

Article

Computing Accurate Probabilistic Estimates of One-D Entropy from Equiprobable Random Samples

Hoshin V. Gupta ^{1,*}, Mohammad Reza Ehsani ¹, Tirthankar Roy ², Maria A. Sans-Fuentes ³, Uwe Ehret ⁴
and Ali Behranghi ¹

¹ Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ 85721, USA; rehsani@email.arizona.edu (M.R.E.); behranghi@email.arizona.edu (A.B.)

² Civil and Environmental Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA; roy@unl.edu

³ GIDP Statistics and Data Science, The University of Arizona, Tucson, AZ 85721, USA; sans@email.arizona.edu

⁴ Institute of Water and River Basin Management, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany; uwe.ehret@kit.edu

* Correspondence: hoshin@email.arizona.edu

Abstract: We develop a simple Quantile Spacing (QS) method for accurate probabilistic estimation of one-dimensional entropy from equiprobable random samples, and compare it with the popular Bin-Counting (BC) and Kernel Density (KD) methods. In contrast to BC, which uses equal-width bins with varying probability mass, the QS method uses estimates of the quantiles that divide the support of the data generating probability density function (pdf) into equal-probability-mass intervals. And, whereas BC and KD each require optimal tuning of a hyper-parameter whose value varies with sample size and shape of the pdf, QS only requires specification of the number of quantiles to be used. Results indicate, for the class of distributions tested, that the optimal number of quantiles is a *fixed fraction* of the sample size (empirically determined to be $\sim 0.25\text{--}0.35$), and that this value is relatively insensitive to distributional form or sample size. This provides a clear advantage over BC and KD since hyper-parameter tuning is not required. Further, unlike KD, there is no need to select an appropriate kernel-type, and so QS is applicable to pdfs of arbitrary shape, including those with discontinuous slope and/or magnitude. Bootstrapping is used to approximate the sampling variability distribution of the resulting entropy estimate, and is shown to accurately reflect the true uncertainty. For the four distributional forms studied (*Gaussian, Log-Normal, Exponential* and *Bimodal Gaussian Mixture*), expected estimation bias is less than 1% and uncertainty is low even for samples of as few as 100 data points; in contrast, for KD the small sample bias can be as large as -10% and for BC as large as -50% . We speculate that estimating quantile locations, rather than bin-probabilities, results in more efficient use of the information in the data to approximate the underlying shape of an unknown data generating pdf.

Keywords: entropy; estimation; quantile spacing; accuracy; uncertainty; bootstrap; small-sample efficiency



Citation: Gupta, H.V.; Ehsani, M.R.; Roy, T.; Sans-Fuentes, M.A.; Ehret, U.; Behranghi, A. Computing Accurate Probabilistic Estimates of One-D Entropy from Equiprobable Random Samples. *Entropy* **2021**, *23*, 740. <https://doi.org/10.3390/e23060740>

Academic Editor: Geert Verdoolaege

Received: 30 March 2021

Accepted: 8 June 2021

Published: 11 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Consider a data generating process $p(x)$ from which a finite size set of N_S random, equiprobable, independent identically distributed (iid) samples $S = \{s_i, i = 1 \dots N_S\}$ is drawn. In general, we may not know the nature and mathematical form of $p(x)$, and our goal is to compute an estimate $\hat{H}_p(X|S)$ of the Entropy $H_p(X)$ of $p(x)$.

In the idealized case, where X is a one-dimensional continuous random variable and the parametric mathematical form of $p(x)$ is known, we can apply the definition of *differential Entropy* [1,2] to compute:

$$H_p(X) = \mathbb{E}_p\{-\ln(p(x))\} = \int_{-\infty}^{+\infty} -\ln(p(x)) \cdot p(x) \cdot dx \quad (1)$$

Explicit, closed form solutions for $H_p(X)$ are available for a variety of probability density functions (pdfs). For a variety of others, closed form solutions are not available, and one can compute $H_p(X)$ via numerical integration of Equation (1). In all such cases, entropy estimation consists of first obtaining estimates $\hat{\theta}|S$ of the parameters θ of the known parametric density $p(x|\theta)$ and then computing the entropy estimate $\hat{H}_{p|\hat{\theta}}(X|S)$ by plugging $p(x|\hat{\theta})$ into Equation (1). Any bias and uncertainty in the entropy estimate will depend on the accuracy and uncertainty of the parameter estimates $\hat{\theta}$. If the form of $p(x|\theta)$ is “assumed” rather than explicitly known, then additional bias will stem from the inadequacy of this assumption.

In most practical situations the mathematical form of $p(x)$ is not known, and S must first be used to obtain a data-based estimate $\hat{p}(x|S)$, from which an estimate $\hat{H}_{\hat{p}}(X|S)$ can be obtained via numerical integration of Equation (1). In generating $\hat{p}(x|S)$, consistency with prior knowledge regarding the nature of $p(x)$ must be ensured—for example, X may be known to take on only positive values, or values on some finite range. Consistency must also be maintained with the information in S . Further, the sample size N_S must be sufficiently large that the information in S provides an accurate characterization of $p(x)$ —in other words, that S is informationally representative and consistent.

