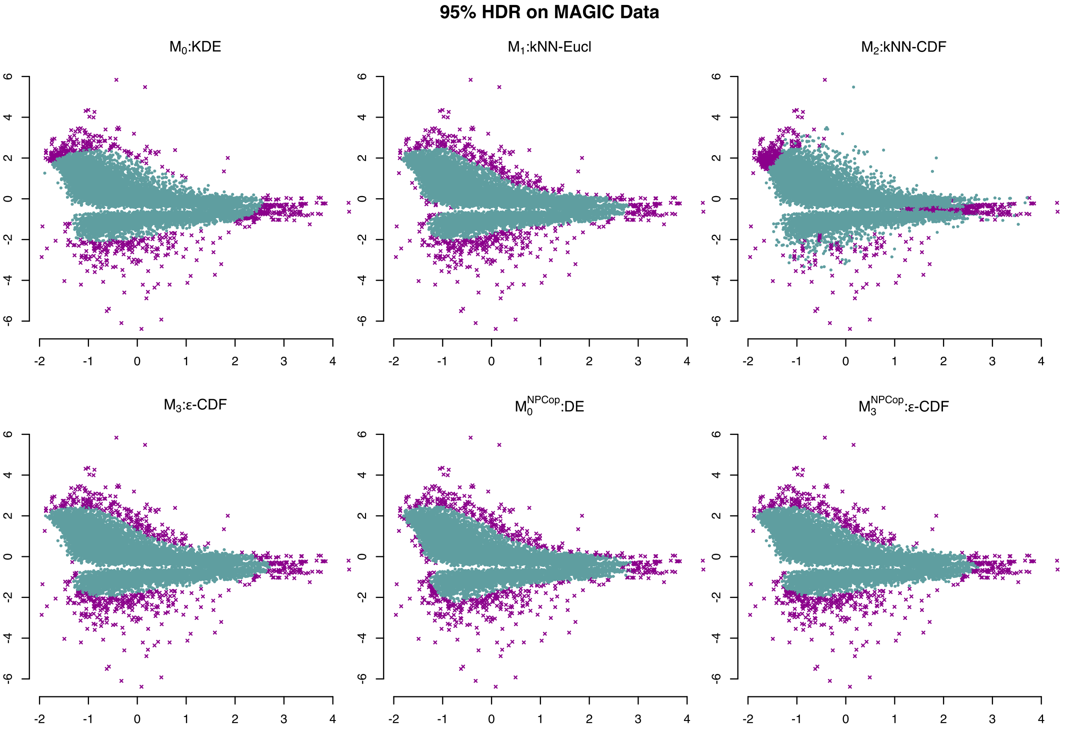

FIGURE 6. *Estimated 95% HDR of two scaled variables ('fConc1' on x-axis and 'fM3Long' on y-axis) of the MAGIC dataset. Only nonparametric measures are evaluated. Cadet-blue points define the estimated 95% HDR; in contrast, purple points are those lying outside the HDR.*

To deal with estimation and data uncertainty, we also explore the potential of a 'measure averaging' in a similar, but simplified, manner to *model averaging* (see, e.g. Hoeting *et al.*, 1999; Hjort & Claeskens, 2003). By forming a consensus between the different available measures, one would expect the resulting averaged HDR to be more robust than the individual estimates. Specifically, in Figure 7, we illustrate a 95% HDR formed by averaging across the different nonparametric measures: the estimated HDR is defined by the set of points identified as highest-density points by more than half of the evaluated metrics. Notably, when comparing the resulting HDR in Figure 7 and those obtained by the individual measures (Figure 6), we identify M_3: **ϵ-CDF** and $M_3^{\textbf{NPCop}}$: **ϵ-CDF** as the regions better resembling the average (both with $<1\%$ classification difference). Therefore, these may result in a possible superior ability to detect anomalous values in scenarios where a parametric assumption is unrealistic.

