entropy

Article

On the Accurate Estimation of Information-Theoretic Quantities from Multi-Dimensional Sample Data

Manuel Álvarez Chaves [1,*], Hoshin V. Gupta [2], Uwe Ehret [3] and Anneli Guthke [1]

[1] Stuttgart Center for Simulation Science, Cluster of Excellence EXC 2075, University of Stuttgart, 70569 Stuttgart, Germany
[2] Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ 85721, USA
[3] Institute of Water and River Basin Management, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
* Correspondence: manuel.alvarez-chaves@simtech.uni-stuttgart.de

Abstract: Using information-theoretic quantities in practical applications with continuous data is often hindered by the fact that probability density functions need to be estimated in higher dimensions, which can become unreliable or even computationally unfeasible. To make these useful quantities more accessible, alternative approaches such as binned frequencies using histograms and *k*-nearest neighbors (*k*-NN) have been proposed. However, a systematic comparison of the applicability of these methods has been lacking. We wish to fill this gap by comparing kernel-density-based estimation (KDE) with these two alternatives in carefully designed synthetic test cases. Specifically, we wish to estimate the information-theoretic quantities: entropy, Kullback–Leibler divergence, and mutual information, from sample data. As a reference, the results are compared to closed-form solutions or numerical integrals. We generate samples from distributions of various shapes in dimensions ranging from one to ten. We evaluate the estimators' performance as a function of sample size, distribution characteristics, and chosen hyperparameters. We further compare the required computation time and specific implementation challenges. Notably, *k*-NN estimation tends to outperform other methods, considering algorithmic implementation, computational efficiency, and estimation accuracy, especially with sufficient data. This study provides valuable insights into the strengths and limitations of the different estimation methods for information-theoretic quantities. It also highlights the significance of considering the characteristics of the data, as well as the targeted information-theoretic quantity when selecting an appropriate estimation technique. These findings will assist scientists and practitioners in choosing the most suitable method, considering their specific application and available data. We have collected the compared estimation methods in a ready-to-use open-source Python 3 toolbox and, thereby, hope to promote the use of information-theoretic quantities by researchers and practitioners to evaluate the information in data and models in various disciplines.

Keywords: information theory; non-parametric estimation; entropy; mutual information; Kullback–Leibler divergence; relative entropy; data; binning; kernel density estimation; *k*-nearest neighbors; *k*-NN

check for updates

Citation: Álvarez Chaves, M.; Gupta, H.V.; Ehret, U.; Guthke, A. On the Accurate Estimation of Information-Theoretic Quantities from Multi-Dimensional Sample Data. *Entropy* **2024**, *26*, 387. https://doi.org/10.3390/e26050387

Academic Editor: Daniel Keren

Received: 28 March 2024
Revised: 24 April 2024
Accepted: 25 April 2024
Published: 30 April 2024

1. Introduction

1.1. The Promise of Information Theory

Entropy, mutual information, and Kullback–Leibler (KL) divergence are fundamental concepts of information theory [1]. Originally introduced in the field of communication [2], information theory has now found uses in a diverse set of disciplines, including artificial intelligence [3], Earth and environmental science [4], experimental design [5], neuroscience [6], and finance and economics [7]. Its wide-ranging applications stem from its solid foundation in probability theory. By analyzing the probability distributions associated with the variables in a given problem, information theory can determine the nature and extent of their relationships. These relationships may exhibit linearity or non-linearity, depending on the specific system involved [8]. In essence, information theory can unveil the intricate connections hidden within complex systems.

1.2. Definition of Information-Theoretic Quantities

For the initial definitions and notation, we follow Cover and Thomas [1] and MacKay [3]. In the discrete case, consider x the outcome of a random variable X, which takes on one of the set of possible values in the alphabet $\mathcal{A}_X = \{a_1, a_2, \ldots, a_n\}$ having probabilities $\mathcal{P}_X = \{p_1, p_2, \ldots, p_n\}$ with the probability mass function (PMF) $P(X = a_i) = p_i$, $p_i \geq 0$, and $\sum_{a_i \in \mathcal{A}_X} P(X = a_i) = 1$. An example of an alphabet is the 27 characters in a random English document (letters from a to z and the space character) [3].

The information content of the outcome x is defined as:

$$h(x) = -\log_2 P(x), \tag{1}$$

with the unit of "bit" due to the base two of the logarithm. The *entropy* $H(X)$ of the random variable X is defined to be the average information content of every possible outcome. Therefore:

$$H(X) = -\sum_{x \in \mathcal{A}_X} P(x) \log_2 P(x). \tag{2}$$

The relative entropy or *KL divergence* D_{KL} between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is:

$$D_{KL}(P\,||\,Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}. \tag{3}$$

A simple interpretation of the KL divergence is a measure of the inefficiency of assuming that the distribution is Q when the true distribution is P.

Introducing a second outcome y of a random variable Y from a different alphabet $\mathcal{A}_Y$, the joint entropy of the two random variables X and Y is:

$$H(X,Y) = -\sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log_2 P(x,y). \tag{4}$$

Entropy is additive only for independent random variables:

$$H(X,Y) = H(X) + H(Y) \iff P(x,y) = P(x)P(y). \tag{5}$$

From this result, *mutual information* $I(X;Y)$ can be introduced as the KL divergence between the joint PMF $P(x,y)$ and the product of the marginal PMFs $P(x)$ and $P(y)$:

$$\begin{aligned} I(X;Y) &= D_{KL}(P(x,y)\,||\,P(x)P(y)) \\ &= \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \end{aligned} \tag{6}$$

Mutual information is positive and symmetric, i.e., $I(X;Y) = I(Y;X)$ and $I(X;Y) = 0$, only when Equation (5) is true. Using the rules of conditional probabilities, namely $P(x,y) = P(y)P(x\,|\,y)$, one arrives at the more common definition for mutual information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{7}$$

As per Equation (7), mutual information measures the average reduction in uncertainty about X that results from the knowledge of Y.

In his seminal paper, Shannon [2] introduced the concept of differential entropy specifically for continuous random variables. In this context, consider a random variable X with values x in $\mathbb{R}^d$ space and a probability density function (PDF) $p(x)$ whose support is a set $\mathcal{X}$. Its differential entropy is given by Equation (8):

$$H(X) = -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \tag{8}$$

In this equation, we have made some changes in the notation to emphasize our interest in the continuous setting. The alphabet $\mathcal{A}_X$ becomes the support set $\mathcal{X}$. PMFs denoted as P become PDFs denoted as p. The $\log_2$ is substituted simply by the log, which is understood to be the natural logarithm, also changing the units in which information is measured from "bit" to "nat" or the natural unit of information and further differentiating between the discrete and continuous case. One bit is equivalent to $\log(2)$ nats.

Similarly, the equations for mutual information and KL divergence can also be adapted for continuous applications (Equations (9) and (10), respectively):

$$D_{KL}(p||q) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{9}$$

$$I(X;Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x,y) \log\left(\frac{p(x,y)}{p(x)\, p(y)}\right) \mathrm{d}x \, \mathrm{d}y \tag{10}$$

Note that Equation (9) requires the support of $q(x)$ to be equal to or larger than the support of $p(x)$ because the integral is over the support of the latter.

As will be described later in Section 1.3, the definition of differential entropy for continuous variables simply replaces a sum with an integral, but this is not a well-defined operation. For the cases of KL divergence and mutual information, the issue of discretization is partially avoided because the argument in the logarithm is a ratio of probabilities over the same space [3], but this is not the case for differential entropy.

1.3. The Challenge of Estimating Information-Theoretic Quantities

Beirlant et al. [9] provide a comprehensive review of various common approaches to estimating differential entropy. In this section, we will follow their classification while noting that these methods can also be extended to the cases of KL divergence and mutual information.

Generally, estimation methods using sample data can be divided into three categories. First are plug-in estimates, which can be further divided into resubstitution and integral estimates. Resubstitution estimates calculate a density for each data point in a sample, and then, the integration term in Equations (8)–(10) is exchanged for a sum over densities. Integral estimates use a density representation of the total sample to perform numerical integration of Equations (8)–(10). Plug-in estimates are often associated with *kernel density estimation* and will be explained further in Section 2.1.

Second are estimates based on sample-spacings, or histograms as they are more commonly known and from this time forward referred to as *binning*. They share the density estimation step used in plug-in estimates, but differ in some other aspects, which will be explored further in Section 2.2.

It becomes obvious that these two categories of estimation methods involve an initial step of density estimation from samples [10] before computing the desired information-theoretic quantities. However, density estimation itself poses challenges [11,12], and common techniques of density estimation often perform poorly on high-dimensional data ($d > 3$). Kernel density methods are often used to address high-dimensional problems, but Joe [13] studied these methods for entropy estimation, concluding that they perform well only when the number of dimensions of the multivariate data is small.

The third category, namely estimates based on *nearest neighbor* distances, differ from the previous two methods in that they do not require an explicit density estimate. Density can be estimated using distance-based methods [14], but this is not necessary in this case.

Hence, they promise to be relatively straightforward to implement and robust, which makes them a good candidate to compare with more traditional density-based estimates. Nearest neighbor distance-based estimates will be detailed further in Section 2.3.

While the literature for each of the listed methods has been well established [9,15] and more recent approaches to density estimation have proven to perform well in higher dimensions [16], in practice, we often find binning to be the most commonly adopted technique [17–20] without further justification. We hypothesize that this preference for a single method with known deficiencies, at least in higher dimensions, is due to the lack of a systematic comparison of methods, and due to a perceived initial hurdle to implementation.

1.4. Contribution and Outline of This Study

In this study, we compare kernel density estimation, binning, and the nearest neighbor approaches for estimating information-theoretic quantities in a practical setting. In a comprehensive and systematic effort, we compare these three most widely recognized non-parametric estimation methods in terms of (1) their theoretical derivation, (2) their sample size requirements, and (3) their accuracy. While our focus primarily centers on estimating entropy, KL divergence, and mutual information due to their foundational significance, it is important to acknowledge that additional measures such as conditional entropy can be derived from these fundamental quantities. To this end, we have designed a set of eight synthetic scenarios to test each estimation method on data coming from different uni- and multi-variate distributions, ranging up to ten dimensions. Usually, when a method is introduced, there is a comparison of its performance against other methods; however, typically, this comparison is limited to two particular algorithms [21], and often, the comparison is performed between algorithms of the same family, e.g., *k*-NN [22]. Also, the comparison of different methods is typically challenging due to the absence of a standardized set of practices across cases [23], making it difficult to ensure comparability between them.

Finally, we provide a free toolbox that collects these methods and test cases, to promote the use of these concepts in a wide range of scientific and practical applications.

The remainder of this article is structured as follows: In Section 2, we introduce the non-parametric estimation methods studied in this article, discuss their properties and hyperparameters, and describe how they are applied for the computation of the information-theoretic quantities of interest. Section 3 first presents the general design of the test cases in which the estimators are evaluated. Each test case is then introduced with a description of the specific probability distribution we sample from, and closed-form solutions to the quantities of interest, if available. The results are presented and analyzed with the help of "evaluation matrix plots", which show the absolute and relative performance of each estimator across the different quantities. Finally, in Section 4, we summarize our findings and discuss how they may guide future applications of information theory in domain-specific workflows.

2. Definition and Theoretical Description of Estimators

This section describes the three non-parametric estimation methods investigated in this study and discusses specific aspects for each of them.

2.1. Kernel Density Estimation (KDE)

Introduced as the Parzen–Rosenblatt window method [24,25], the KDE method consists of estimating a PDF based on kernels as weights, with the kernel being a non-negative window function. The density $p(x)$ at a point x is estimated as:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(u), \tag{11}$$

where

$$u = \frac{(x - x_i)^\intercal \Sigma^{-1} (x - x_i)}{h^2} \tag{12}$$

and n is the total number of samples, K is a multivariate kernel function, $x_i = [x_{1,i}, x_{2,i}, \ldots, x_{d,i}]^\intercal$ is a d-dimensional vector of samples, Σ is the covariance matrix of the samples, and h is a smoothing parameter. For the results presented further in this paper, we chose to always use a multivariate Gaussian kernel function and the Silverman bandwidth estimate, as suggested by Moon et al. [26].

The multivariate Gaussian kernel takes the form:

$$K(u) = \frac{1}{(2\pi)^{d/2} h^d \det(\Sigma)^{1/2}} e^{-u/2}, \tag{13}$$

where u was previously defined in Equation (12) and h is Silverman's bandwidth estimate [11], given by:

$$h = \left(\frac{n(d+2)}{4} \right)^{-1/(d+4)}. \tag{14}$$

Using the density estimate from Equation (11), entropy, mutual information, and KL divergence can be calculated directly as:

$$H(X) = -\frac{1}{n} \sum_{i=1}^{n} \log\left(\hat{p}(x_i)\right) \tag{15}$$

$$D_{KL}(\hat{p}||\hat{q}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\hat{p}(x_i)}{\hat{q}(x_i)}\right) \tag{16}$$

$$I(X;Y) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\hat{p}(x_i, y_i)}{\hat{p}(x_i)\hat{p}(y_i)}\right) \tag{17}$$

In this form, these can be called resubstitution estimates of each quantity. Additionally, if numerical integration using the kernel PDF is used to approximate the result of Equations (8)–(10), this is known as an integral estimate [9]. For example, Equation (8) would be written as:

$$H(X) = -\int_{\mathcal{X}} \hat{p}(x) \log \hat{p}(x) \, \mathrm{d}x \tag{18}$$

Both resubstitution and integral estimates are plug-in estimates.

2.2. Binning

The binning method relies on obtaining an estimate of the density using a histogram [27,28]. Given an origin at x_0 and a bin width of delta (Δ), the bins of the histogram are set to the intervals $[x_0 + m\Delta, x_0 + (m+1)\Delta]$ for a set number of positive and negative integers of m. Then, considering the total number of observations n and the number of observations c_i in the same bin as x_i, the frequency estimate at x_i in the histogram can be written as:

$$\hat{f}(x_i) = \frac{c_i}{n\Delta}. \tag{19}$$

In this case, we have chosen to use f instead of p to distinguish between a frequency count and a probability density estimate. To obtain probabilities from this frequency estimate, one must account for the specified bin width $p(x) = \Delta f(x)$. In Equation (19), the parameter Δ controls the amount of smoothing applied to the frequency estimate, similar to the bandwidth used in Section 2.1. To ease the readability, we have chosen to show equations that use a bin spacing that is uniform, but every bin could have a particular size, independent of the rest (Δ_i).

In terms of probability, we follow the notation of Cover and Thomas [1] in which X^{Δ} represents the quantized (binned) version of a random variable. Then, following their derivation, the entropy of this quantized version can be written as:

$$H(X^{\Delta}) = -\sum \Delta \hat{f}(x_i) \log \hat{f}(x_i) - \sum \hat{f}(x_i) \Delta \log \Delta, \tag{20}$$

where $\sum \Delta f(x) = \int p(x) = 1$. The second term on the right-hand side of Equation (20) is a correction factor due to the bin spacing of the chosen quantization of x.

As KL divergence and mutual information evaluate ratios between distributions, this correction factor cancels out, and then, their expressions can be written similarly to Equations (16) and (17):

$$D_{KL}(\hat{f}||\hat{g}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\hat{f}(x_i)}{\hat{g}(x_i)}\right) \tag{21}$$

$$I(X;Y) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\hat{f}(x_i, y_i)}{\hat{f}(x_i)\hat{f}(y_i)}\right) \tag{22}$$

Here, the second PDF $\hat{g}(x_i)$ is estimated using the same binning scheme as $\hat{f}(x_i)$.

Even though the correction factor cancels out for the previous two equations, the final result does depend on the particular binning scheme chosen. A particular choice of binning controls the trade-off between a resulting histogram that has too much detail ("undersmoothing") or a histogram that has too little detail ("oversmoothing") with respect to the true distribution. The resulting estimate of the KL divergence or mutual information, therefore, also depends on the selection of a particular binning scheme. Because of this reason, several rules-of-thumb have been developed over time for the selection of an "optimal" bin width or binning scheme.