To summarize, for the case that X is a continuous random variable, entropy estimation from data involves two steps; (i) Use of S to estimate $\hat{p}(x|S)$, and (ii) Numerical integration to compute an estimate of entropy using Equation (2):

$$\hat{H}_{\hat{p}}(X|S) = \mathbb{E}_{\hat{p}}\{-\ln(\hat{p}(x|S))\} = \int_{-\infty}^{+\infty} -\ln(\hat{p}(x|S)) \cdot \hat{p}(x|S) \cdot dx \quad (2)$$

Accordingly, the estimate $\hat{H}_{\hat{p}}(X|S)$ has two potential sources of error. One is due to the use of $\hat{p}(x|S)$ to approximate $p(x)$, and the other is due to imperfect numerical integration. To maximize accuracy, we must ensure that both these errors are minimized. Further, $\hat{H}_{\hat{p}}(X|S)$ is a statistic that is subject to inherent random variability associated with the sample S , and so it will be useful to have an uncertainty estimate, in some form such as confidence intervals.

For cases where X is discrete and can take on only a finite set of values $\{x^{(j)}, j = 1, \dots, N_X\}$, if the mathematical form of $p(x) = \{p(x^{(j)}), j = 1, \dots, N_X\}$ is known, then $H_p(X)$ can be computed by applying the mathematical definition of *discrete* Entropy:

$$H_p(X) = \mathbb{E}_p\{-\ln(p(x))\} = -\sum_{j=1}^{N_X} \ln(p(x^{(j)})) \cdot p(x^{(j)}) \quad (3)$$

Here, given a data sample S , pdf estimation amounts simply to counting the number $n^{(j)}$ of data points in S that take on the value $x^{(j)}$, and setting $\hat{p}(x|S) = \{\hat{p}(x^{(j)}|S) = \frac{n^{(j)}}{N_S}, j = 1, \dots, N_X\}$ and $\hat{p}(x|S) = 0$ otherwise. Entropy estimation then consists of applying the equation:

$$\hat{H}_{\hat{p}}(X|S) = -\sum_{j=1}^{N_X} \ln(\hat{p}(x^{(j)}|S)) \cdot \hat{p}(x^{(j)}|S) \quad (4)$$

In this case, there is no numerical integration error; any bias in the estimate is entirely due to $\hat{p}(x^{(j)}|S) \neq p(x^{(j)})$, which occurs due to S not being perfectly informative about $p(x)$, while uncertainty is due to S being a random sample drawn from $p(x)$. If S is a representative sample, as $N_S \rightarrow \infty$ then $\hat{p}(x^{(j)}|S) \rightarrow p(x^{(j)})$ and hence $\hat{H}_{\hat{p}}(X|S) \rightarrow H_p(X)$, so that estimation bias and uncertainty will both tend towards zero as the sample size is increased.

When the one-dimensional random variable X is some hybrid combination of discrete and continuous, the relative fractions of total probability mass associated with the discrete

and continuous portions of the pdf must also be estimated. The general principles discussed herein also apply to the hybrid case, and we will not consider it further in this paper; for a relevant discussion of estimating entropy for mixed discrete-continuous random variables, see [3].

2. Popular Approaches to Estimating Distributions from Data

We focus here on the case of a one-dimensional continuous random variable X for which the mathematical form of $p(x)$ is unknown. [4] provides a summary of methods for the estimation of pdfs from data, while [5] provides an overview of methods for estimating the differential entropy of a continuous random variable. The three most widely used “non-parametric” methods for estimating differential entropy by pdf approximation are the: (a) Bin-Counting (BC) or piece-wise constant frequency histogram method, (b) Kernel Density (KD) method, and (c) Average Shifted Histogram (ASH) method. Ref [6] points out that these can all be asymptotically viewed as “Kernel” methods, where the bins in the BC and ASH approaches are understood as treating the data points falling in each bin as being from a locally uniform distribution. For a theoretical discussion of the basis for non-parameteric estimation of a density function see also [7].

As discussed by [8,9], appropriate selection of the bin-width (effectively a smoothing hyper-parameter) is critical to success of the BC and ASH histogram-based methods. Bin-widths that are too small can result in overly rough approximation of the underlying distribution (increasing the variance), while bin-widths that are overly large can result in an overly smooth approximation (introducing bias). Therefore, one typically has to choose values that balance variance and bias errors. [8] and [10] present expressions for “optimal” bin width when using BC, including the “normal reference rule” that is applicable when the pdf is approximately Gaussian, and the “oversmoothed bandwidth rule” that places an upper-bound on the bin-width. Similarly, [6] shows that while KD is more computationally costly to implement than BC, its accuracy and convergence are better, and they derive optimal values for the KD smoothing hyper-parameter. [11] also proposed the ASH method, which refines BC by sub-dividing each histogram bin into sub-bins, with computational cost similar to BC and accuracy approaching that of KD. Note that, if prior information on the shape of $p(x)$ is available, or if a representation with the smallest number of bins is desired, then variable bin width methods may be more appropriate (e.g., [12–16]).

The BC, KD and ASH methods all require hyper-parameter tuning to be successful. BC requires selection of the histogram bin-widths (and thereby the number of bins), KD requires selection of the form of the Kernel function and tuning of its parameters, and ASH requires selection of the form of a Kernel function and tuning of the coarse bin width, and number of sub-bins. While recommendations are provided to guide the selection of these “hyper-parameters”, these recommendations depend on theoretical arguments based in assumptions regarding the typical underlying forms of $p(x)$. Based on empirical studies, and given that we typically do not know the “true” form of $p(x)$ to be used as a reference for tuning, [3] recommend use of BC and KD rather than the ASH method.