5 Discussion and Conclusions

In this work, we have discussed some generalisations of the standard density-based approach for estimating highest-density regions, using neighbourhood measures. Various measures, with distinct properties, have been introduced and compared with the classical kernel density estimator, across different scenarios and sample sizes. Furthermore, the use of copula models for representing a multivariate distribution through a combination of its marginals and their dependence structure has been investigated. Our results suggest that such a generalised approach may provide great advantages to HDR estimation when considering several alternative measures to KDE, especially those based on copulae. In fact, compared with traditional KDE,

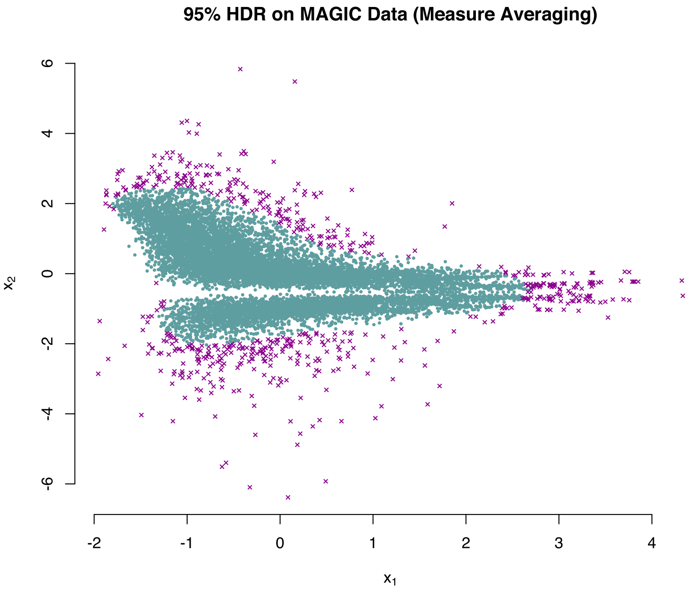

FIGURE 7. *Estimated 95% HDR based on a 'measure averaging' approach of two scaled variables ('fConc1' on x-axis and 'fM3Long' on y-axis) of the MAGIC dataset. Cadet-blue points define the estimated 95% HDR; in contrast, purple points are those lying outside the HDR.*

copula-based HDR resulted in greater accuracy and lower FPR and FNR in a number of simulation scenarios and possibly in real data (as suggested by the MAGIC data application). Such performances are particularly important when the interest is in balancing different types of errors, minimising both *false positives* and *false negatives*, and maximising therefore the probability of detecting atypical or anomalous values. In a doping detection problem, for example, this would directly translate into an enhanced ability of identifying doping abuse.

It is important to emphasise that among the various copula-based measures, the parametric ones generally outperformed the non-parametric measures. This outcome was expected because all the considered scenarios, despite their complexity, were generated from parametric models and were evaluated under the assumption of no model misspecification for the marginals. In real-world applications, one would first need to select the most appropriate family using a criterion of fit, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Notably, an important advantage of using copulae is that marginal model selection is implemented separately for each marginal, restricting the analysis to simpler univariate cases. Concerning the copula model, we account for potential misspecification, with model selection performed using the AIC (note that built-in functions are available in R software, for example, the `BiCopSelect()` function from the `VineCopula` package; Nagler *et al.*, 2023).

Relative to the scenarios and measures considered in this work, the main recommendation for the neighbourhood measure is thus a copula-based alternative. In particular, parametric variants should be preferred when distributions show a pattern traceable to a parametric family. When parametric measures do not appear appropriate for the data at hand, we would recommend the nonparametric copula-based options, with the exception of scenarios with heavy-tailed marginals. Here, $M_1 : k$ **NN-Eucl** achieves the best results, and the nonparametric $M_0^{\mathbf{NPCop}}$**:DE** measure is among the most inferior ones, especially in small samples. On the contrary, the $M_3^{\mathbf{NPCop}} : \epsilon$ **-CDF** measure using the ϵ-neighbourhood multivariate CDF distance shows promising results. Because the main characteristic of the latter is that it is based on the CDF rather than the PDF, it may suggest that focusing on estimating the former may be advantageous. Similar findings are reported by Magdon-Ismail & Atiya (2002), who acknowledge that, compared with directly estimating the PDF with, for example, KDE, approximating the CDF is less sensitive to statistical fluctuations and its convergence rate is faster than the convergence rate of KDE methods. We also emphasise that although heuristic considerations for each specific measure are made in terms of their hyperparameters, for example, the choice of k in the kNN-based approaches, further work may investigate the existence of optimal theoretical values in a similar fashion to the asymptotic works on the optimal bandwidth choice in KDE (see, e.g. Chacón *et al.*, 2011; Wand & Jones, 1994b).