2.2.1. Rules-of-Thumb for Bin Width Selection

For the purpose of obtaining a good representation of the underlying distribution of the data, there are several methods to estimate an adequate number of bins or, more specifically, the bin width Δ to be used when building a histogram. Some rules-of-thumb exist to estimate Δ for one-dimensional data. In the context of multi-dimensional data, methods for optimal bin width estimation have been proposed, but they often require the use of combinatorial methods [29] or the solution of an optimization problem [30]. Therefore, we chose the simpler approach of estimating Δ independently for each dimension of the data and built a multi-dimensional histogram based on the estimated Δs, meaning that binning is uniform in each dimension, but not across dimensions.

Sturges' rule: The width of each bin is the base two logarithm of the number of samples in the data (n), $\Delta = \log_2(n) + 1$. With this estimate of the number of bins, there is an inherent assumption that the data follows a normal distribution [31].

Scott's rule: The width of each bin is proportional to the standard deviation (σ) of the data, and inversely proportional to the cube root of the number of samples (n), $\Delta = 3.49\,\sigma n^{-1/3}$. Although there is still an assumption that the data follow a normal distribution, this assumption is not as strong as with Sturges' rule [27].

Freedman and Diaconis' rule: The width of each bin is proportional to the interquartile range (IQR) of the data, and inversely proportional to the number of samples (n), $\Delta = 2\,\mathrm{IQR}\ /\ n^{1/3}$. Although this is similar to Scott's rule, in using the IQR, this estimator for the bin width is more robust to outliers of non-normal distributions [32].

2.2.2. The Quantile Spacing Approach

Taking a different perspective on optimal binning for estimating entropy from a sample, Gupta et al. [10] recently introduced the Quantile Spacing (QS) approach for the case where X is a one-dimensional continuous random variable and the mathematical form for the distribution of the data-generating process is unknown. The approach is based in the assumption that the PDF can be approximated as piecewise constant on the intervals between quantile locations where the values for the quantiles have been determined from the sample S.

The method assumes that the PDF can be approximated as piecewise constant on the intervals between quantiles $Z = \{z_0, z_1, z_2, \ldots, z_{N_Z}\}$, where N_Z represents the total number of quantiles and $z_0 = x_{min}$ and $z_{N_Z} = x_{max}$. Using the quantiles Z, the PDF can be approximated as:

$$p(x) \approx \hat{p}(x|Z) = p_{j-1}^{j} = \frac{K}{\Delta_j} \text{ for } z_{j-1} \leq X \leq z_j \text{ and } j = 1, \ldots, N_Z, \tag{23}$$

where $\Delta_j = z_j - z_{j-1}$ and $K = \frac{1}{N_Z}$. From this definition of the quantiles, the entropy estimate is given by the sum of the individual entropies of each uniform distribution across quantiles. More specifically, the estimate depends on the logs of the spacings between quantiles and is defined by the average of these values:

$$\hat{H}_{\hat{p}}(X|Z) = \frac{1}{N_Z} \cdot \sum_{i=1}^{N_Z} \log\left(N_Z \cdot \Delta_j\right) \tag{24}$$

To determine the empirical quantiles, N_k sample subsets are sampled without replacement from the full sample S, each of size $N_Z - 1$. These subsamples are then sorted to obtain an estimate of the quantile locations, thereby obtaining N_k estimates of each quantile z_j. Finally, the average of the value of the location is taken to obtain $\hat{Z} = \{x_{\min}, \hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{N_Z - 1}, x_{\max}\}$.

As described, the QS approach has two hyperparameters: N_Z and N_k. The choice of N_Z is suggested to be 25% of the total number of points in the sample N_S, as it minimizes the bias in the estimate for the normal, exponential, and log-normal distributions [10]. N_k is recommended to be a large number as a greater number of subsamples and repetitions more accurately estimate the correct value for each quantile, with $N_k = 500$ found to be optimal [10].

2.3. *k*-Nearest Neighbors (*k*-NN)

The k-nearest neighbors (k-NN) algorithm is a non-parametric supervised learning method originally developed for classification and regression [33]. It aims at local function approximation by assuming similarity between neighboring sample points. Since the distance of a sample point to its k-nearest neighbors can be interpreted as a local density estimate, it is not a surprise that also k-NN-based estimates of information-theoretic quantities have been proposed. However, the charm of these estimates is precisely that they do *not* require an explicit evaluation of a probability density estimate.

The k-NN-based estimator for entropy was introduced by Kozachenko and Leonenko [34] and serves as the basis for k-NN-based estimations of KL divergence and mutual information:

$$\hat{H}(X) = \psi(N) - \psi(k) + \log(c_1(d)) + \frac{d}{N} \sum_{i=1}^{N} \log(\rho_k^d(i)), \tag{25}$$

where ψ is the digamma function defined as the logarithmic derivative of the gamma function $(\frac{\mathrm{d}}{\mathrm{d}z} \log(\Gamma(z)))$. N is the total number of points in the sample. k is a hyperparameter specifying the number of nearest neighbors used in the estimate. $c_1(d)$ is the volume of a d-dimensional unit ball with d being the number of dimensions of the sample. $\rho_k^d(i)$ is the distance between x_i and its kth nearest neighbor. In calculating the distance, the length between two points (x and y) is given by a p-norm function, where $p \geq 1$, as follows:

$$\|x - y\|_p = \left(|x_1 - y_1|^p + |x_2 - y_2|^p + \cdots + |x_n - y_n|^p\right)^{\frac{1}{p}}, \tag{26}$$

with a suggested $p = 2$, i.e., using the Euclidean norm for the entropy estimate [34].

Kozachenko and Leonenko [34] demonstrated that the suggested estimator has square consistency for any number of dimensions, meaning that, as the size of the sample increases, the mean-squared error of the estimate tends to 0:

$$\lim_{n\to\infty} \mathbb{E}\left\{ \left(\hat{H} - H(f)\right)^2 \right\} = 0, \tag{27}$$

where $\hat{H}$ is the estimate of $H(f)$. Additionally, Delattre and Fournier [35] studied the bias and variance of the k-NN estimator for entropy, suggesting that it is unbiased up to even a strong form of consistency, but variance increases as the number of dimensions d becomes higher. Such a type of consistency has also been claimed for the KDE estimator described in Section 2.1 by Ahmad and Lin [36].

The estimator for KL divergence was proposed by Wang et al. [23]. Considering p and q as continuous PDFs in $\mathbb{R}^d$, let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be independent and identically distributed (i.i.d.) d-dimensional samples drawn from p and q. Then, the proposed estimator for KL divergence is:

$$\hat{D}_{KL\,n,m}(p||q) = \frac{d}{n}\sum_{i=1}^{n} \log\left(\frac{\nu_k(i)}{\rho_k(i)}\right) + \log\left(\frac{m}{n-1}\right) \tag{28}$$

where $\rho_k(i)$ is the distance between X_i and its k-NN in $\{X_j\}_{j\neq i}$ and $\nu_k(i)$ is the distance between X_i and its k-NN in $\{Y_j\}$. The authors also demonstrate the mean-squared consistency of the estimator, as in Equation (27). $p = 2$ or the Euclidean norm is also suggested to calculate the distance.

Finally, the estimator for mutual information was proposed by Kraskov et al. [21]:

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\psi(n_{i,x}+1) + \psi(n_{i,y}+1)\right] + \psi(N) \tag{29}$$

where $n_{i,x}$ and $n_{i,y}$ are the number of neighbors in the X and Y spaces inside a radius given by the distance up to the kth nearest neighbor in the joint X-Y space. The authors require that the distance be calculated using $p = \infty$, i.e., the infinite or maximum norm.

Gao et al. [37] analyzed the properties of the proposed k-NN estimator of mutual information, finding it consistent, as it was with the previous two estimators. Further, they determined upper bounds of the rate of convergence in the estimate as a function of the dimensions of the two random variables involved.

To illustrate the somewhat abstract idea of using k-NN for estimating information-theoretic quantities, assume we wish to determine the mutual information between two univariate random variables X and Y. The algorithm for Equation (29) then works as follows: for each pair (x_i, y_i) in the data, find its kth nearest neighbor in the joint X-Y space, then count the number of neighbors inside a radius of distance $\rho_i^k/2$ in the X and Y spaces. Because the maximum norm is used ($p = \infty$), the neighbors are found strictly inside a row (X) and a column (Y) of width ρ_i^k. The sketch in Figure 1 serves to illustrate this procedure. In the case of multiple dimensions, the rows and columns for X and Y become hyper-stripes [15].

As the k-NN equations do not require an explicit evaluation of a probability density estimate, they look nothing like the canonical formulae shown in Equations (8)–(10). Nevertheless, an estimate of density at a point x_i can be calculated using k-NN:

$$\hat{p}_k(x_i) = \frac{k}{N-1} \cdot \frac{1}{c_1(d)\cdot \rho_k^d(i)} \tag{30}$$

This k-NN-based density estimation could be used as a plug-in estimate for entropy, for example, but this is not recommended, as this density estimate has some issues, which will be explored further in Section 3.4.2.

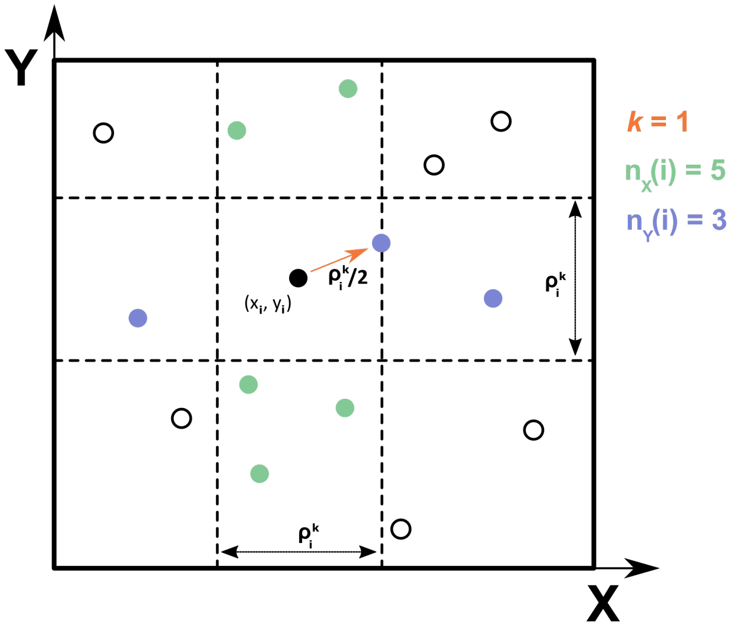

Figure 1. Determining ρ_i^k, $n_x(i)$, and $n_y(i)$ for a single pair (x_i, y_i) in the algorithm for Equation (29). Adapted from Kraskov et al. [21].

3. Comparison of Estimator Performance in Test Cases of Increasing Complexity

3.1. Test Case Design

We investigated the efficiency and accuracy of the different estimators described in Section 2 on a range of test cases summarized in Table 1. For each test case, we generated a sample of a fixed size from a given distribution and test the estimators' ability to quantify the three information-theoretic measures entropy, KL divergence, and mutual information.

We varied the sample size in order to reveal how data-hungry the different methods are in producing a reasonably good estimate of the true measure for the underlying distribution. Note that the difficulty is not only to have a large enough sample for the estimation method to work reliably, but, by its nature, a limited sample can only represent its corresponding theoretical distribution to a limited extent. Hence, we expect the estimates based on small sample sizes to be of mixed quality partly because they are not representative enough of the true distribution. We repeatedly drew samples of a certain size and report the average and the confidence interval of the estimates. We expect the performance and the confidence of the estimators to generally increase with the sample size due to this representation effect; yet, it remains to be observed how the different estimators behave for the different sample sizes and how they compare with each other in terms of accuracy and efficiency. The robustness of an estimator might be increased by bootstrapping; however, this will not resolve the issue of (non-)representativeness. It is beyond the scope of this study to design and recommend the best-suited bootstrapping approaches for the individual estimators; further, initial tests not shown here have confirmed the expectation that the variability resulting from bootstrapping is dominated by the much larger sampling variability, which is accounted for in this study, and hence, we neglected the effect of bootstrapping in this investigation.

The complexity of the test cases increases with respect to dimensionality (ranging from one to ten) and skewness (ranging from symmetrical to highly skewed) of the chosen distributions.

For the simplest 1D cases, there is no dependent variable to calculate mutual information on; so, in Cases 1 to 3, we only investigated the estimators' performance in calculating entropy and KL divergence. We deliberately started our investigation with low dimensions and well-behaved distribution shapes such that we can rely on analytical reference solutions for the information-theoretic quantities. Then, for cases of increased complexity, analytical equations no longer exist (or only for some quantities of interest), and we have to switch to computationally heavy numerical integration to serve as a reference for the true value of the target quantity (Cases 3, 5, and 6).

Table 1. Summary of test cases.

Case ID	Distribution	No. of Dimensions
1	Uniform	1
2	Normal	1
3	Normal mixture	1
4	Bivariate normal	2
5	Bivariate normal mixture	2
6	Gamma-exponential	2
7	Multivariate normal	4
8	Multivariate normal	10

3.2. Choice of Hyperparameters

Each estimation method needs the tuning of certain hyperparameters or making methodological choices. In the case of binning, the hyperparameter is the bin size Δ, and we investigated the differences in performance due to the four different methods to select the bin size presented in Section 2.2. In histogram-based methods, it is important to clarify how bins with zero probability, or empty bins, are handled because of the log function in all quantities of interest. In our specific application, where Equations (20)–(22) are sums, bins that have zero probability are excluded before the summation operation takes place. The hyperparameters of the QS approach are fixed to the recommended values of the original proposal [10] as described in Section 2.2.2.

For KDE, we show the difference between performing a resubstitution estimate as opposed to an estimate through numerical integration. In both cases, a kernel representation of the PDF of the data is created using a Gaussian kernel and Silverman’s bandwidth estimate, as described in Section 2.1. Then, the resubstitution estimate is obtained by averaging over the kernel PDF evaluated at all available data points in the sample, while the integration estimate is obtained by taking the expected value. Specifically, for entropy, Equations (15) and (18) show how these two processes are different.

As for k-NN, we investigated the effect of varying the number of neighbors used, k, between 1 and 15. The value of k was cut off there because, for larger values (e.g., $k = 50$), the results showed very similar behavior to $k = 15$, but with larger variability in small sample sizes.

3.3. Implementation

The sample sizes ranged between 100 and 100,000 samples, with the upper limit being reduced to 5000 in some cases related to the computationally expensive KDE-based estimator. The test case procedure was repeated across 300 different fixed seeds for random sampling to show confidence intervals related to the effect of sampling variability on an estimate.

Numerical integration as a reference when there is no analytical solution was achieved through the QUADPACK [38] (quad and nquad for higher dimensions) interface in *SciPy* [39]. For the integration scheme, default settings were used where both the absolute and relative error tolerances were set to 1.49×10^{-8}. Integration limits were set according to the true support of the underlying distribution.

Each of the estimators was implemented as a Python 3 [40] function. The binned estimators use the histogram, histogram-dd, and digitize functions from *NumPy* [41], the latter two used for multi-dimensional histograms with uneven and even binning widths, respectively. The KDE and k-NN estimators rely on the Gaussian KDE and the k-d Tree data structures available in *SciPy* [39], respectively.

The code for the setup of the test cases presented here to allow for reproducing our results and, e.g., future testing of further alternative methods is available at the *Data Repository of the University of Stuttgart (DaRUS)*. Further, all of the estimation methods have been collected as part of the open-source *UNITE toolbox*. The published toolbox itself contains the functions for the estimators to be applied to any case study of interest.