Finally, since BC effectively treats the pdf as being discrete, and therefore uses Equation (4) with each of the indices j corresponding to one of the histogram bins, the loss of entropy associated with implementing the discrete constant bin-width approximation is approximately $\ln(\Delta)$, where Δ is the bin width, provided Δ is sufficiently small [2]. This fact allows conversion of the discrete entropy estimate to differential entropy simply by adding $\ln(\Delta)$ to the discrete entropy estimate.

In summary, while BC and KD can be used to obtain accurate estimates of entropy for pdfs of arbitrary form, hyper-parameter tuning is required to ensure that good results are obtained. In the next section, we propose an alternative method to approximate $p(x)$ that does not require counting the numbers of samples in “bins”, and is instead based on estimating the quantile positions of $p(x)$. We first compare the properties and performance of the method with BC (rather than with KD or ASH) for a range of distributions because (i) BC is arguably the most straightforward and popular method, (ii) BC shares important

similarities with the proposed approach, but also shows distinct differences worthy of discussion, and (iii) we seek good performance for a wide range of distributions. At the request of a reviewer, we further report performance of KD on the same distributions, reinforcing the findings of [17] that KD performance can depend strongly on distribution shape and sample size.

3. Proposed Quantile Spacing (QS) Approach

We present an approach to computing an estimate $\hat{H}_p(X|S)$ of Entropy $H_p(X)$ given a set of available samples S for the case where X is a one-dimensional continuous random variable and the mathematical form of the data generating process $p(x)$ is unknown. The approach is based in the assumption that $p(x)$ can be approximated as piecewise constant on the intervals between quantile locations, and consists of three steps.

3.1. Step 1—Assumption about Support Interval

The first step is to assume that $p(x)$ exists only on some finite support interval $[x_{min}, x_{max}]$, where $x_{min} \leq \min\{S\}$ and $x_{max} \geq \max\{S\}$; i.e., we treat $p(x)$ as being 0 everywhere outside of the interval $[x_{min}, x_{max}]$. Given that the true support of X may, in reality, be as extensive as $[-\infty, +\infty]$, we allow the selection of this interval (based on prior knowledge, such as physical realism) to be as extensive as appropriate and/or necessary. However, as we show later, the impact of this selection can be quite significant and will need special attention.

3.2. Step 2—Assumption about Approximate Form of $p(X)$

The second step is to assume that $p(x)$ can be approximated as piecewise constant on the intervals between quantiles $Z = \{z_0, z_1, z_2, \dots, z_{N_Z}\}$ associated with the $\left\{0, \frac{1}{N_Z}, \frac{2}{N_Z}, \dots, \frac{N_Z-1}{N_Z}, 1\right\}$ non-exceedance probabilities of $p(x)$, where N_Z represents the number of quantiles, $z_0 = x_{min}$, and $z_{N_Z} = x_{max}$. This corresponds to making the *minimally informative* (maximum entropy) assumption that $p(x)$ is 'uniform' over each of the quantile intervals $[z_{j-1}, z_j]$ for $j = 1, \dots, N_Z$, which is equivalent to assuming that the corresponding cumulative distribution function $P(x)$ is piecewise linear (i.e., increases linearly between z_{j-1} and z_j).

Assuming perfect knowledge of the locations of the quantiles Z , this approximation corresponds to:

$$p(x) \approx \hat{p}(x|Z) = p_{j-1}^j = \frac{K}{\Delta_j} \text{ for } z_{j-1} \leq X < z_j, j = 1, \dots, N_Z \quad (5)$$

where $\Delta_j = z_j - z_{j-1}$. To ensure that $\hat{p}(x|Z)$ integrates to 1.0 over the support region $[x_{min}, x_{max}]$ we have $K = \frac{1}{N_Z}$. Accordingly, our entropy estimate is given by:

$$\hat{H}_{\hat{p}}(X|Z) = \sum_{j=1}^{N_Z} \int_{z_{j-1}}^{z_j} -\ln(p_{j-1}^j) \cdot p_{j-1}^j \cdot dx \quad (6)$$

$$= \frac{1}{N_Z} \cdot \sum_{j=1}^{N_Z} \ln(N_Z \cdot \Delta_j) \quad (7)$$

$$= \ln(N_Z) + \frac{1}{N_Z} \cdot \sum_{j=1}^{N_Z} \ln(\Delta_j) \quad (8)$$

From Equation (8) we see that the estimate depends on the logs of the spacings between quantiles, and is defined by the average of these values. Further, we can define the error due to piecewise constant approximation of $p(x)$ as $\Delta H_{p, \hat{p}}(X|Z) = \hat{H}_{\hat{p}}(X|Z) - H_p(X)$.

3.3. Step 3—Estimation of the Quantiles of $p(x)$

The third step is to use the available data S to compute estimates of the quantiles Z to be plugged into Equation (6). Of course, given a finite sample size N_S , the number of quantiles N_Z that can be estimated will, in general, be much smaller than the sample size N_S (i.e., $N_Z \ll N_S$).