In this work, we focused on estimating HDR for continuous distributions that are dominated by the Lebesgue measure. Furthermore, despite providing a general multidimensional framework for building HDRs, we evaluated the proposed approach in a bivariate context. Future lines of research may examine the problem in the underappreciated setting of discrete probability distributions and in higher dimensions. In the first case, possible connections could be made with the work of O'Neill (2022), which established some theory and an algorithm for HDRs for discrete distributions. In the second case, and more specifically with reference to copula-based measures, our aim is to explore the use of *vine copulae* (Nagler & Czado, 2016) to construct flexible dependence models for an arbitrary number of variables using only bivariate building blocks. We expect, in fact, to see remarkable advantages in using copulae over an increased number of variables, as the extension of the common KDE to high dimensions has proven challenging in terms of both computational efficiency and statistical inference.

ACKNOWLEDGEMENTS

Open access publishing facilitated by Universita degli Studi di Roma La Sapienza, as part of the Wiley - CRUI-CARE agreement.

Supporting information

Additional supporting information is provided in the supporting information. The implementation of the algorithms is possible with the R package `HDR2D` provided in the Github repository https://github.com/nina-DL/HDR2D.

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.*, **44**(2), 139–160. https://academic.oup.com/jrsssb/article/44/2/139/7027742

Bock, R.K., Chilingarian, A., Gaug, M., Hakl, F., Hengstebeck, T., Jirina, M., Klaschka, J., Kotrc, E., Savicky, P., Towers, S., Vaiciulis, A. & Wittek, W. (2004). Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instrum. Methods Phys. Res. Sect. A: Acceler., Spectromet., Detect. Assoc. Equip.*, **516**(2), 511–528.

Box, G.E.P. & Tiao, G.C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley: New York.

Cai, Y. (2010). Multivariate quantile function models. *Stat. Sin.*, **20**(2), 481–496.

Chacón, J.E., Duong, T. & Wand, M.P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Stat. Sin.*, **21**(2), 807–840.

Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**(1), 6.

Coblenz, M., Dyckerhoff, R. & Grothe, O. (2018). Nonparametric estimation of multivariate quantiles. *Environmetrics*, **29**(2), e2488.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, **13**(1), 21–27.

Cramér, H. (1946). *Mathematical Methods of Statistics*, Mathematical Methods of Statistics. Princeton, NJ, US: Princeton University Press.

Deliu, N. & Liseo, B. 2024. A Multivariate Copula-based Bayesian Framework for Doping Detection. arXiv preprint arXiv:2404.12499, https://arxiv.org/abs/2404.12499

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer New York: New York, NY.

Doss, C.R. & Weng, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electron. J. Stat.*, **12**(2), 4313–4376.

Dvořák, J. & Savický, P. (2007). Softening splits in decision trees using simulated annealing. In *Adaptive and Natural Computing Algorithms, Lecture Notes in Computer Science*, Eds. B. Beliczynski, A. Dzielinski, M. Iwanowski & B. Ribeiro, 721–729, Berlin, Heidelberg: Springer.

Figalli, A. (2018). On the continuity of center-outward distribution and quantile functions. *Nonlinear Anal.*, **177**, 413–421.

Fix, E. & Hodges, J.L. (1951). Discriminatory analysis. *Nonparam. Discrim.: Small Sample Perfor. Report A*, 193008.

Grazian, C., Dalla Valle, L. & Liseo, B. (2022). Approximate Bayesian conditional copulas. *Comput. Stat. Data Anal.*, **169**, 107417.

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biometr. J. Biometrische Zeitschrift*, **50**(5), 678–692.

Hjort, N.L. & Claeskens, G. (2003). Frequentist model average estimators. *J. Am. Stat. Assoc.*, **98**(464), 879–899.

Hjort, N.L. & Jones, M.C. (1996). Locally parametric nonparametric density estimation. *The Ann. Stat.*, **24**(4), 1619–1647.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Stat. Sci.*, **14** (4), 382–401.

Hyndman, R.J. (1995). Highest density forecast regions for nonlinear and non-normal time series models. *J. Forecast.*, **14**(5), 431–441.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *The Am. Stat.*, **50**(2), 120–126.