3.4. Case 1: 1D, Uniform Distribution

To begin our evaluation of the different estimation methods presented in this paper, we start with the uniform distribution. In the continuous case of the uniform distribution, the total mass of the distribution lies uniformly between two arbitrary bounds. As such, it is the maximum entropy distribution for a random variable X under no other constraints [3]. If the bounds of the uniform distribution are a and b, where $b > a$, then the PDF is:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \tag{31}$$

3.4.1. Analytical Reference Solution

The entropy of a uniform distribution $p(x)$ is given by

$$H(X) = \log(b-a). \tag{32}$$

If $q(x)$ is a second uniform distribution with bounds c and d where $(a,b) \subset (c,d)$, the KL divergence between the two distributions $p(x)$ and $q(x)$ is:

$$D_{KL}(p\,||\,q) = \log\left(\frac{d-c}{b-a}\right). \tag{33}$$

Figure 2 shows the PDFs of the distributions used to estimate entropy and KL divergence, together with the estimators' analytical reference value. For KL divergence, the approximating distribution ($q(x)$ in red) must have greater or equal support for the true distribution ($p(x)$ in black), which is the case for the example shown.

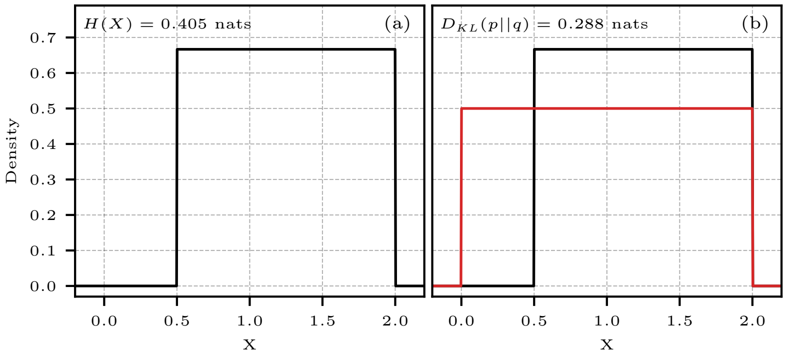

Figure 2. Case 1: 1D, uniform distribution. (**a**) PDF of a uniform distribution with a, b = 0.5, 2.0 and a reference value for entropy H; (**b**) same as (a) including the PDF of an approximating uniform distribution (in red) with c, d = 0.0, 2.0 and a reference value for KL divergence D_{KL}.

3.4.2. Density Estimation

For illustrative purposes, we start off by showing the ability of the different estimation methods to make an estimate of this one-dimensional uniform distribution. Figure 3 shows the true underlying PDF and the three PDFs derived by each method applied for density estimation, i.e., Equations (11), (19), and (30). The solid line for each method represents the mean of the estimated density across 300 random seeds of 10,000 samples of the uniform distribution, and the density was calculated for 500 evenly spaced points between 0.0 and

2.5 along the x axis. Additionally, the 2.5^{th} and 97.5^{th} percentile of the density estimated at each point is shown as the shaded area in the same color as the mean for all methods.

As expected, KDE gives a very accurate density estimation within the limits of the underlying distribution, but the effect of the Gaussian kernel is apparent close and outside of the limits. A boxcar-function-based kernel might be most suitable for this particular example, but said recommendation becomes more difficult when the data are higher dimensional; therefore, the Gaussian kernel is the default option [11]. Although binning produces estimates with higher variance than KDE, it seems to be the most suitable method for this example. The k-NN-based method was not explicitly designed to perform density estimation, and the example in Figure 3 graphically shows the large variance that the method has even in a 1D case [35]. The variance is affected by the number of neighbors used for density estimation; for smaller values of k, the variance is much higher than for larger values of k. In Figure 3, $k = 50$ was used. This is a much larger value than what is typically recommended, as we will see later, but appears adequate for density estimation, and larger values of k would indeed reduce the variance further. Nevertheless, the uniform distribution remains challenging due to the steps at the limits of the distribution, where only density estimation through binning is able to identify the "hard" cutoff limits.

Recall that density estimation is not the primary goal of our comparison; instead, we are interested in the estimation quality of the three information-theoretic quantities. Only the KDE and binning methods rely on this initial density estimation step; k-NN skips this step and is only shown here for a better intuition about the characteristics of this approach.

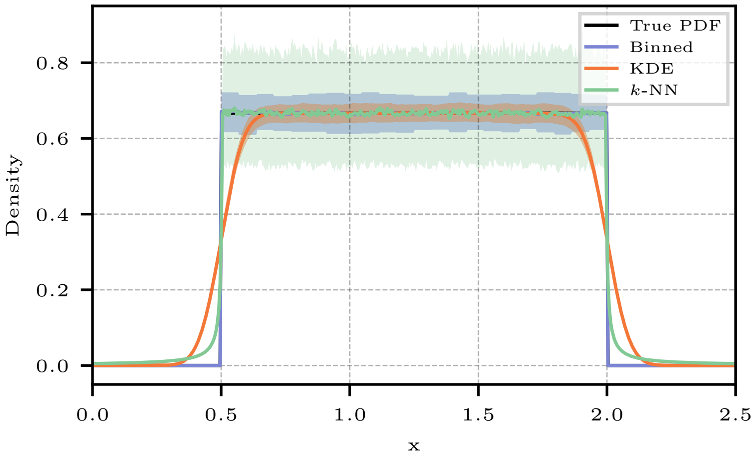

Figure 3. Comparison of all estimation methods for density estimation in Case 1 (1D, uniform distribution).

3.4.3. Estimation Results

The estimation results for the case of the uniform distribution are shown in Figure 4. The estimates of the three methods (columns) for the two quantities entropy and KL divergence (rows) are shown as absolute values in nats on the left-hand axes, whereas the relative errors with respect to the reference solution can be read off from the right-hand axes. Note that the axis scaling is consistent within each row (i.e., between estimates for a specific quantity), but not necessarily across the rows (i.e., between the different quantities).

This first test case serves as an intuitive starting point for the interpretation of Figure 4, as the results for all further test cases follow the same template.

The KDE estimator gives a greatly biased result when estimating entropy due to the effect of using a Gaussian kernel to approximate a uniform PDF (as seen in Figure 3). This is because the Gaussian kernel widens the true support of the uniform distribution, and therefore, it leads to an overestimation of entropy. Integration reduces this effect and also reduces the variance of the estimate in smaller sample sizes.

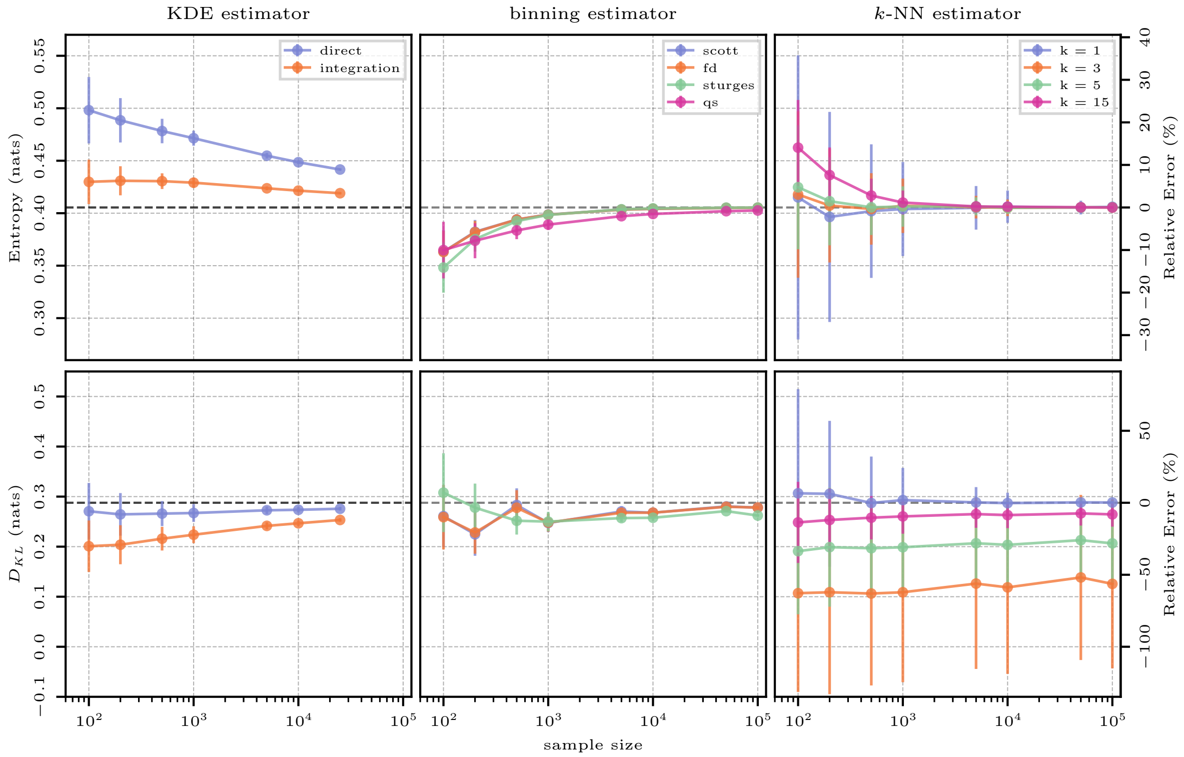

Figure 4. Evaluation of all estimation methods (columns) for entropy (**top**) and KL divergence (**bottom**), in Case 1 (1D, uniform distribution).

Binning underestimates entropy for sample sizes up to 1000 and then yields very accurate and precise estimates. This observation is well in line with the theory, since a too-small-to-be-representative sample will not be perfectly uniform and, hence, the true entropy value of the maximum-entropy distribution is approached from below with increasing sample size. The QS approach typically underestimates entropy when compared to a histogram-based approach and one of the considered "rules-of-thumb".

The k-NN estimator for entropy gives a similarly accurate result as binning, but with a much higher variance. As discussed in the context of density estimation, the k-NN method does not rely on any smoothing via kernels or bins, and hence, the sampling variability hits its result hard. As expected, this variability is reduced with increasing number of neighbors; however, a smaller number of neighbors yields more accurate results for smaller sample sizes.

Estimating KL divergence directly using the KDE estimator gives a result very close to the truth even in small sample sizes, with an increase in the number of samples only reducing the variance. Across different samples, the ratio between the true and the approximating distribution seems to stay consistent, leading to this result. Numerical integration leads to a worse outcome as the effect of the Gaussian kernel becomes more apparent (both distributions are first approximated, and then, their ratio is assessed, "doubling" the smearing effect of the kernel).

For the binning estimator, we see a "reversal effect" in the estimate for KL divergence where Sturges' rule initially computes a larger KL divergence, but starts to produce results that are lower than Scott's and FD's rule at 500 samples. The reason is the resulting number of bins obtained from the different methods described in Section 2.2.1: across larger sample sizes, Scott's and FD's rules result in a very similar number of bins with FD's estimate being usually slightly larger, and both being larger than Sturges' rule. This situation is reversed in smaller sample sizes, with Sturges' rule producing the largest number of bins.

The k-NN approach yields the best results for KL divergence with a single neighbor ($k = 1$). While, again, the variance is high for small sample sizes, the estimate is highly accurate across all sample sizes. Obviously, a one-to-one comparison of estimated densities is preferable for an accurate estimate of KL divergence, as also seen for KDE (a direct estimate being much superior to integration). While intermediate numbers of neighbors perform poorly, the highest number investigated here ($k = 15$) comes close to the accuracy of $k = 1$, but fails to reach it. It is also remarkable that k-NN-based estimation of D_{KL} shows very little variation across sample sizes, again emphasizing that a local estimate of the ratio between the two distributions is less affected by unrepresentative sampling or distribution reconstruction.

3.5. Case 2: 1D, Normal Distribution

The normal or Gaussian distribution is the second of our test cases and also serves as the basis for the applications in higher dimensions, where the estimators will be tested on multivariate normal distributions (Sections 3.10 and 3.11). In the most basic one-dimensional case, the PDF of the normal distribution is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{34}$$

The normal distribution is often denoted as $\mathcal{N}(\mu, \sigma^2)$, where μ is the expected value of the distribution and σ is the standard deviation.

3.5.1. Analytical Reference Solution

Given the PDF in Equation (34), the entropy of the normal distribution is:

$$H(X) = \frac{1}{2} \ln\left(2\pi\sigma^2\right) + \frac{1}{2} \tag{35}$$

Given two normal distributions, where $p(x) = \mathcal{N}(\mu_p, \sigma_p^2)$ and $q(x) = \mathcal{N}(\mu_q, \sigma_q^2)$, the KL divergence between these two distributions is:

$$D_{KL}(p||q) = \frac{1}{2}\left\{ \left(\frac{\sigma_p}{\sigma_q}\right)^2 + \frac{(\mu_q - \mu_p)^2}{\sigma_q^2} - 1 + \ln\left(\frac{\sigma_q^2}{\sigma_p^2}\right) \right\} \tag{36}$$

Figure 5 shows the distributions used in this application case.

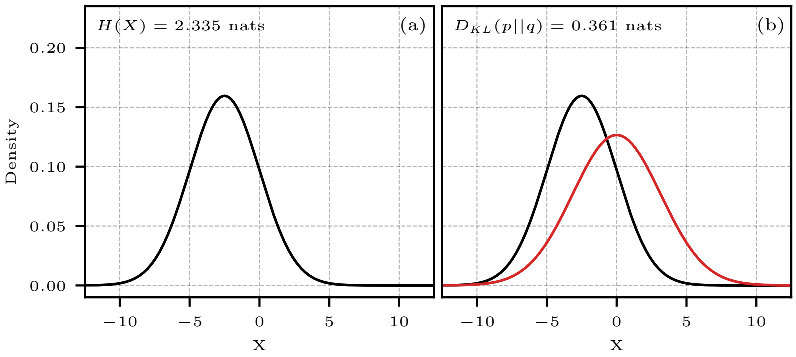

Figure 5. Case 2: 1D, normal distribution. (**a**) PDF of a normal distribution $\mathcal{N}(-2.5, 2.5^2)$ and a reference value for entropy H; (**b**) same as (**a**) including the PDF of an approximating normal distribution (in red) $\mathcal{N}(0, 3.15^2)$ and a reference value for KL divergence D_{KL}.

3.5.2. Estimation Results

The estimation results for the case of the normal distribution are shown in Figure 6. Estimating the density of a normal distribution with a Gaussian kernel is a prime task for KDE, so the relative errors in estimating the two information-theoretic quantities are expected to be small. Across all sample sizes and quantities, a direct evaluation of the KDE estimator performs better than performing numerical integration over the kernel representation. The latter gives additional importance to the overall shape of the distribution, which is an approximation in both cases, rather than focusing on the specific points, which support the kernel representation. This can be seen more clearly for KL divergence, where ratios are considered. Only at large sample sizes, where the overall representation of the underlying distribution is very close to the true distribution, numerical integration gives a result that is almost as good as direct estimation.