Various methods for computing estimates of the quantiles are available. Here, we use a relatively simple approach in which N_K sample subsets S_k , $k = 1, \dots, N_K$, each of size $N_Z - 1$ (i.e., $S_k = \{s_1^k, s_2^k, \dots, s_{N_Z-1}^k\}$) are drawn from the available sample set S , where the samples in each subset are drawn from S without replacement so that the values obtained in each subset are unique (not-repeated). For each subset, we sort the values in increasing order to obtain $Z_k = \{z_1^k, z_2^k, \dots, z_{N_Z-1}^k\}$, thereby obtaining N_K estimates $\{z_j^1, z_j^2, \dots, z_j^{N_K}\}$ for each z_j , $j = 1, \dots, N_Z - 1$. This procedure results in an empirical estimate of the sample distribution $p(z_j|S)$ for each quantile z_j , $j = 1, \dots, N_Z - 1$. Finally, we compute $\hat{z}_j = \frac{1}{N_K} \sum_{k=1}^{N_K} z_j^k$, and set $\hat{Z} = \{x_{min}, \hat{z}_1, \hat{z}_2, \dots, \hat{z}_{N_Z-1}, x_{max}\}$. Plugging these values into Equations (7) and (8), we get:

$$\hat{H}_p(X|\hat{Z}) = \frac{1}{N_Z} \cdot \sum_{j=1}^{N_Z} \ln(N_Z \cdot \hat{\Delta}_j) \tag{9}$$

$$= \ln(N_Z) + \frac{1}{N_Z} \cdot \sum_{j=1}^{N_Z} \ln(\hat{\Delta}_j) \tag{10}$$

where $\hat{\Delta}_j = \hat{z}_j - \hat{z}_{j-1}$. For practical computation, to avoid numerical problems as N_Z becomes large so that $\hat{\Delta}_j$ becomes very small and $\ln(\hat{\Delta}_j)$ approaches $-\infty$, we will actually use Equation (9). Further, we define the additional error due purely to imperfect quantile estimation to be $\Delta H_p(X|\hat{Z}, Z) = \hat{H}_p(X|\hat{Z}) - H_p(X|Z)$.

3.4. Random Variability Associated with the QS-Based Entropy Estimate

Given that the quantile spacing estimates $\{\hat{\Delta}_j, j = 1, \dots, N_Z\}$ are subject to random sampling variability associated with (i) the sampling of S from $p(x)$, and (ii) estimation of the quantile positions \hat{z}_j , the entropy estimate $\hat{H}_p(X|\hat{Z})$ will also be subject to random sampling variability. As shown later, we can generate probabilistic estimates of the nature and size of this error from the empirical estimates of $p(z_j|S)$ obtained in Step 3, and by bootstrapping on S .

4. Properties of the Proposed Approach

The accuracy of the estimate $\hat{H}_p(X|\hat{Z})$ obtained using the QS method outlined above depends on the following four assumptions, each of which we discuss below:

- (i) **A1:** The piecewise constant approximation $\hat{p}(x|Z)$ of $p(x)$ on the intervals between the quantile positions is adequate
- (ii) **A2:** The quantile positions $Z = \{z_0, z_1, z_2, \dots, z_{N_Z}\}$ of $p(x)$ have been estimated accurately.
- (iii) **A3:** The pdf $p(x)$ exists only on the support interval $[x_{min}, x_{max}]$, which has been properly chosen
- (iv) **A4:** The sample set S is consistent, representative and sufficiently informative about the underlying nature of $p(x)$

4.1. Implications of the Piecewise Constant Assumption

Assume that $\hat{p}(x|Z)$ provides a piecewise-constant estimate of $p(x)$ and that the quantile positions $Z = \{z_0, z_1, z_2, \dots, z_{N_Z}\}$ associated with a given choice for N_Z are perfectly known. Since the continuous shape of the cumulative distribution function (cdf)

$P(x)$ can be approximated to an arbitrary degree of accuracy by a sufficient number of piecewise linear segments, we will have $\hat{P}(x|Z) \rightarrow P(x)$ as $N_Z \rightarrow \infty$.

However, an insufficiently accurate approximation will result in a pdf estimate that is not sufficiently smooth, so that the entropy estimate will be biased. This bias will, in general, be positive (overestimation) because the piecewise-constant form $\hat{p}(x|Z)$ used to approximate $p(x)$ will always be shifted slightly in the direction of larger entropy; i.e., each piecewise constant segment in $\hat{p}(x|Z)$ is a maximum-entropy (uniform distribution) approximation of the corresponding segment of $p(x)$. However, the bias can be reduced and made arbitrarily small by increasing N_Z until the *Kullback-Leibler* divergence between $\hat{p}(x|Z)$ and $p(x)$ is so small that the information loss associated with use of $\hat{p}(x|Z)$ in place of $p(x)$ in Equation (1) is negligible.