Kim, J.H., Fraser, I. & Hyndman, R.J. (2011). Improved interval estimation of long run response from a dynamic linear model: a highest density region approach. *Comput. Stat. Data Anal.*, **55**(8), 2477–2489.

Korpela, J., Oikarinen, E., Puolamaki, K. & Ukkonen, A. (2017). Multivariate Confidence Intervals. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 696–704. Society for Industrial and Applied Mathematics: Houston, Texas, USA.

Krishnamoorthy, K. & Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*, Wiley series in probability and statistics. Wiley: Hoboken, N.J.

Liu, H., Lafferty, J. & Wasserman, L. (2007). Sparse Nonparametric Density Estimation in High Dimensions Using the Rodeo. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pp. 283–290. PMLR.

Loftsgaarden, D.O. & Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. *The Ann. Math. Stat.*, **36**(3), 1049–1051.

Magdon-Ismail, M. & Atiya, A. (2002). Density estimation and random variate generation using multilayer networks. *IEEE Trans. Neural Netw.*, **13**(3), 497–520.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. et Biophys. Acta (BBA) - Protein Struct.*, **405**(2), 442–451.

Meeker, W.Q., Hahn, G.J. & Escobar, L.A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*, Second edition, Wiley series in probability and statistics. Wiley: Hoboken, New Jersey.

Munoz, A. & Moguerza, J.M. (2006). Estimation of high-density regions using one-class neighbor machines. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**(3), 476–480.

Nagler, T. & Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *J. Multivar. Anal.*, **151**, 69–89.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B. & Erhardt, T. 2023. Vinecopula: Statistical inference of vine copulas. https://github.com/tnagler/VineCopula, R package version 2.5.0.

Nelsen, R.B. (2006). *An Introduction to Copulas*, 2nd ed, Springer series in statistics. Springer: New York.

O'Neill, B. (2022). Smallest covering regions and highest density regions for discrete distributions. *Comput. Stat.*, **37** (3), 1229–1254.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Ann. Math. Stat.*, **33**(3), 1065–1076.

Ramaswamy, S., Rastogi, R. & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, **29**(2), 427–438.

Saavedra-Nieves, P. (2022). Nonparametric estimation of highest density regions for COVID-19. *J. Nonparam. Stat.*, **34**(3), 663–682.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. & Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.*, **13**(7), 1443–1471.

Silverman, B.W. (1996). *Density Estimation for Statistics and Data Analysis*, Monographs on statistics and applied probability. Chapman & Hall/CRC: Boca Raton.

Sklar, M. (1959). Fonctions de répartition á N dimensions et leurs marges. *Annales de l'ISUP*, **VIII**(3), 229–231.

Sottas, P.-E., Baume, N., Saudan, C., Schweizer, C., Kamber, M. & Saugy, M. (2007). Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio. *Biostatistics*, **8**(2), 285–296. https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxl009

Stefanucci, M. & Mazzuco, S. (2022). Analysing cause-specific mortality trends using compositional functional data analysis. *J. Royal Stat. Soc. Ser. A: Stat. Soc.*, **185**(1), 61–83. https://academic.oup.com/jrsssa/article/185/1/61/7068446

Steinwart, I., Hush, D. & Scovel, C. (2005). A classification framework for anomaly detection. *J. Mach. Learn. Res.*, **6** (8), 211–232.

Terrell, G.R. & Scott, D.W. (1992). Variable kernel density estimation. *The Ann. Stat.*, **20**(3), 1236–1265.

Turkkan, N. & Pham-Gia, T. (1993). Computation of the highest posterior density interval in Bayesian analysis. *J. Stat. Comput. Simul.*, **44**(3-4), 243–250.

Venturini, M. (2015). Statistical distances and probability metrics for multivariate data, ensembles and probability distributions. PhD Thesis, Universidad Carlos III de Madrid.

WADA (2021). The World Anti-Doping Code, World Anti-Doping Agency https://www.wada-ama.org/en/what-we-do/world-anti-doping-code

Wand, M.P. & Jones, M.C. (1994a). *Kernel Smoothing*. Chapman and Hall/CRC.

Wand, M.P. & Jones, M.C. (1994b). Multivariate plug-in bandwidth selection. *Comput. Stat.*, **9**(2), 97–116.

[*Received November 2023; accepted July 2024*]