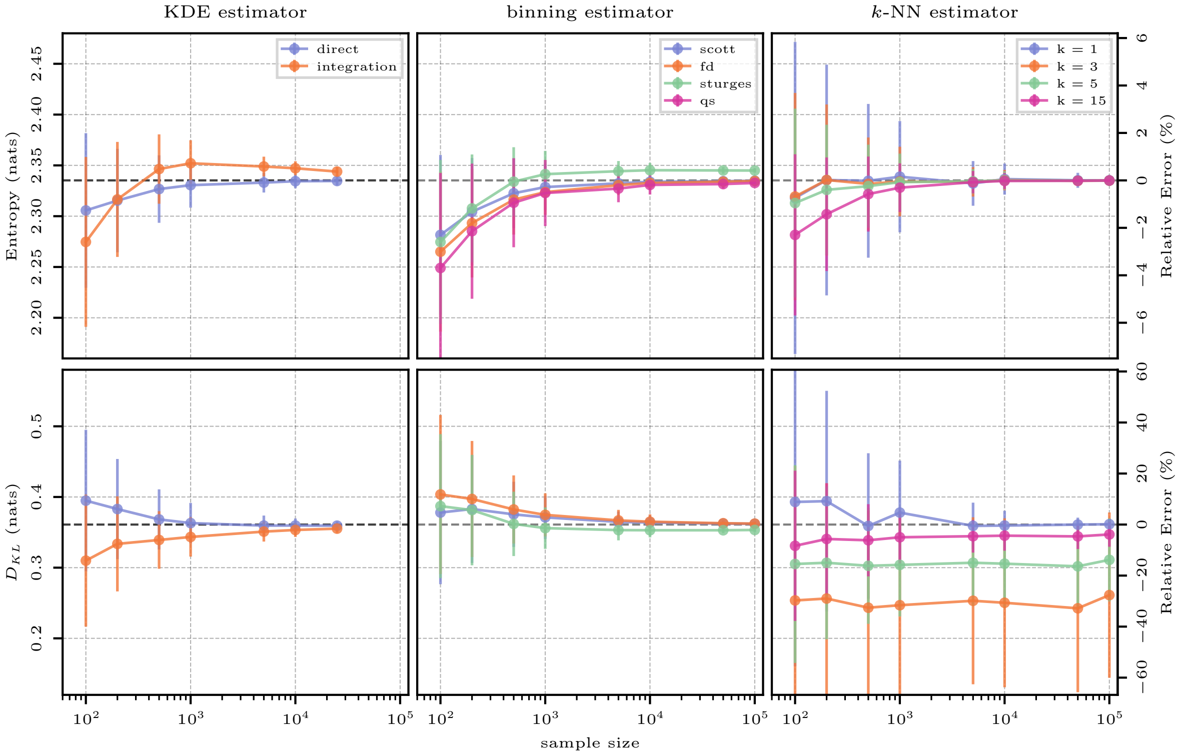

Figure 6. Evaluation of all estimation methods (columns) for entropy (**top**) and KL divergence (**bottom**), in Case 2 (1D, normal distribution).

All binning methods, with the exception of Sturges' rule, give very similar results, which are highly accurate for large sample sizes. This is expected as the rules-of-thumb were all derived with more or less strict assumptions on the distribution being Gaussian. The bin width given by Scott's works best because it is optimal for random samples of normally distributed data, in the sense that it minimizes the integrated mean-squared error of the density estimate [27]. Neither FD's and Sturges' estimate follow this same property and allow for a more flexible representation of the underlying distribution. On normally distributed data and in large sample sizes, FD's rule equally matches Scott's rule, but Sturges' rule overestimates entropy and underestimates Kullback–Leibler divergence because it calculates a larger Δ and a smaller number of bins. The QS approach does not assume normally distributed data (and hence, also cannot benefit from that knowledge); as in the previous case, it tends to slightly underestimate entropy, but its performance is generally very similar to the other binning methods. Remember that it is only derived for estimating entropy, not KL divergence or other multivariate quantities.

The results of using k-NN to estimate entropy and KL divergence appear consistent with those of the previous case: for small sample sizes, the bias decreases with decreasing k, but the variance increases. However, for sample sizes > 1000, the bias and variance almost diminish; also, for any choice of hyperparameter and sample size, the relative error of the entropy estimates is generally small, with the largest relative error being $\approx$6%. This demonstrates the effectiveness of this estimation method when applied to normally distributed data. For KL divergence, again, $k = 1$ is the best choice of hyperparameter, and given a specific k, the mean estimates of KL divergence are very consistent across sample sizes.

3.6. Case 3: 1D, Normal Mixture Distribution

To exploit the statistical properties of normal distributions while making them more flexible in fitting non-normal real-word data sets, it is common to use weighted mixtures of normal distributions in approximation tasks. Such a mixture offers the possibility to test the estimators presented in this paper in a setting where the data come from a distribution that is not uni-modal and more skewed. Further, the main interest in evaluating a mixture of normal distributions is that a normal mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a normal mixture model with enough components [33]. The PDF of a normal mixture distribution is:

$$p(x) = \sum_{i=1}^{n} \omega_i \mathcal{N}(\mu_i, \sigma_i^2), \tag{37}$$

where ω_i is the individual weight of each of the n components of the mixture, with their individual expected values μ_i and standard deviations σ_i, and $\sum_{i=1}^{n} \omega_i = 1$.

3.6.1. Numerical Reference Solution

Because of the logarithm of a sum of exponential functions, the entropy of a mixture of normal distributions cannot be calculated [42]. The same applies to KL divergence. Therefore, numerical integration was used to obtain the reference value presented in the results.

Figure 7a shows the distribution used for this application case. This is an equally weighted mixture of the two normal distributions $\mathcal{N}(-2.5, 2.5^2)$ and $\mathcal{N}(2.5, 1^2)$.

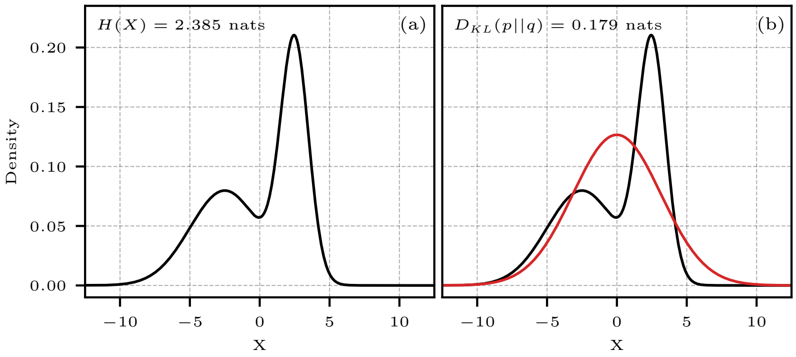

Figure 7. Case 3: 1D, normal mixture distribution. (**a**) PDF of a normal mixture distribution (see the parameters in the text) and a reference value for entropy H; (**b**) same as (**a**) including the PDF of an approximating normal distribution (in red) $\mathcal{N}(0, 3.15^2)$ and a reference value for KL divergence D_{KL}.

3.6.2. Estimation Results

The estimation results of the mixture of normal distributions are shown in Figure 8 and they are very similar to those of the single normal distribution in the previous case. KDE again performs well due to the usage of a Gaussian kernel. Given the two peaks of the distribution as seen in Figure 7a, for a low number of samples, the kernel density at specific points is underestimated, and therefore, the entropy is overestimated. This is the opposite as in Case 2 (pure normal distribution), in which small samples tend to over-represent a narrower high-probability range of the underlying distribution, which leads to an underestimation of entropy. In both cases, direct estimation gives a very accurate result for sample sizes larger than 1000; integration struggles much more in the case of the normal mixture distribution.

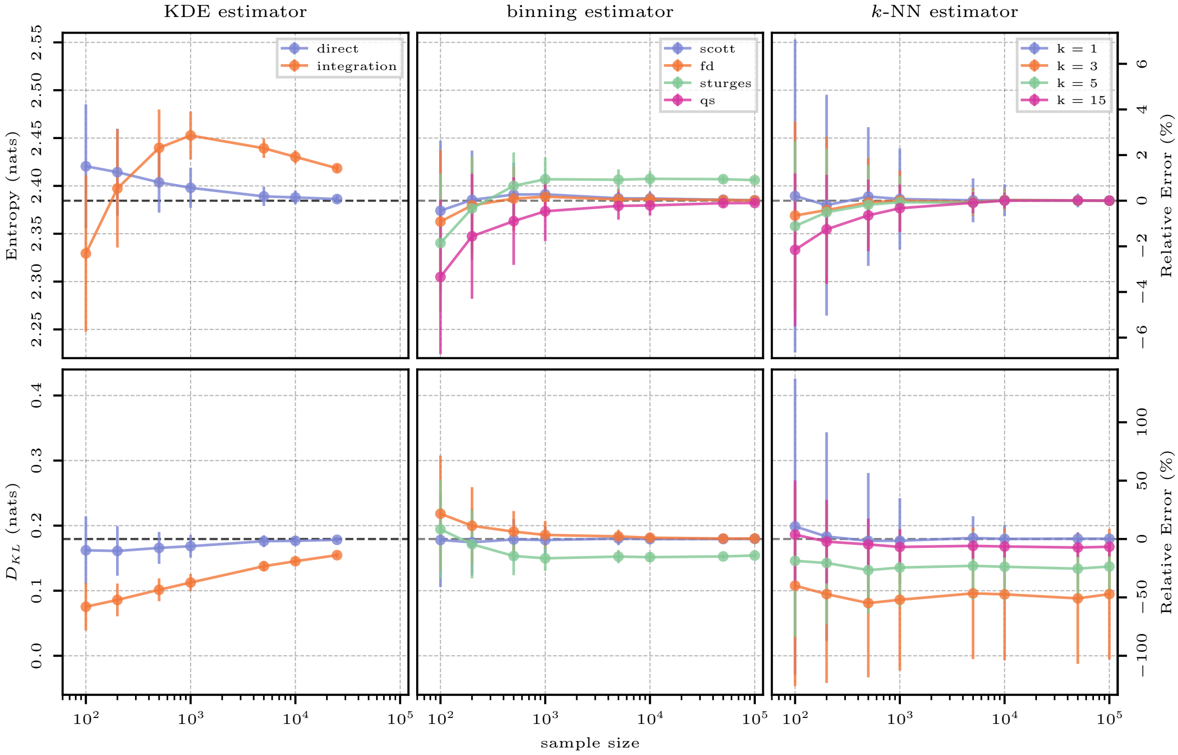

Figure 8. Evaluation of all estimation methods (columns) for entropy (**top**) and KL divergence (**bottom**), in Case 3 (1D, normal mixture distribution).

For the binning estimator, the results are also very similar to the previous case: both Scott's and FD's estimate perform very well; the QS approach converges to their result from below; Sturges' rule struggles even more than before due to an inadequate number of bins for both estimating entropy and KL divergence. Scott's rule appears to be best to calculate an adequate number of bins, particularly for KL divergence where the estimate based on said rule appears to consistently match the true value across all sample sizes.

k-NN-based estimation behaves similarly to all previous one-dimensional cases. Here, once again, values of $k = 3$ or $k = 5$ give an estimate of entropy that is very close to the true value while having a smaller variance than $k = 1$. $k = 15$ gives the estimate with the smallest variance, but largest bias for small sample sizes. For estimating KL divergence, again, $k = 1$ performs best and $k = 15$ approaches this result (but still shows a small bias even for large sample sizes).

3.7. Case 4: 2D, Normal Distribution

For extending Case 2 to a normal distribution with d dimensions, we use the notation $\mathcal{N}_d(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ is a d-dimensional mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is a $d \times d$ covariance matrix, where Σ is positive-definite. For this distribution, the PDF is:

$$p(x) = \frac{1}{(2\pi)^{d/2}(\det(\Sigma))^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right) \tag{38}$$

3.7.1. Analytical Reference Solution

The entropy of this multivariate normal distribution is given by [1]:

$$H(X) = \frac{1}{2}\ln\left((2\pi e)^d \cdot \det(\Sigma)\right) \tag{39}$$

The KL divergence between two d-dimensional normal distributions $p(x) = \mathcal{N}_d(\mu_p, \Sigma_p)$ and $q(x) = \mathcal{N}_d(\mu_q, \Sigma_q)$ [43]:

$$D_{KL}(p||q) = \frac{1}{2}\left[\ln\left(\frac{\det\left(\Sigma_q\right)}{\det\left(\Sigma_p\right)}\right) + \mathrm{Tr}\left(\Sigma_q^{-1}\Sigma_p\right) + \left(\mu_q - \mu_p\right)^{\mathsf{T}}\Sigma_q^{-1}\left(\mu_q - \mu_p\right) - d\right] \tag{40}$$

In particular, for $d = 2$, the parameters for the normal distribution become a 2 × 1 vector for the means and a 2 × 2 matrix for the covariance. These parameters can be written as:

$$\mu = \begin{bmatrix}\mu_0 & \mu_1\end{bmatrix}^{\mathsf{T}} \quad \Sigma = \begin{bmatrix}\sigma_0^2 & \rho\,\sigma_0\,\sigma_1 \\ \rho\,\sigma_0\,\sigma_1 & \sigma_1^2\end{bmatrix}$$

This notation is similar to that of Equation (34) with the addition of ρ being the Pearson correlation coefficient between X_0 and X_1. Using this parametrization, the expression for mutual information for the bi-variate normal distribution becomes:

$$I(X_0, X_1) = -\frac{1}{2}\log\left(1 - \rho^2\right) \tag{41}$$

This means that, for a bivariate normal distribution, there is an exact relationship between the correlation coefficient ρ and mutual information [44].

Figure 9a shows a bivariate normal distribution $p(x_1, x_2) = \mathcal{N}_2(\mu_p, \Sigma_p)$, and Figure 9b shows the same, with a second approximating distribution $q(x_1, x_2) = \mathcal{N}_2(\mu_q, \Sigma_q)$ in red. The parameters for both distributions are:

$$\mu_p = \begin{bmatrix}-2 & 0\end{bmatrix}^{\mathsf{T}} \qquad \mu_q = \begin{bmatrix}0 & 0\end{bmatrix}^{\mathsf{T}}$$

$$\Sigma_p = \begin{bmatrix}1 & -0.5 \\ -0.5 & 1\end{bmatrix} \quad \Sigma_q = \begin{bmatrix}5 & 0 \\ 0 & 1\end{bmatrix}$$

Additionally, in Σ_p, the value of ρ is -0.5 and can be directly used in Equation (41). This result is also shown in Figure 9a.

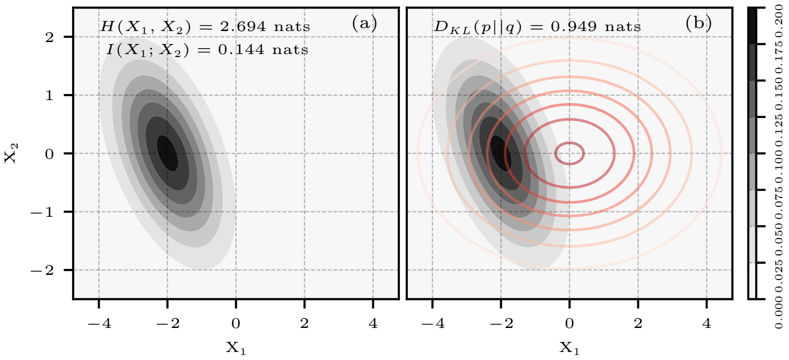

Figure 9. Case 4: 2D , normal distribution. (**a**) PDF of a bivariate normal distribution (see the parameters in the text) and reference values for entropy H and mutual information I; (**b**) same as (**a**) including the PDF of an approximating bivariate normal distribution (in red; see the parameters in the text) and a reference value for KL divergence D_{KL}.

3.7.2. Estimation Results

This initial case in two dimensions starts to introduce the challenges of density estimation in higher dimensions. The estimation results are shown in Figure 10. More specifically, for KDE-based estimation, direct evaluation of a kernel representation of the probability distribution is very computationally expensive for sample sizes larger than 25,000, and the waiting time to obtain an estimate becomes unreasonable. Therefore, this method is tested no further than this number of samples. Computational costs will be discussed further in Section 3.12. Nevertheless, estimates for entropy, KL divergence, and now, also mutual information using this technique are quite accurate with a behavior similar to the one-dimensional case in Section 3.5. Numerical integration of the KDE density estimate further increases the computational cost with the procedure becoming unfeasible for a number of samples larger than 10,000. Further, integration only notably improves performance for mutual information.