The left panel of Figure 1 shows how this bias in the estimate of $H_p(X)$, due solely to piecewise constant approximation of the pdf (no sample data are involved), declines with increasing N_Z for three pdf forms of varying functional complexity (*Gaussian*, *Log-Normal* and *Exponential*), each using a parameter choice such that its theoretical entropy $H_p(X) = 1$. Also shown, for completeness, are results for the *Uniform* pdf where only one piecewise constant bin is theoretically required. Note that because $H_p(k \cdot (X - \mu_x)) = H_p(X) + \ln k$, the entropy can be changed to any desired value simply by rescaling on X . For these theoretical examples, the quantile positions are known exactly, and the resulting estimation bias is due *only* to the piecewise constant approximation of $p(x)$. However, since the theoretical pdfs used for this example all have infinite support, whereas the piecewise approximation requires specification of a finite support interval, for the latter we set $[x_{min}, x_{max}]$ to be the theoretical locations where $P(z_0) = \epsilon$ and $P(z_{N_Z}) = 1 - \epsilon$ respectively, with ϵ chosen to be some sufficiently small number (we used $\epsilon = 10^{-5}$). We see empirically that bias due to the piecewise constant approximation declines to zero approximately as an exponential function of $\log N_Z$ so that the absolute percent bias is less than $\sim 10\%$ for $N_Z > 10$ –30, less than $\sim 5\%$ for $N_Z > 50$, and less than $\sim 1\%$ for $N_Z > 200$.

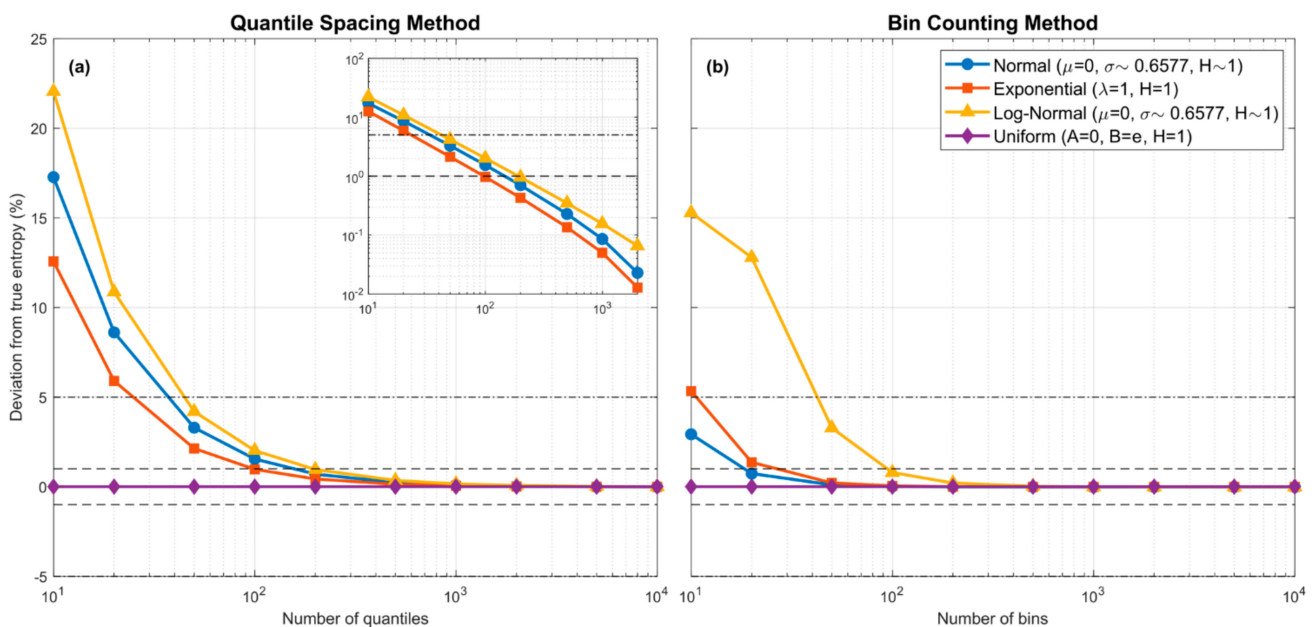


Figure 1. Plots showing how entropy estimation bias associated with the piecewise-constant approximation of various theoretical pdf forms varies with the number of quantiles ((a); QS method) or number of equal-width bins ((b); BC method) used in the approximation. The dashed horizontal lines indicate $\pm 1\%$ and $\pm 5\%$ bias error. No sampling is involved and the bias is due purely to the piecewise constant assumption. For QS, the locations of the quantiles are set to their theoretical values. To address the “infinite support” issue, $[x_{min}, x_{max}]$ were set to be the locations where $P(z_0) = \epsilon$ and $P(z_{N_Z}) = 1 - \epsilon$ respectively, with $\epsilon = 10^{-5}$. In both cases, bias approaches zero as the number of piecewise-constant units is increased. For the QS method, the decline in bias is approximately linear in the log-log space (see inset in the left subplot).

In practice, given a finite sample size N_S , our ability to increase the value of N_Z will be constrained by the size of the sample (i.e., $N_Z < N_S$). This is because when the form of $p(x)$ is unknown, the locations of the quantile positions must be estimated using the information provided by S . Further, what constitutes a sufficiently large value for N_Z will depend the complexity of the underlying shape of $p(x)$.

4.2. Implications of Imperfect Quantile Position Estimation

Assume that N_Z is large enough for the piecewise constant pdf approximation to be sufficiently accurate, but that the estimates $\{\hat{z}_0, \hat{z}_2, \dots, \hat{z}_{N_Z}\}$ of the locations of the quantiles are imperfect. Clearly, this can affect the estimate of entropy computed via Equation (9) by distorting the shape of $\hat{p}(x|\hat{Z})$ away from $\hat{p}(x|Z)$, and therefore away from $p(x)$. Further, the uncertainty associated with the quantile estimates will translate into uncertainty associated with the estimate of entropy.