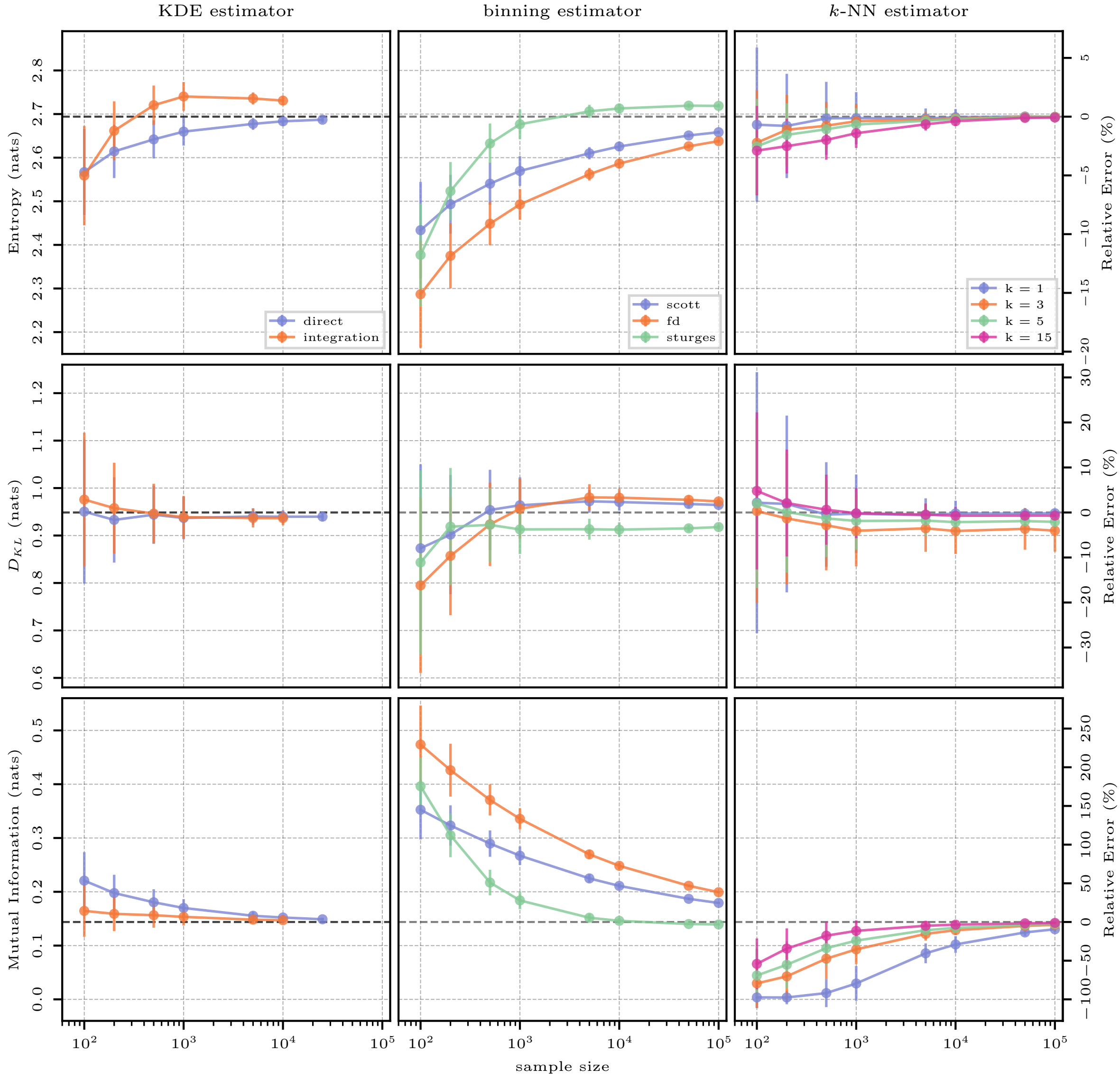

Figure 10. Evaluation of all estimation methods (columns) for entropy (**top**), KL divergence (**middle**), and mutual information (**bottom**), in Case 4 (2D, normal distribution).

Binning-based estimation does not suffer from the increase in computational cost as much as KDE-based estimation, but there is a small increase (see Section 3.12). As opposed to the one-dimensional cases investigated so far, binning yields large relative errors when estimating entropy even for larger sample sizes. The main drawback of the different binning rules is their inability to account for the multiple dimensions of the data, making the method less effective for dimensions higher than one. Here, Sturges' rule is superior to the other methods, as it yields relatively accurate estimates for entropy and mutual information for samples of size 1000 or larger. For KL divergence, Scott's and FD's rules seem more effective. It appears that, for entropy and mutual information, a smoother histogram created using a larger Δ is more favorable, while for KL divergence, a more detailed histogram with a finer Δ is preferred. Recall that the QS approach has not been extended to higher dimensions yet and, hence, cannot be analyzed in this and the following multi-dimensional test cases.

Finally, for k-NN estimation, the findings from the previous one-dimensional cases seem to hold, at least for two dimensions. Namely, estimation performs best for smaller values of k both for entropy and KL divergence. This is not true for mutual information, where a large value for $k = 15$ gives better estimates across all sample sizes. Importantly, for certain cases, particularly in Figure 10 shown for small sample sizes and $k = 1$, the second term in Equation (29) becomes larger than all other terms, resulting in an estimation of negative mutual information. As mutual information cannot be negative per its definition (Equation (7)), the estimator reports 0.0 as the result. Hence, the k-NN estimate converges to the true mutual information systematically from below.

3.8. Case 5: 2D, Normal Mixture Distribution

To increase the level of complexity while exploiting the fact that two dimensions still allow for visualization, we again tested on a mixture of now bivariate normal distributions using the scheme in Equation (37).

3.8.1. Numerical Reference Solution

As in Case 3, analytical expressions for entropy, KL divergence, or mutual information do not exist for this type of distribution. We, therefore, relied on numerical integration to obtain the reference values presented in the results. Figure 11a shows an equally weighted mixture of $\mathcal{N}_2(\mu_0, \Sigma_0)$ and $\mathcal{N}_2(\mu_1, \Sigma_1)$.

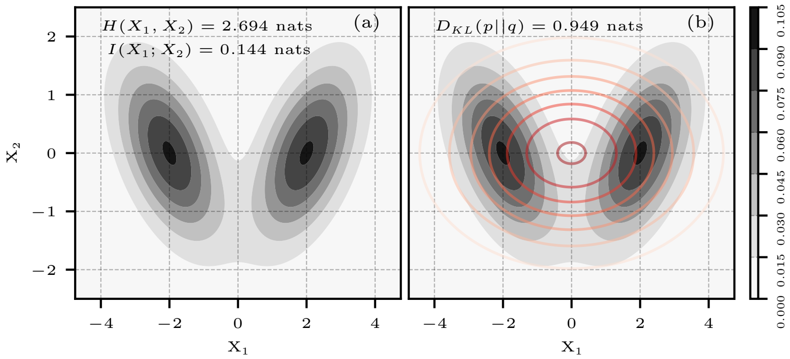

Figure 11. Case 5: 2D, normal mixture distribution. (**a**) PDF of a bivariate normal mixture distribution (see the parameters in the text) and reference values for entropy H and mutual information I; (**b**) same as (**a**) including the PDF of an approximating bivariate normal distribution (in red; see the parameters in the text) and a reference value for KL divergence D_{KL}.

The parameters of the mixture are:

$$\mu_0 = \begin{bmatrix} -2 & 0 \end{bmatrix}^\intercal \qquad \mu_1 = \begin{bmatrix} 2 & 0 \end{bmatrix}^\intercal$$

$$\Sigma_0 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Figure 11b shows an approximating bivariate normal distribution $\mathcal{N}_2(\mu_2, \Sigma_2)$. The parameters of this distribution come from making the incorrect assumption that the x_0, x_1 pairs shown in Figure 11a are independent.

$$\mu_2 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\intercal \quad \Sigma_2 = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$

3.8.2. Estimation Results

The results of this test case are found to be very similar to the previous case of a bivariate normal distribution, with the most notable difference being that KDE seems to struggle more when moving away from a pure normal distribution, meaning that the performance is overall worse in this case.

Note that, for estimating mutual information with the k-NN method, further tests were conducted using $k = 50$, but there were no noticeable differences between this value of k and the maximum shown in Figure 12 of $k = 15$.

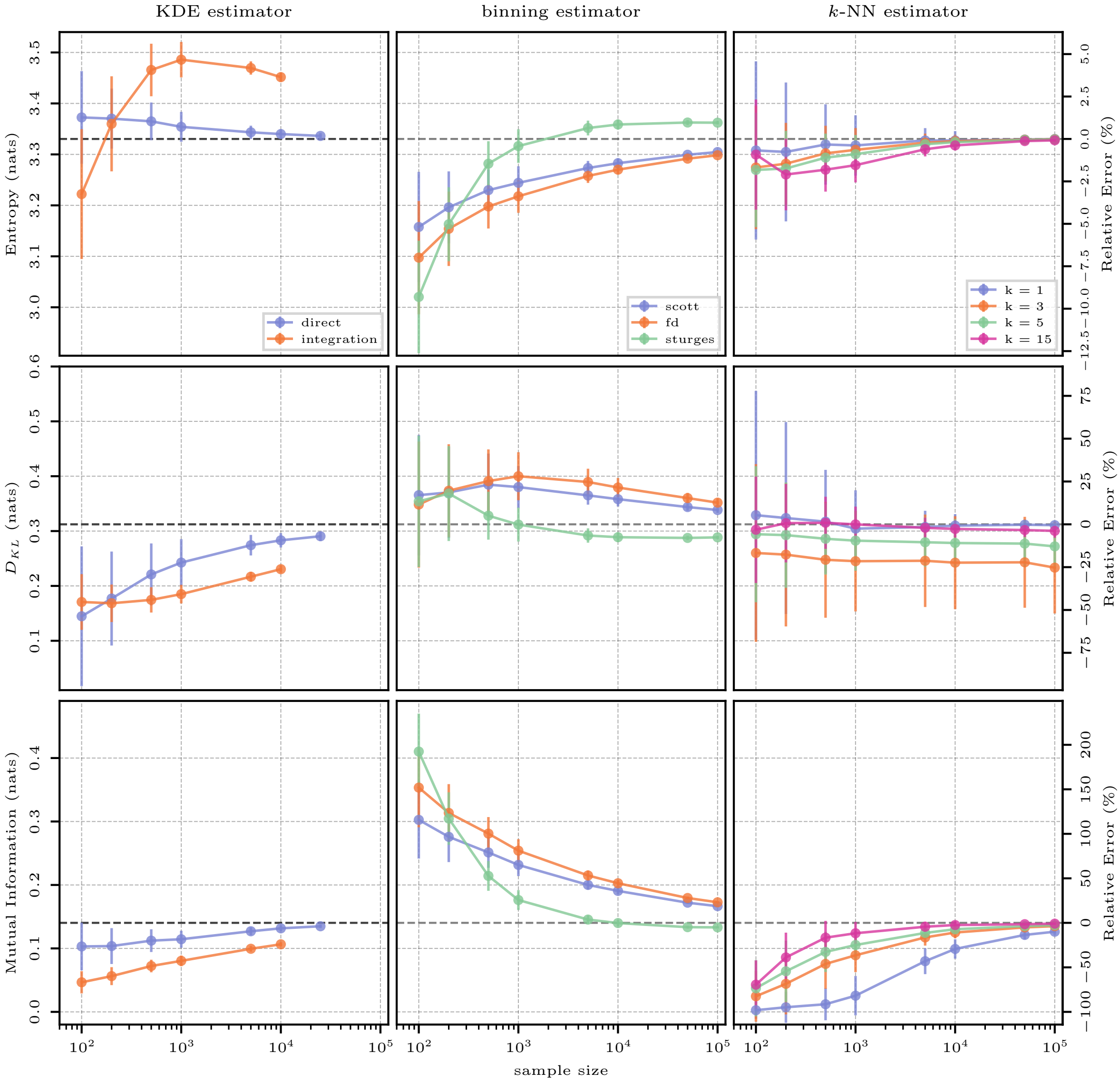

Figure 12. Evaluation of all estimation methods (columns) for entropy (**top**), KL divergence (**middle**), and mutual information (**bottom**), in Case 5 (2D, normal mixture distribution).

3.9. Case 6: 2D, Gamma-Exponential Distribution

An example of a skewed distribution in two dimensions is the gamma-exponential distribution, which has the following PDF [21], defined for $x_1, x_2 > 0$:

$$p(x_1, x_2) = \frac{x_1^{\theta} e^{-x_1 - x_1 \cdot x_2}}{\Gamma(\theta)} \tag{42}$$

where θ is the scaling parameter of the distribution >0 and Γ is the gamma function.

3.9.1. Analytical and Numerical Reference Solutions

Darbellay and Vajda [45] presented a list of multivariate differential entropies and mutual information. According to them, the (joint) entropy of the gamma-exponential distribution is:

$$H(X_1, X_2) = 1 + \theta - \theta \cdot \psi(\theta) + \ln \Gamma(\theta) - \ln(1) \tag{43}$$

where ψ is the digamma function or $\frac{d}{d\theta} \ln \Gamma(\theta)$. The mutual information is:

$$I(X_1, X_2) = \psi(\theta) - \ln(\theta) + \frac{1}{\theta} \tag{44}$$

For calculating entropy and mutual information in this case, the scaling parameter was set to $\theta = 3$. For determining KL divergence, the parameter of the approximating distribution was set to $\theta = 4$. Because no reference in the literature was found to provide a theoretical result of relative entropy between two gamma-exponential distributions, numerical integration was used to obtain the true value of D_{KL} presented in the results. Figure 13 shows a plot of the two PDFs used in this test case.

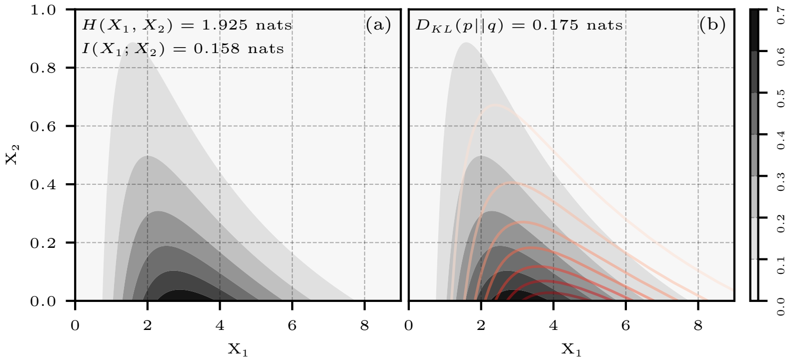

Figure 13. Case 6: 2D, gamma-exponential distribution. (**a**) PDF of the gamma-exponential distribution, where $\theta = 3$, and reference values for entropy H and mutual information I; (**b**) same as (**a**) including the PDF of an approximating function (in red), where $\theta = 4$, and a reference value for KL divergence D_{KL}.

3.9.2. Estimation Results

The estimation results for the case of the gamma-exponential distribution are shown in Figure 14. Here, we have the case of a distribution that has a different shape than the normal distribution; therefore, the limitations of the usage of a Gaussian kernel in the KDE-based estimator become more apparent. The KDE-based entropy estimate only achieves an accurate result for large sample sizes and using numerical integration. This seems to contradict the previous cases, but this can be explained by how numerical integration is implemented on the KDE estimator. Numerical integration requires a set of limits, and as implemented, the limits were chosen as the maximum and minimum values available in the sample for each dimension, plus and minus the specific bandwidth calculated for the Gaussian kernel. Therefore, these limits constrain the region where numerical integration happens, even though the kernel-based representation of the distribution has support everywhere due to the usage of the Gaussian kernel. This effect also applies to KL divergence and mutual information, but is more apparent in the latter. Both of these estimates remain biased to a significant degree even with the largest sample sizes tested here.

Binning estimates applying Scott's rule give the best results, while the FD rule drastically underestimates entropy and overestimates mutual information. Sturges' rule shows the opposite behavior. As previously commented on, this is due to the FD rule calculating a smaller Δ and forcing a greater number of bins than Sturges'. This overestimation on the required number of bins causes the representation of the joint distribution $p(x, y)$ to be sparser and to have a larger number of empty bins, leading to a smaller variability in the representation of $p(x,y)$. Generally, estimating entropy with bins using Scott's rule leads to small relative errors for practically all considered sample sizes, whereas the mutual information estimate using this same rule still contains quite high relative errors even for the largest sample size of 100,000. Because KL divergence compares the ratios of $p(x)$ and $q(x)$, very similar results are obtained across the binning hyperparameters as the absolute bin size does not matter as much as the fact that equal binning schemes are defined for $p(x)$ and $q(x)$.

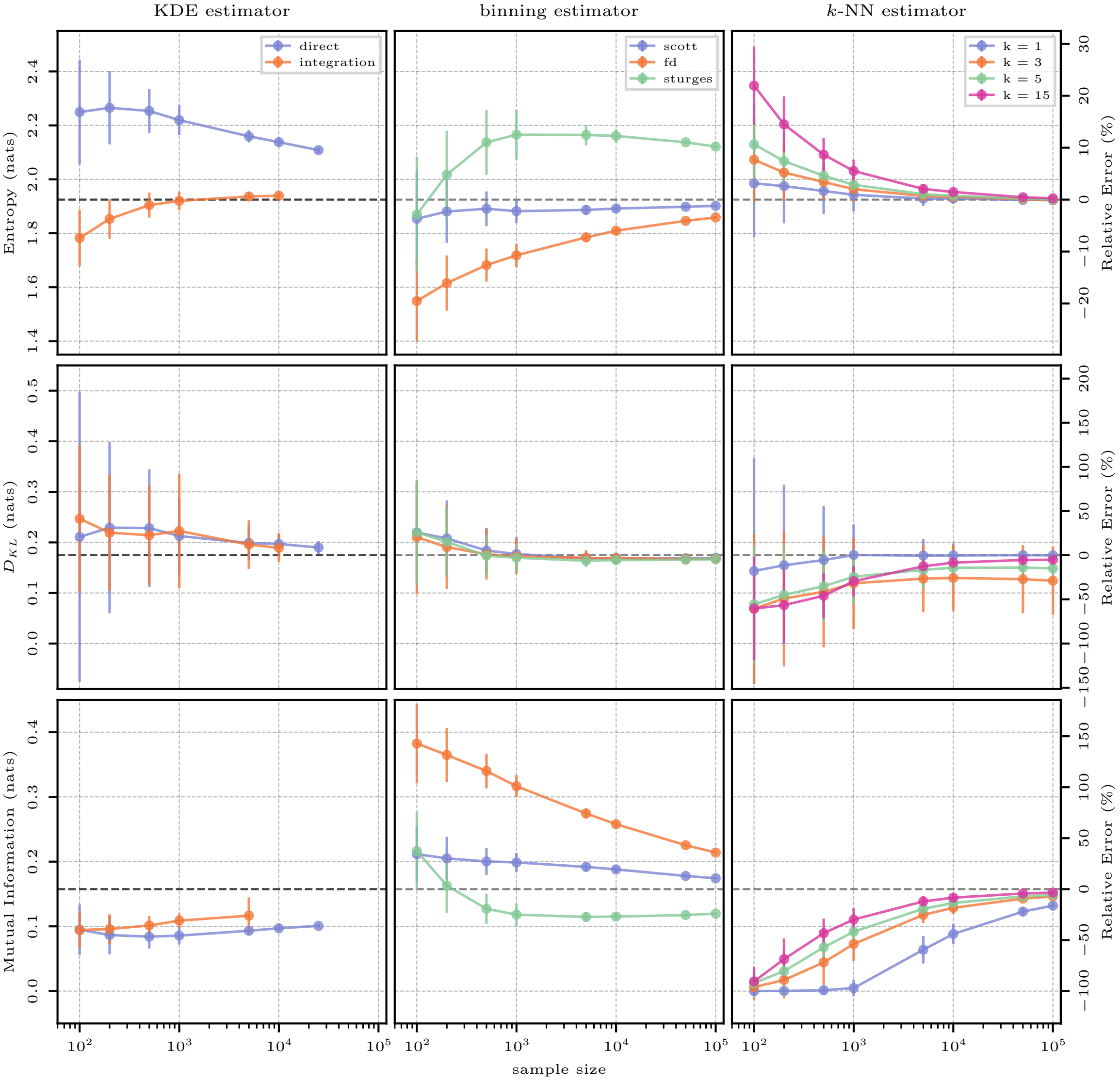

Figure 14. Evaluation of all estimation methods (columns) for entropy (**top**), KL divergence (**middle**), and mutual information (**bottom**), in Case 6 (2D, gamma-exponential distribution).

As with the other cases, k-NN-based estimates improve in accuracy with increasing sample size, and $k = 1$ performs best for entropy and KL divergence. In the case of entropy, larger k leads to a worse approximation, while still, all investigated values of k produce very small errors with large sample sizes. For KL divergence, however, the results become more mixed with larger sample sizes, where the highest considered k of 15 approaches the highly accurate result of $k = 1$, while the smaller k values in between seem to stabilize at large relative errors, underestimating the true KL divergence significantly. In the case of mutual information, larger k is favorable for all sample sizes, but all k larger than one achieve acceptable relative errors for sample sizes larger than 1000.

3.10. Case 7: 4D, Multivariate Normal Distribution

For higher dimensions, we tested the estimators' ability to quantify the entropy, KL divergence, and mutual information of a four-dimensional normal distribution, since analytical solutions are available. We adopted the experiments by Wang et al. [23] and replicated their results, while going further with their experiment by quantifying the sampling uncertainty of their results, as well as investigating additional estimation methods. For entropy and mutual information, data were sampled from $\mathcal{N}_4(\mu_0, \Sigma_0)$, and for KL divergence, data were also sampled from $\mathcal{N}_4(\mu_1, \Sigma_1)$.

$$\mu_0 = \begin{bmatrix}0.1 & 0.3 & 0.6 & 0.9\end{bmatrix}^\intercal \qquad \mu_1 = \begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}^\intercal$$

$$\Sigma_0 = \begin{bmatrix}1 & 0.5 & 0.5 & 0.5\\0.5 & 1 & 0.5 & 0.5\\0.5 & 0.5 & 1 & 0.5\\0.5 & 0.5 & 0.5 & 1\end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix}1 & 0.1 & 0.1 & 0.1\\0.1 & 1 & 0.1 & 0.1\\0.1 & 0.1 & 1 & 0.1\\0.1 & 0.1 & 0.1 & 1\end{bmatrix}$$

3.10.1. Analytical Reference Solution

In this case, we can use Equations (39) and (40) as analytical expressions for entropy and KL divergence, respectively.

For mutual information, Arellano-Valle et al. [46] suggested breaking down the distribution into marginal distributions with n and m dimensions, where $d = n + m$ and:

$$\begin{pmatrix}X\\Y\end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\\Sigma_{YX} & \Sigma_{YY}\end{pmatrix}\right) \tag{45}$$

In this form, using the expression for mutual information in the first row of Equation (7) and considering the expression for entropy obtained in Equation (39), mutual information can be written as:

$$I(X,Y) = \frac{1}{2}\ln\left(\frac{\det(\Sigma_{XX})\det(\Sigma_{YY})}{\det(\Sigma)}\right) = -\frac{1}{2}\ln\left(\det\left(\mathrm{I}_n - \Sigma_{XX}^{-1}\Sigma_{XY}\cdot\Sigma_{YY}^{-1}\Sigma_{YX}\right)\right) \tag{46}$$

As we cannot present plots of the PDF of the distributions used in this application case, we show here the reference true values:

$$H(X) = 5.09 \text{ nats} \quad D_{KL}(p||q) = 0.90 \text{ nats} \quad I(X;Y) = 0.24 \text{ nats}$$

The reference solution for mutual information comes from applying Equation (45) to the distribution $\mathcal{N}_4(\mu_0, \Sigma_0)$. The distribution is split so that the distribution of X contains the first three dimensions of the original distribution; therefore $n = 3$; the distribution of Y has dimensions of the highest order ($d = 4$) and $m = 1$. More succinctly, mutual information is calculated as: $I([x_1, x_2, x_3]; x_4)$.

3.10.2. Estimation Results

The estimation results for the case of the 4D normal distribution are shown in Figure 15. In higher dimensions than those of the previous cases, the main drawbacks of some of the estimation methods start to become apparent. To begin, numerical integration of a KDE-based probability distribution becomes too computationally expensive, so this method was not considered in this experiment. Further, as was also the case for some of the previous cases, the evaluation of a KDE-based representation also becomes too expensive for larger sample sizes; therefore, the KDE-based direct resubstitution estimate was only calculated up to 25,000 samples. For entropy and mutual information, it seems that the estimator would converge for a larger number of samples, with the estimates having very little variance after 1000 samples. The estimate for KL divergence, however, is extremely stable with a highly biased result (40% relative error).

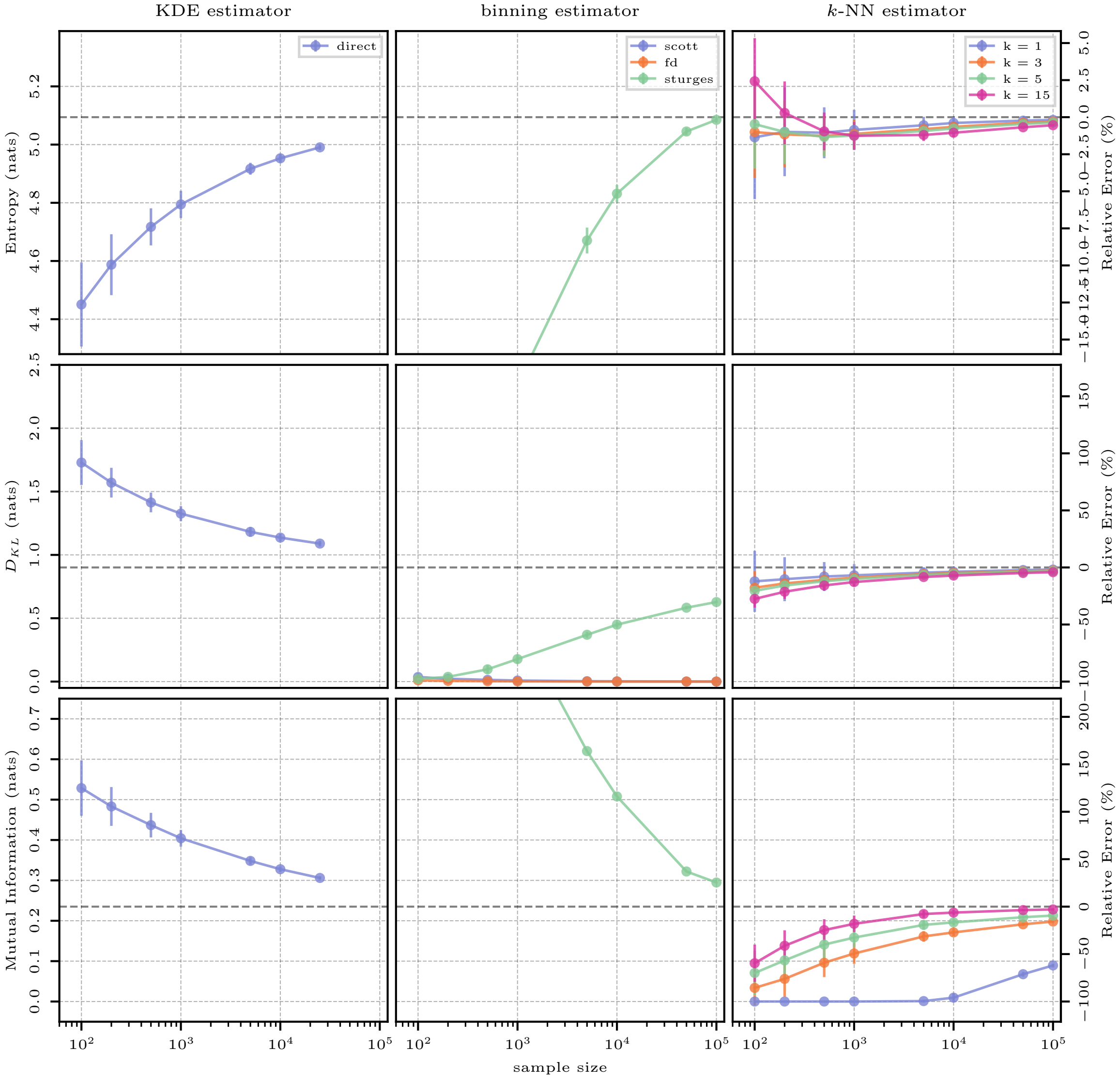

Figure 15. Evaluation of all estimation methods (columns) for entropy (**top**), KL divergence (**middle**), and mutual information (**bottom**), in Case 7 (4D, normal distribution).

Binning-based estimation performs acceptably well only when using Sturges' rule. Sturges' rule typically results in the largest bin width, with Scott's and FD's rules giving smaller bin widths and, therefore, a larger amount of bins in the histogram-based representation of the PDF. In the case of entropy, for smaller bins, the contribution of each bin to the computation of entropy is very small and the correction factor described in Section 2.2 and Equation (20) dominates the calculation, typically underestimating entropy and resulting in values that are not shown in the limits of the plot in Figure 15. Only Sturges' rule with its larger bin sizes converges to the true value with the highest investigated sample size, however, approaching the true value very steeply, meaning that all smaller sample sizes exhibit large relative errors. The same is true for mutual information, but this quantity is usually overestimated with smaller sample sizes. Finally, also for KL divergence, Sturges' is the only rule that shows a reasonable convergence behavior. The other rules result in many more bins, making it difficult for samples to match the same bin as those that are evaluated for the computation of KL divergence between the histogram of $\hat{p}(x)$ and $\hat{q}(x)$. Contrary to the previous case, in higher dimensions, not only equal binning schemes matter, but also the quality of the estimate depends on the size of the bin. As the data are sparser in higher dimensions, samples that are close together should be accounted for in the same bin, and this becomes difficult when the number of bins is too large.

k-NN-based estimation performs best among all investigated methods and quantities with similar behavior as in the previous lower-dimensional cases, where, again, $k = 1$ is the best hyperparameter for entropy and KL divergence, while $k = 15$ seems best for mutual information. For mutual information, it can also be noted that $k = 1$ is not an adequate parameter for estimation, especially for small sample sizes, as the estimator typically reports a value of 0.0.

3.11. Case 8: 10D, Multivariate Normal Distribution

Following the experiments by Wang et al. [23] and, therefore, using a similar notation as the previous case, for entropy and mutual information, data were sampled from $\mathcal{N}_{10}(\mu_0, \Sigma_0)$, and for KL divergence, data were sampled from $\mathcal{N}_{10}(\mu_1, \Sigma_1)$. In this case, both distributions are centered at 0, and we write the covariance matrices using the notation i for rows and j for columns, then $\Sigma_0^{i,i} = 1$, $\Sigma_0^{i,j} = 0.9$ and $\Sigma_1^{i,i} = 1$, $\Sigma_1^{i,j} = 0.1$, for i and $j = 1, \ldots, 10$.

3.11.1. Analytical Reference Solution

We can again use Equations (39), (40), and (46). Similar to the previous case, to calculate mutual information, Equation (45) is applied to the distribution $\mathcal{N}_{10}(\mu_0, \Sigma_0)$, and once again, the dimensions are separated so that $n = 9$ and $m = 1$, where m is the dimension of the highest order; then, mutual information is calculated as $I([x_1, \ldots, x_9]; x_{10})$. The following solutions are obtained as reference values:

$$H(X) = 4.93 \text{ nats} \quad D_{KL}(p||q) = 7.00 \text{ nats} \quad I(X;Y) = 1.10 \text{ nats}$$

3.11.2. Estimation Results

KDE-based estimation performs poorly for all quantities. For estimating entropy, the kernel-based representation of the PDF seems to not be as smooth as the true distribution, resulting in a lower estimation of entropy. Nevertheless, the representation does improve with additional samples, and it would be expected that, given enough samples, the KDE PDF would resemble the true distribution and the estimator would converge. Then, again, more samples would mean added computational cost in evaluating the kernel representation of the PDF, and this procedure becomes a limiting factor. As with the previous case, the maximum number of samples for entropy and KL divergence has been limited to 25,000. For KL divergence, the mean estimates across sample sizes are very consistent, but the kernel-based representation of both the true $p(x)$ and approximating $q(x)$ distributions do not resemble the distributions from which the data were sampled; therefore, a biased result is obtained. Finally, mutual information is consistently overestimated across all sample sizes, with added computational cost, as three kernel-based approximations of the true distribution have to be evaluated: $p(x, y)$, $p(x)$, and $p(y)$.