In general, as the number of quantiles N_Z is increased, the inter-sample spacings associated with each ordered subset $Z_k = \{z_1^k, z_2^k, \dots, z_{N_Z-1}^k\}$, $k = 1, \dots, N_K$ will decrease, so that the distribution of possible locations for each quantile $z_j^k, j = 1, \dots, N_Z - 1$ will progressively become more tightly constrained. This means that the bias associated with each estimated quantile \hat{z}_j will reduce progressively towards zero as N_Z is increased (constrained only by sample size N_S) and the variance of the estimate \hat{z}_j will decline towards zero as the number of subsamples N_K is increased.

Figure 2 illustrates how bias and uncertainty associated with estimates of the quantiles diminish with increasing N_Z and N_K . Experimental results are shown for the *Log-Normal* density with $\mu = 0$ and $\sigma^2 = 0.6577$ (theoretical entropy $H_p(X) = 1$), with the y -axis indicating percent error in the quantile estimates corresponding to the 90% (green), 95% (purple) and 99% (turquoise) non-exceedance probabilities. In these plots, there is no distorting effect of sample size N_S (the sample size is effectively infinite), since when computing the estimates of the quantiles (as explained in Section 3.3) we draw subsamples of size N_Z directly from the theoretical pdf.

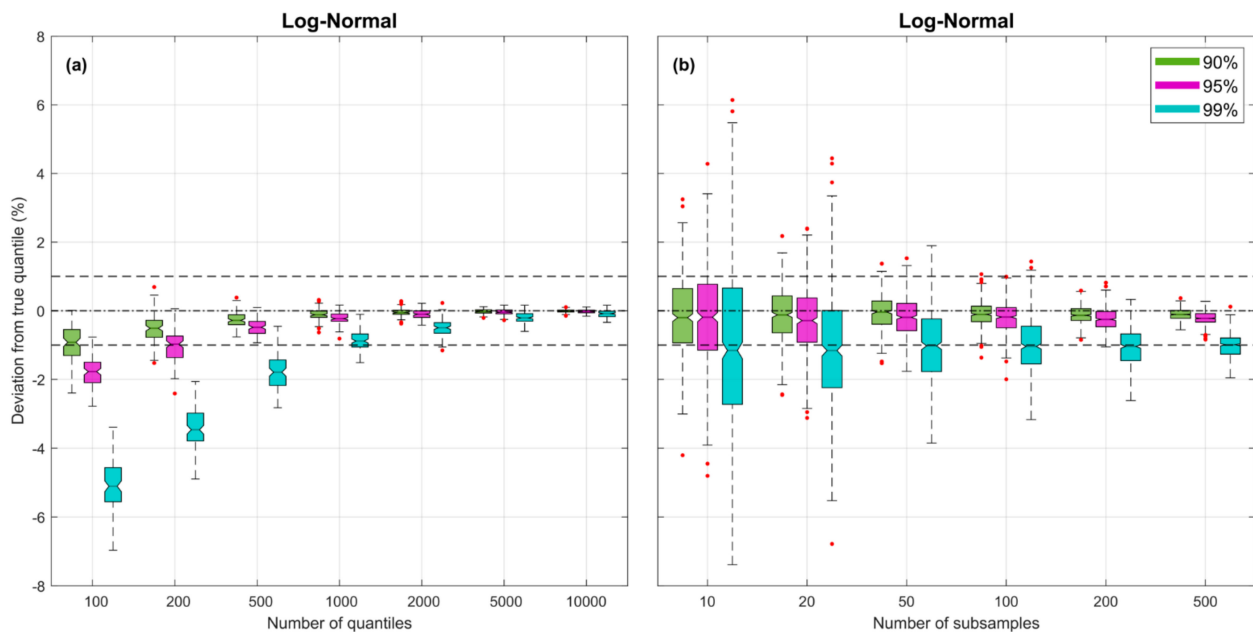


Figure 2. Plots showing bias and uncertainty associated with estimates of the quantiles derived from random samples, for the *Log-Normal* pdf. Uncertainty associated with random sampling variability is estimated by repeating each experiment 500 times. In both subplots, for each case, the box plots are shown side by side to improve legibility. (a) Subplot showing results varying $N_Z = [100, 200, 500, 1000, 2000, 5000, 10000]$ for fixed $N_K = 500$. (b) Subplot showing results varying $N_K = [10, 20, 50, 100, 200, 500]$ for fixed $N_Z = 1000$.

The left-side plot shows, for $N_K = 500$ subsamples, how the biases and uncertainties diminish as N_Z is increased. The boxplots reflect uncertainty due to random sampling variability, estimated by repeating each experiment 500 times (by drawing new samples from the pdf). As expected, for smaller N_Z the quantile location estimates tend to be negatively biased, particularly for those in the more extreme tail locations of the distribution. However, for $N_Z = 150$ the bias associated with the 99% non-exceedance probability quantile is less than -5% , for $N_Z \approx 500$ the corresponding bias is less than -2% , and for $N_Z = 1500$ it is less than -1% . The right-side plot shows, for a fixed value of $N_Z = 1000$, how the uncertainties diminish but the biases remain relatively constant as the number of subsamples N_K is increased. Overall, the uncertainty becomes quite small for $N_K > 200$.

4.3. Implications of the Finite Support Assumption

Assume that N_Z has been chosen large enough for the piecewise constant pdf approximation to be sufficiently accurate, and that the exact quantile positions associated with this choice for N_Z are known. The equation for estimating entropy (Equation (9)) can be decomposed into three terms:

$$H_{\hat{p}}(X|Z) = \frac{1}{N_Z} \cdot \sum_{j=1}^{N_Z} \ln(N_Z \cdot \Delta_j) = H_{x_{min}}^{z_1} + H_{z_1}^{z_{N_Z-1}} + H_{N_Z-1}^{x_{max}} \tag{11}$$

where $H_{x_{min}}^{z_1} = \frac{\ln(N_Z \cdot \Delta_1)}{N_Z}$, $H_{z_1}^{z_{N_Z-1}} = \frac{1}{N_Z} \cdot \sum_{j=2}^{N_Z-1} \ln(N_Z \cdot \Delta_j)$ and $H_{N_Z-1}^{x_{max}} = \frac{\ln(N_Z \cdot \Delta_{N_Z})}{N_Z}$, and

where Δ_j indicates the true inter-quantile spacings. Only the first and last terms $H_{x_{min}}^{z_1}$ and $H_{N_Z-1}^{x_{max}}$ are affected by the choices for x_{min} and x_{max} through $\Delta_1 = z_1 - x_{min}$ and $\Delta_{N_Z} = x_{max} - z_{N_Z-1}$.

Clearly, if $p(x)$ is bounded both above and below by specific known values, then there is no issue. However, if the support of $p(x)$ is not known, or if one or both bounds can reasonably be expected to extend to $\pm\infty$ (as appropriate), then the choice for the relevant limiting value (x_{min} or x_{max}) can significantly affect the computed value for $\hat{H}_{\hat{p}}$. To see this, note that the first term $H_{x_{min}}^{z_1}$ can be made to vary from $-\infty$ when $\Delta_1 = 0$, to $+\infty$ when $\Delta_1 = \infty$, passing through zero when $\Delta_1 = \frac{1}{N_Z}$; and similarly for the last term $H_{N_Z-1}^{x_{max}}$. Therefore, the error associated with $\hat{H}_{\hat{p}}$ can be made arbitrarily negatively large by choosing Δ_1 and Δ_{N_Z} to be too small, or arbitrarily positively large by choosing Δ_1 and Δ_{N_Z} to be too large.

In practice, when dealing with samples S from some unknown data generating process, we will often have only the samples themselves from which to infer the support of $p(x)$, and therefore can only confidently state that $x_{min} \leq \min\{S\}$ and $x_{max} \geq \max\{S\}$. One possibility could be to ignore the fractional contributions of the terms $H_{x_{min}}^{z_1}$ and $H_{N_Z-1}^{x_{max}}$ corresponding to the (unknown) portions of the pdf and instead use as our estimate $H_{\hat{p}}^*(X|Z) = H_{z_1}^{z_{N_Z-1}}$. This would be equivalent to setting $\Delta_1 = \Delta_{N_Z} = \frac{1}{N_Z}$, so that $X_{min} = z_1 - \frac{1}{N_Z}$ and $X_{max} = z_{N_Z-1} + \frac{1}{N_Z}$. By doing so, we would be ignoring a portion of the overall entropy associated with the pdf and can therefore expect to obtain an underestimate. However, this bias error $BE = H_{\hat{p}}(X|Z) - H_{\hat{p}}^*(X|Z)$ will tend to zero as N_Z is increased.

An alternative approach, that we recommend in this paper, is to set $x_{min} = \min\{S\}$ and $x_{max} = \max\{S\}$. In this case, there will be random variability associated with the sampled values for $\min\{S\}$ and $\max\{S\}$ and so the bias in our estimate $H_{\hat{p}}^*(X|Z)$ can be either negative or positive. Nonetheless, this bias error BE will still tend to zero as N_Z is increased.

Note that the percentage contributions of the entropy fractions $H_{x_{min}}^{z_1}$ and $H_{N_Z-1}^{x_{max}}$ to the total entropy $H_p(X|Z)$ will depend on the nature of the underlying pdf. Figure 3 illustrates this for three pdfs (*Gaussian* which has infinite extent on both sides, and the *Exponential* and *Log-Normal* which have infinite extent on only one side), assuming no estimation error associated with the quantile locations. For the *Gaussian* (blue) and *Exponential* (red)

densities, the largest fractional entropy contributions are clearly from the tail regions, whereas for the *Log-Normal* (orange) density this is not so. So, the entropy fractions can be proportionally quite large or small at the extremes, depending on the form of the pdf. Nonetheless, the overall entropy fraction associated with each quantile spacing diminishes with increasing N_Z . For the examples shown, when $N_Z = 100$ (left plot) the maximum contributions associated with a quantile spacing are less than 6% and when $N_Z = 1000$ (right plot) become less than 1%. This plot illustrates clearly the most important issue that must be dealt with when estimating entropy from samples.