Considering binning methods, similar to the previous case, only Sturges' rule is able to capture an adequate representation of the underlying distribution with its larger bin widths. For entropy, as the sample size increases, the estimate made by Sturges moves toward the true value. But, even for the largest sample size of 100,000, the estimate made using this rule underestimates the true value of entropy by approximately 20%. Furthermore, when using rules with smaller bin widths, bin occupations and, hence, densities become even lower, resulting in the estimated entropy being typically negative. This can be seen in Figure 16 as the trajectory that Scott's rule follows, and this was true also for the previous case in Section 3.10. Further, for KL divergence, not even the large bins produced by Sturges' rule prevent the mismatch between the bin occupations for high-dimensional data, making the result of estimating KL divergence equivalent to infinity and not shown in Figure 16. Finally, in the case of mutual information, once again, only Sturges' rule produces an estimate that follows the expected behavior of improving as the sample size is increased. Both Scott's and FD's rule produce more bins, which typically have higher estimated densities in the joint distribution of X and Y than the product of both marginal distributions X and Y. As mutual information measures the distance between a joint distribution and the product of its marginals, as described in Equation (6), mutual information increases with the sample size, as can be seen in Figure 16, and said increase is not compensated by dividing by the size of the sample, as Equation (22) indicates.

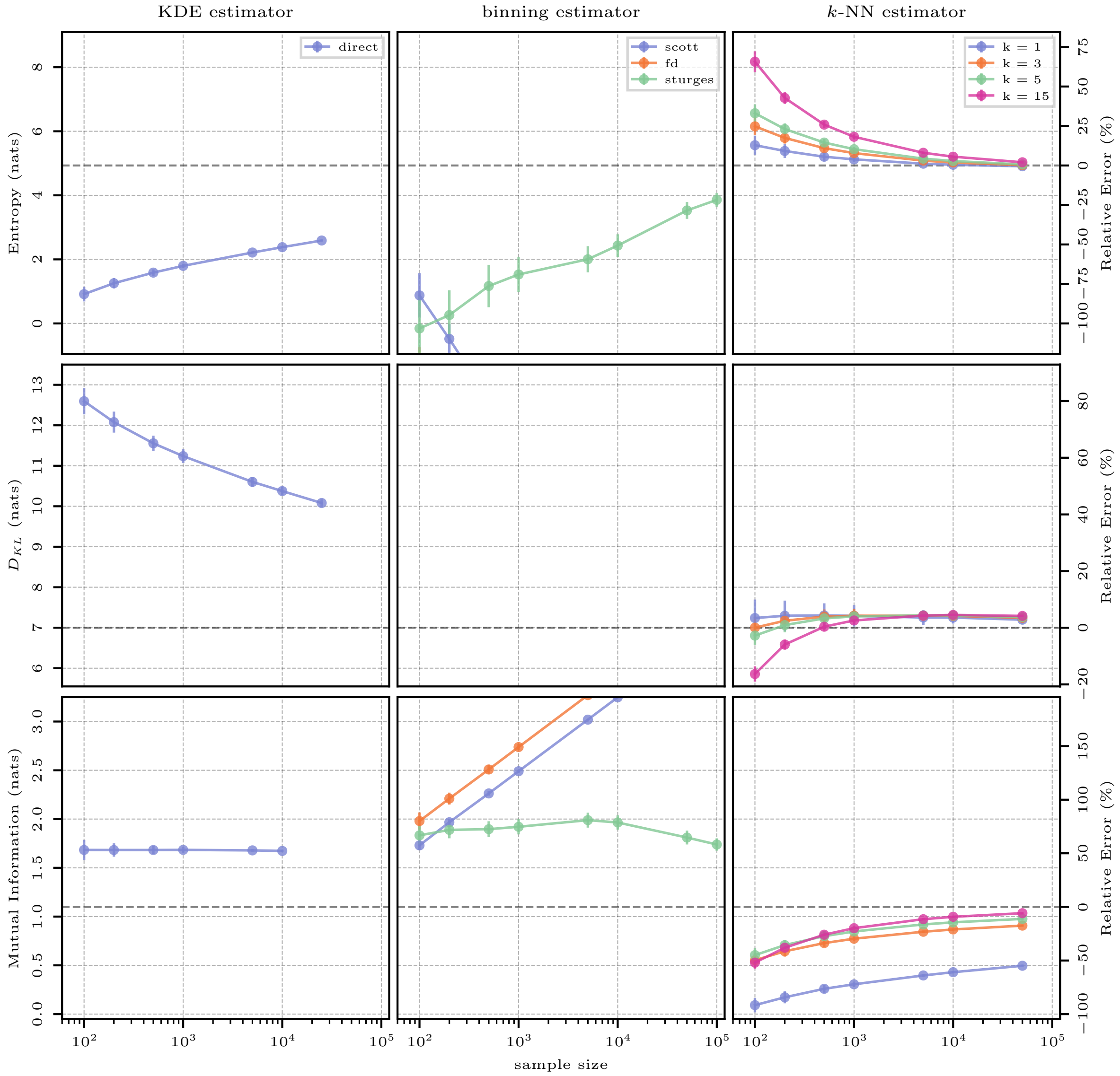

Figure 16. Evaluation of all estimation methods (columns) for entropy (**top**), KL divergence (**middle**), and mutual information (**bottom**), in Case 8 (10D, normal distribution).

k-NN-based estimation continues to perform well, but the challenges of estimation in higher dimensions also affect the results based on this method. Here, the number of samples was also limited from the typical 100,000 as the maximum evaluated to 50,000. Although not as markedly as with other methods, the computational cost increased from seconds to minutes in comparison to the previous four-dimensional case, even considering that the number of available samples was restricted. Notably, the estimator gives a much more accurate result for entropy across all sample sizes than KDE or binning for the largest implemented sample size, using the k-NN hyperparameter $k = 1$. For KL divergence and mutual information, the performance is not as accurate, even when considering an optimal choice of hyperparameters. For KL divergence, estimates using $k = 1$ show a downward trend, which would suggest that the estimator will eventually converge, but at a much higher number of available samples. The same is true for mutual information using $k = 15$.

However, the relative errors of both quantities are much smaller than those produced by KDE or binning and remarkably similar in order of magnitude to all lower-dimensional test cases.

3.12. Computational Cost

A desirable property of an estimator is a low computational time. With such a property, the estimator would lend itself to practical purposes and processes that require a result to be estimated many times such as Monte Carlo methods or bootstrapping [10].

To assess the computational cost of the different estimation methods, we chose to evaluate the computation time required for each test case for a set number of samples and one hyperparameter of each specific method. Given the results presented above, we chose to use 10,000 samples to evaluate computation time because, in most cases, at this sample size, each estimator has (almost) converged to the true result. In terms of hyperparameters, for the binned estimator, we chose to use Scott's method to calculate an appropriate bin width (see Section 2.2), Silverman's bandwidth for the KDE estimator (Equation (14)), and $k = 1$ for the k-NN estimator to calculate entropy and KL divergence, while for mutual information, we chose $k = 15$. The results in Table 2 show the mean, maximum, and minimum computation time (in seconds) for each experiment. All results were obtained on a single thread of an Intel(R) Xeon(R) CPU E5-26280 v2 with a clock speed of 2.80 GHz.

Across the board, Table 2 shows that binning- and k-NN-based methods have the lowest required computational time with a single estimate taking less than a second in most cases. This makes them ideal for tasks that require repeating calculations for different samples of data or using different hyperparameters.

For the binning estimator, the process of creating, filling, and applying a specific equation as described in Section 2.2.1 typically takes milliseconds with an increase depending on the number of dimensions of the data. While, for lower dimensions, the required histogram is calculated, for the 4- and 10-dimensional cases, we employed a procedure in which every point in the sample is replaced by the specific bin it occupies, greatly limiting the amount of memory and computational time required for estimation. Interestingly, there is a decrease in the time required to estimate KL divergence between the 4- and 10-dimensional cases. This is the product of having to find the matching and occupied bins given by Scott's rule-of-thumb for estimating bin size. As the bins are smaller, they are more sparsely populated in the 10-dimensional case, resulting in less matching bins that are also occupied. Having to calculate the estimate of KL divergence for less bins makes this process faster in the 10-dimensional case when compared against the 4-dimensional case.

As introduced by Gupta et al. [10], the QS estimator uses bootstrapping twice. From a single sample, it uses bootstrapping to determine the theoretical quantiles that best approximate the true distribution, and then, it uses bootstrapping again to estimate the bootstrapped confidence intervals of the estimate. Because the experiments wanted to address the effect of sampling variability on the estimators, the second step of the QS estimator was not performed and bootstrapping was only performed to determine the quantiles. This is reflected in the computational times shown in Table 2 as the QS estimator takes longer than histogram-based methods using rules-of-thumb to determine the bin width. As described in Section 2.2.2, bootstrapping was performed $N_k = 500$ times to determine the ideal quantiles.

KDE-based estimation provides the estimates with the highest computation time out of all the estimation methods evaluated. Calculating a direct resubstitution estimate for all quantities is one or two orders of magnitude slower than using any of the previously discussed methods. This becomes even worse when numerical integration is used to calculate an estimate. For a resubstitution estimate, each data point in the sample has to be evaluated once in the KDE-based representation of the PDF. This is opposed to performing numerical integration where the number of evaluations is unknown and they continue until a certain tolerance for error in the method is met. This is apparent as, for lower-dimensional cases, the number of required evaluations for the numerical integration estimate is low, making this method faster than evaluating all available samples in the kernel-based PDF for a resubstitution estimate. Nevertheless, due to its high computation time, the KDE-based estimator does not seem suitable for multiple evaluations.

Finally, k-NN estimation requires about 10-times the computational effort of binning in low dimensions, but it's faster than direct KDE evaluation by about a factor of 1000. In high-dimensional cases, the effort of k-NN increases, but still is well below the effort of the direct evaluation of KDE.

Table 2. Comparison of computational time between cases and methods (time in seconds).

Quantity	Case	Bins		KDE		k-NN
		Scott	**QS**	**Direct**	**Integration**	
Entropy	Uniform	$0.001^{0.012}_{0.001}$	$0.263^{0.556}_{0.137}$	$3.408^{4.078}_{1.643}$	$0.122^{0.290}_{0.091}$	$0.018^{0.046}_{0.014}$
	Normal	$0.001^{0.014}_{0.001}$	$0.278^{0.455}_{0.210}$	$3.467^{4.058}_{2.652}$	$0.286^{0.672}_{0.197}$	$0.018^{0.043}_{0.014}$
	Normal Mixture	$0.001^{0.001}_{0.001}$	$0.285^{0.505}_{0.225}$	$3.486^{4.196}_{2.795}$	$0.245^{0.604}_{0.156}$	$0.018^{0.040}_{0.014}$
	2D Normal	$0.003^{0.012}_{0.002}$	-	$3.513^{4.176}_{2.861}$	$80.09^{112.1}_{56.75}$	$0.023^{0.077}_{0.017}$
	2D Normal Mixture	$0.002^{0.014}_{0.002}$	-	$3.525^{4.048}_{2.667}$	$37.23^{49.74}_{21.06}$	$0.022^{0.056}_{0.018}$
	Gamma-Exponential	$0.003^{0.005}_{0.002}$	-	$3.548^{4.012}_{3.338}$	$279.2^{633.6}_{76.25}$	$0.023^{0.064}_{0.012}$
	4D Normal	$0.008^{0.015}_{0.006}$	-	$4.052^{4.896}_{2.825}$	-	$0.043^{0.115}_{0.034}$
	10D Normal	$0.016^{0.039}_{0.010}$	-	$5.193^{6.357}_{4.120}$	-	$0.597^{0.937}_{0.442}$
D_{KL}	Uniform	$0.002^{0.011}_{0.001}$	-	$4.712^{4.755}_{4.668}$	$0.524^{0.952}_{0.365}$	$0.034^{0.087}_{0.027}$
	Normal	$0.002^{0.011}_{0.001}$	-	$4.715^{4.783}_{4.596}$	$0.837^{1.395}_{0.554}$	$0.033^{0.083}_{0.027}$
	Normal Mixture	$0.002^{0.002}_{0.001}$	-	$4.713^{4.782}_{4.597}$	$0.581^{1.087}_{0.365}$	$0.033^{0.074}_{0.017}$
	2D Normal	$0.004^{0.014}_{0.002}$	-	$4.877^{4.921}_{4.767}$	$250.5^{373.3}_{187.3}$	$0.044^{0.135}_{0.023}$
	2D Normal Mixture	$0.004^{0.016}_{0.004}$	-	$4.875^{4.930}_{4.756}$	$168.4^{238.4}_{122.5}$	$0.042^{0.084}_{0.034}$
	Gamma-Exponential	$0.005^{0.014}_{0.003}$	-	$4.942^{4.985}_{4.907}$	$381.4^{494.7}_{228.4}$	$0.043^{0.094}_{0.023}$
	4D Normal	$0.192^{0.412}_{0.128}$	-	$5.363^{5.409}_{5.304}$	-	$0.093^{0.179}_{0.067}$
	10D Normal	$0.057^{0.132}_{0.044}$	-	$6.960^{6.994}_{6.900}$	-	$1.012^{1.077}_{0.949}$
Mutual Information	2D Normal	$0.011^{0.035}_{0.006}$	-	$10.40^{11.43}_{7.998}$	$224.6^{361.8}_{133.9}$	$0.175^{0.342}_{0.082}$
	2D Normal Mixture	$0.012^{0.044}_{0.008}$	-	$10.27^{11.31}_{9.553}$	$165.2^{362.3}_{94.73}$	$0.170^{0.236}_{0.082}$
	Gamma-Exponential	$0.016^{0.035}_{0.006}$	-	$10.19^{11.59}_{9.919}$	$938.9^{3271.}_{244.7}$	$0.177^{0.256}_{0.141}$
	4D Normal	$0.017^{0.025}_{0.015}$	-	$8.347^{12.02}_{5.171}$	-	$0.601^{1.060}_{0.466}$
	10D Normal	$0.025^{0.040}_{0.024}$	-	$6.232^{6.260}_{6.213}$	-	$3.218^{3.959}_{2.473}$

3.13. Synthesis of Findings from Test Cases

Histogram-based estimation or binning is the most common approach for calculating entropy, KL divergence, or mutual information. In this study, we found that it is an accurate method especially for data in one dimension and using Scott's rule-of-thumb to determine an adequate binning scheme for the data. Nevertheless, the method loses accuracy for data in higher dimensions, where there are no established good practices for selecting a particular bin size, and the rules-of-thumb extrapolate poorly to higher dimensions. Only Sturges' rule, which calculates the largest bin widths, seems to generalize to higher dimensions, as the available data for estimation are sparser and less bins are able to capture an adequate number of points in a particular bin.

As an alternative, KDE-based estimation uses a kernel to create a smooth representation of a probability density function for each point in the available sample. The performance of this method is good when compared to the other estimation methods, but it might not be practical due to its high computational cost and the fact that, as implemented in this study, it used a Gaussian kernel and it was applied to data sampled from normal distributions (multivariate normal distributions in higher dimensions). Extrapolating from the lower-dimensional test cases, we reckon that, for high-dimensional distributions with different shapes, this estimation method might not perform as well.

Across all of the experiments, k-NN-based estimation was among the top-scoring methods, independent of the shape and dimensionality of the distribution sampled from, and the given sample size. We found a very consistent best-performing choice of the hyperparameter k, i.e., the number of neighbors to consider in the formulation of the estimator. Notably, the optimal choice of k depends on the specific quantity to be estimated: for entropy and KL divergence, $k = 1$ performed best, while for mutual information, a high number of $k = 15$ was shown to be optimal. Higher values of k were tested, but did not show significant improvement, while using $k = 1$ for mutual information is generally not recommended. These findings suggest that a direct estimation in the immediate vicinity of each sample point is beneficial for estimating entropy (thereby, not smearing out the sample distribution too much) and KL divergence (where the ratios of probabilities are estimated per sampling point, thereby not accumulating approximation errors before calculating the ratios). For mutual information, a more aggregated view of the sample is required, favoring a higher number of neighbors to consider.

Moreover, the computational cost is relatively low, as described in Section 3.12. Hence, k-NN-based estimation is readily implementable in practical applications and lends itself to repeated sampling such as in Monte Carlo or bootstrapping analyses.

We found that KDE and binning both performed well in some cases, but by far not as consistently as k-NN and with varying methods (e.g., direct estimation vs. integration for KDE and Scott's rule vs. Sturges' and FD's rule for binning). Hence, the clear recommendation on how to set the value of k-NN's hyperparameter is a further advantage of this method.

4. Summary, Conclusions, and Outlook

Estimating information-theoretic quantities from sample data, possibly in higher dimensions, poses a challenge. Most methods rely on an initial density estimation step, but density estimation itself is known to be computationally demanding or even prohibitive, and to produce unreliable results with unknown accuracy and precision in practical settings. Further, most methods assume a certain shape of distribution that the sample stems from, with an unknown impact on the estimator's performance if the true distribution deviates from that assumption. As an alternative, nearest-neighbor-based methods to directly estimate specific information-theoretic quantities have been proposed, skipping the initial density estimation step altogether. Hence, they show promise to also perform well for higher dimensions and arbitrary distributions, but have never been systematically compared to density-based methods such as kernel density estimation (KDE) or binning (histogram-based schemes). In fact, most users in science and practice seem to favor binning for its straightforward implementation, but with no further justification. We hypothesize that this is due to a lack of systematic evaluation and guidance on which method to choose in what settings. With this investigation, we aimed to close this research gap.

Typically, a new method was introduced individually without systematic comparative analysis in regard to other methods. In some cases, a brief comparison was made against another method that uses the same basic principle, such as histograms or k-NN. In response, we presented and discussed three of the most widely used non-parametric estimation methods for information-theoretic quantities, namely binning, KDE, and k-nearest neighbors (k-NN). To evaluate the estimators' performance, we designed test cases that used data samples from distributions with different shapes and with different numbers of dimensions to quantify the information-theoretic quantities entropy, KL divergence, and mutual information. Depending on the level of complexity of each case, analytical solutions existed, or were approximated with a high-quality numerical reference solution. We tested the estimation methods on each case and reported the performance in a chart that serves for intercomparing all methods for all target quantities as a function of sample size. The true distributions used to generate the data for our experiments ranged from simple 1D uniform or normal distributions over more skewed or bimodal shapes up to multivariate distributions in 4 and 10 dimensions. Sample sizes ranged from 100 to 100,000, with a reduction of the maximum sample size in cases where computational effort exploded. As described, this is a very practical and easy-to-understand basis for our study. We also accounted for sampling variability as each experiment was repeated with 300 random seeds for the data-generating sampling procedure. Further, we considered different choices of hyperparameters for each estimation method. Finally, we assessed and compared the computation time required to obtain an estimate for a sample size of 10,000, which produced a well-converged estimate for most estimation methods. All methods showed larger relative errors for sample sizes typically below 1000, pointing to the fact that samples of such a small size are not representative enough of the underlying true distribution, and hence, any estimation method necessarily fails in reliably quantifying the underlying distribution's particular property. For larger sample sizes, however, estimation results typically converged to a more or less biased estimate, with distinct differences between the methods.

For binning, the most important parameter is the bin width or Δ used to build the histogram because it controls the trade-off between oversmoothing and undersmoothing the data with respect to the true distribution. Typical rules-of-thumb were implemented and tested in this study (namely Scott's, Sturges', and FD's rule), as well as the Quantile Spacing method by Gupta et al. [10], but alternative methods using piecewise constant approximations for density [47], as well as methods that calculate an estimate of Δ based on the minimization of the mean-integrated-squared error (MISE) should be explored as well in the future. These latter methods aim to reduce the error between the true distribution $f(x)$ and the histogram-based distribution $\hat{f}(x|\Delta)$ [48].

Our KDE-based estimator used a Gaussian kernel as a typical choice; however, the choice of kernel is a hyperparameter that was not analyzed in this study, and it is expected that the performance of the estimator would vary depending on the choice of kernel used. A boxcar kernel, for example, would improve estimation in cases where the true distribution has distinct bounds such as a uniform distribution in one dimension, or a gamma-exponential distribution in two dimensions. However, such improvement would be accompanied by the known high computational cost of kernel-based methods, and as such, if only the particular information-theoretic quantity is of interest and no other property related to the kernel-based representation of the PDF, other more efficient methods seem better suited.

Finally, the evaluation of the k-NN-based estimator proved its performance and served as a showcase for its computational efficiency. The speed of the estimator comes from the usage of the k-d Tree data structure, which is in charge of the look-up operation required to identify nearest neighbors. While we found that, in its original form, the estimator performs well in the diverse set of experiments we tested it on, recent variations have been proposed to further improve the estimation of entropy by combining k-NN with normalizing flows [49] and the estimation of mutual information by combining k-NN with KDE [22] or using neural networks [50].

In general, across all test cases, k-NN produced among the lowest relative errors in estimating entropy, KL divergence, and mutual information, independent of the sample size and shape and dimensionality of the distribution sampled from. Also, a clear identification of the best-performing value of the hyperparameter k, i.e., the number of neighbors to be considered in the vicinity of a sampling point, was possible; interestingly, the optimal choice of k depends on the specific quantity to be estimated. For entropy and KL divergence, $k = 1$ performed best, while for mutual information, a more integral approximation with a high number of $k = 15$ was shown to be optimal (even higher values were tested, but did not show significant further improvement).

All tested methods might benefit from a specifically tailored bootstrapping approach to increase the stability of the estimator for large samples that are (sufficiently) representative of the true underlying distribution; this is recommended for further analysis. For small sample sizes, however, our study has confirmed the dominant impact of non-representativeness.

Based on our evaluation of the performance, ease of implementation, and computational effort, we recommend k-NN-based estimation for estimating information-theoretic quantities from sample data, especially in higher dimensions, due to its clear advantages. However, the significance of this work extends beyond a theoretical comparison and practical performance assessment: we have also collected these methods in a publicly available Python 3 toolbox, ensuring transparency and accessibility for the wider research community. This toolbox shall serve as a valuable resource, enabling researchers and practitioners to integrate information-theoretic concepts into their data analysis and modeling workflows. Specifically, this study in combination with the toolbox shall enable them to make a well-informed decision on the choice of estimation method.

Author Contributions: Conceptualization, M.Á.C., H.V.G., U.E. and A.G.; methodology, M.Á.C., H.V.G. and A.G.; software, M.Á.C. and U.E.; validation, M.Á.C.; formal analysis, M.Á.C., H.V.G., U.E. and A.G.; investigation, M.Á.C., H.V.G., U.E. and A.G.; writing—original draft preparation, M.Á.C.; writing—review and editing, M.Á.C., H.V.G., U.E. and A.G.; visualization, M.Á.C. and A.G.; supervision, A.G.; project administration, A.G.; funding acquisition, A.G. and U.E. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) both under Germany's Excellence Strategy—EXC 2075–390740016 and the project 507884992.

Data Availability Statement: The different estimation methods were implemented using Python 3 and are collected in the *UNITE toolbox*, which can be found in the following public repository: https://pypi.org/project/unite-toolbox (accessed on 24 April 2024). The scripts used in the workflow to generate the data for each of the experiments, the evaluation of each estimator on the data for each experiment, and results, as well as the log files can be found in this repository: https://doi.org/10.18419/darus-4087 (accessed on 24 April 2024).

Acknowledgments: We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech) and thank the two anonymous reviewers for their valuable feedback on our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KDE	kernel density estimation
D_{KL}	Kullback–Leibler divergence
KL divergence	Kullback–Leibler divergence
k-NN	k-nearest neighbors

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
2. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
3. MacKay, D.J.C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2003.

4. Ruddell, B.L.; Brunsell, N.A.; Stoy, P. Applying Information Theory in the Geosciences to Quantify Process Uncertainty, Feedback, Scale. *Eos Trans. Am. Geophys. Union* **2013**, *94*, 56. [CrossRef]
5. Nowak, W.; Guthke, A. Entropy-Based Experimental Design for Optimal Model Discrimination in the Geosciences. *Entropy* **2016**, *18*, 409. [CrossRef]
6. Timme, N.M.; Lapish, C. A Tutorial for Information Theory in Neuroscience. *eNeuro* **2018**, *5*. [CrossRef] [PubMed]
7. Yang, J. Information Theoretic Approaches in Economics. *J. Econ. Surv.* **2018**, *32*, 940–960. [CrossRef]
8. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **2014**, *36*, 119–140. [CrossRef]
9. Beirlant, J.; Dudewicz, E.J.; Györfi, L.; Van der Meulen, E.C. Nonparametric Entropy Estimation: An Overview. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17–39.
10. Gupta, H.V.; Ehsani, M.R.; Roy, T.; Sans-Fuentes, M.A.; Ehret, U.; Behrangi, A. Computing Accurate Probabilistic Estimates of One-D Entropy from Equiprobable Random Samples. *Entropy* **2021**, *23*, 740. [CrossRef]
11. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Number 26 in Monographs on Statistics and Applied Probability; Chapman & Hall/CRC: Boca Raton, FL, USA, 1998.
12. Scott, D.W.; Sain, S.R. Multidimensional Density Estimation. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2005; Volume 24, pp. 229–261. [CrossRef]
13. Joe, H. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Stat. Math.* **1989**, *41*, 683–697. [CrossRef]
14. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics (ISS); Springer: New York, NY, USA, 2006.
15. Bossomaier, T.; Barnett, L.; Harré, M.; Lizier, J.T. *An Introduction to Transfer Entropy*; Springer International Publishing: Cham, Switzerland, 2016. [CrossRef]
16. Liu, H.; Lafferty, J.; Wasserman, L. Sparse Nonparametric Density Estimation in High Dimensions Using the Rodeo. In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; PMLR: San Juan, PR, USA 2007; pp. 283–290.
17. Thiesen, S.; Vieira, D.M.; Mälicke, M.; Loritz, R.; Wellmann, J.F.; Ehret, U. Histogram via entropy reduction (HER): An information-theoretic alternative for geostatistics. *Hydrol. Earth Syst. Sci.* **2020**, *24*, 4523–4540. [CrossRef]
18. Goodwell, A.E.; Bassiouni, M. Source Relationships and Model Structures Determine Information Flow Paths in Ecohydrologic Models. *Water Resour. Res.* **2022**, *58*, e2021WR031164. [CrossRef]
19. Kim, J.; Kim, G.; An, S.; Kwon, Y.K.; Yoon, S. Entropy-Based Analysis and Bioinformatics-Inspired Integration of Global Economic Information Transfer. *PLoS ONE* **2013**, *8*, e51986. [CrossRef] [PubMed]
20. Madulara, M.D.; Francisco, P.A.B.; Nawang, S.; Arogancia, D.C.; Cellucci, C.J.; Rapp, P.E.; Albano, A.M. EEG transfer entropy tracks changes in information transfer on the onset of vision. *Int. J. Mod. Phys. Conf. Ser.* **2012**, *17*, 9–18. [CrossRef]
21. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
22. Piras, D.; Peiris, H.V.; Pontzen, A.; Lucie-Smith, L.; Guo, N.; Nord, B. A robust estimator of mutual information for deep learning interpretability. *Mach. Learn. Sci. Technol.* **2023**, *4*, 025006. [CrossRef]
23. Wang, Q.; Kulkarni, S.R.; Verdu, S. Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances. *IEEE Trans. Inf. Theory* **2009**, *55*, 2392–2405. [CrossRef]
24. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
25. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
26. Moon, Y.I.; Rajagopalan, B.; Lall, U. Estimation of mutual information using kernel density estimators. *Phys. Rev. E* **1995**, *52*, 2318–2321. [CrossRef]
27. Scott, D.W. On optimal and data-based histograms. *Biometrika* **1979**, *66*, 605–610. [CrossRef]
28. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*; Springer Texts in Statistics; Springer: New York, NY, USA, 2004.
29. Devroye, L.; Lugosi, G. Bin width selection in multivariate histograms by the combinatorial method. *Test* **2004**, *13*, 129–145. [CrossRef]
30. Muto, K.; Sakamoto, H.; Matsuura, K.; Arima, T.; Okada, M. Multidimensional Bin-Width Optimization for Histogram and Its Application to Four-Dimensional Neutron Inelastic Scattering Data. *J. Phys. Soc. Jpn.* **2019**, *88*, 044002. [CrossRef]
31. Sturges, H.A. The Choice of a Class Interval. *J. Am. Stat. Assoc.* **1926**, *21*, 65–66. [CrossRef]
32. Freedman, D.; Diaconis, P. On the histogram as a density estimator: L2 theory. *Z. FüR Wahrscheinlichkeitstheorie Und Verwandte Geb.* **1981**, *57*, 453–476. [CrossRef]
33. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2016.
34. Kozachenko, L.F.; Leonenko, N.N. A statistical estimate for the entropy of a random vector. *Probl. Inf. Transm.* **1987**, *23*, 95–101.
35. Delattre, S.; Fournier, N. On the Kozachenko–Leonenko entropy estimator. *J. Stat. Plan. Inference* **2017**, *185*, 69–93. [CrossRef]
36. Ahmad, I.; Lin, P.E. A nonparametric estimation of the entropy for absolutely continuous distributions (Corresp.). *IEEE Trans. Inf. Theory* **1976**, *22*, 372–375. [CrossRef]
37. Gao, W.; Oh, S.; Viswanath, P. Demystifying Fixed k-Nearest Neighbor Information Estimators. *IEEE Trans. Inf. Theory* **2018**, *64*, 5629–5661. [CrossRef]

38. Piessens, R. (Ed.) *QUADPACK: A Subroutine Package for Automatic Integration*; Number 1 in Springer Series in Computational Mathematics; Springer: Berlin/Heidelberg, Germany, 1983. [CrossRef]
39. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
40. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
41. Harris, C.R.; Millman, K.J.; Walt, S.J.v.d.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
42. Huber, M.F.; Bailey, T.; Durrant-Whyte, H.; Hanebeck, U.D. On entropy approximation for Gaussian mixture random vectors. In Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, Seoul, Republic of Korea, 20–22 August 2008; IEEE: Seoul, Republic of Korea, 2008; pp. 181–188. [CrossRef]
43. Zhang, Y.; Pan, J.; Li, L.K.; Liu, W.; Chen, Z.; Liu, X.; Wang, J. On the Properties of Kullback-Leibler Divergence between Multivariate Gaussian Distributions. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 58152–58165.
44. Gelfand, I.M.; Yaglom, A.M. Amount of Information and Entropy for Continuous Distributions. In *Selected Works of A. N. Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*; Shiryayev, A.N., Ed.; Mathematics and Its Applications; Springer: Dordrecht, The Netherlands, 1993; pp. 33–56. [CrossRef]
45. Darbellay, G.; Vajda, I. Entropy expressions for multivariate continuous distributions. *IEEE Trans. Inf. Theory* **2000**, *46*, 709–712. [CrossRef]
46. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Genton, M.G. Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions. *Scand. J. Stat.* **2013**, *40*, 42–62. [CrossRef]
47. Knuth, K.H. Optimal data-based binning for histograms and histogram-based probability density models. *Digit. Signal Process.* **2019**, *95*, 102581. [CrossRef]
48. Wand, M.P. Data-Based Choice of Histogram Bin Width. *Am. Stat.* **1997**, *51*, 59. [CrossRef]
49. Ao, Z.; Li, J. Entropy Estimation via Normalizing Flow. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 9990–9998. [CrossRef]
50. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR: San Juan, PR, USA, 2018; pp. 531–540.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.