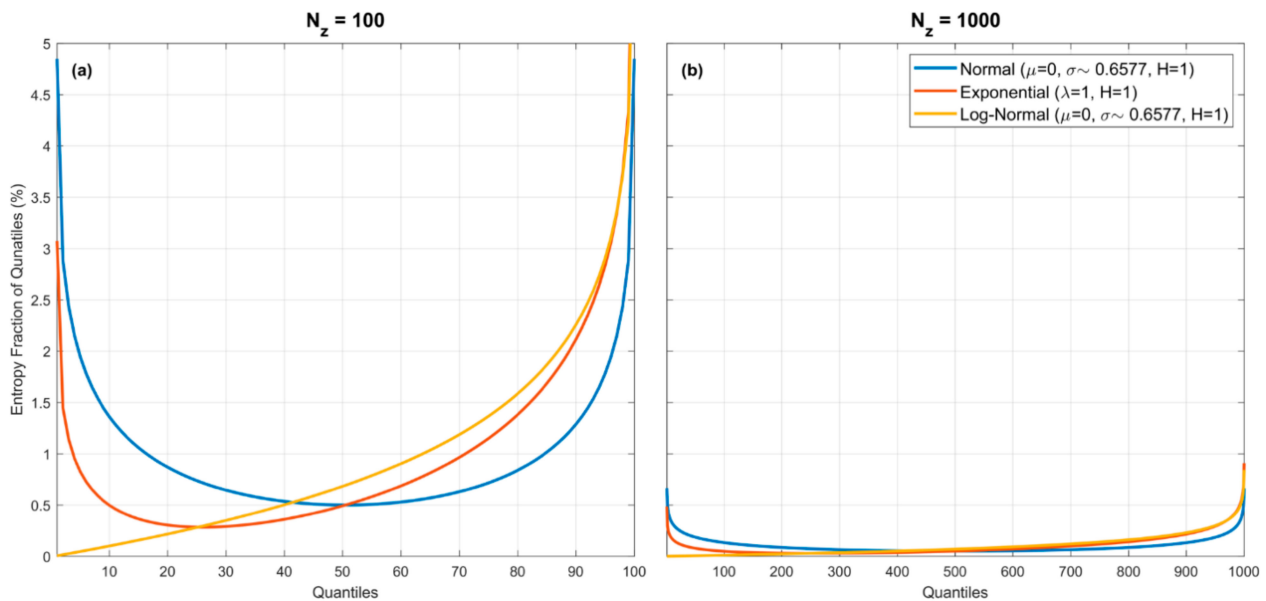


Figure 3. Plots showing percentage entropy fraction associated with each quantile spacing for the *Gaussian*, *Exponential* and *Log-Normal* pdfs, for $N_Z = 100$ (a), and $N_Z = 1000$ (b). For the *Uniform* pdf (not shown to avoid complicating the figures) the percentage entropy fraction associated with each quantile spacing is a horizontal line (at 1% in the left panel, and at 0.1% in the right panel). Note that the entropy fractions can be proportionally quite large or small at the extremes, depending on the form of the pdf. However, the overall entropy fraction associated with each quantile spacing diminishes with increasing N_Z . For the examples shown, the maximum contributions associated with a quantile spacing are less than 6% for $N_Z = 100$ (a), and become less than 1% for $N_Z = 1000$ (b).

So, on the one hand, the cumulative entropy fractions associated with the tail regions of $p(x)$ that lie beyond $\min\{S\}$ and $\max\{S\}$ are impossible to know. On the other, the individual contributions of these fractions associated with the extreme quantile spacings Δ_1 and/or Δ_{N_Z} where $p(x)$ is small can be quite a bit larger than those associated with the contributions from intermediate quantile spacings. Overall, the only real way to control the estimation bias and uncertainty associated with these extreme regions is to use a sufficiently large value for N_Z so that the relative contribution of the extreme regions is small. This will in turn, of course, be constrained by the sample size.

4.4. Combined Effect of the Piecewise Constant Assumption, Finite Support Assumption, and Quantile Position Estimation Using Finite Sample Sizes

In Section 4.1, we saw that the effect of the piecewise constant assumption on the QS-based estimate of entropy is *positive* bias that diminishes with increasing N_Z . Similarly, Section 4.2 showed that the biases associated with the quantiles diminish with increasing N_Z , while the corresponding uncertainties diminish with increasing N_K . As mentioned earlier, the bias in each quantile position will be towards the direction of locally higher probability mass (since more of the equiprobable random samples will tend to drawn from this region), and therefore the estimate $\hat{p}(x|\hat{Z})$ of $p(x)$ will be distorted in the direction of having smaller “dispersion” (i.e., $\hat{p}(x|\hat{Z})$ will tend to be more ‘peaked’ than $p(X)$), resulting

in *negative* bias in the corresponding estimate of entropy. Finally, Section 4.3 discussed the implications of the finite support assumption, given that x_{min} and x_{max} will often not be known.

Figure 4 illustrates the combined effect of these assumptions. Here we show how the overall percentage error in the QS-based estimate of entropy varies as a function of $\alpha = (N_Z/N_S)$, where α expresses the number of quantiles N_Z as a fraction of the sample size N_S . Sample sets of given size N_S are drawn from the *Gaussian* (left panel), *Exponential* (middle panel) and *Log-Normal* (right panel) densities, the quantiles are estimated using the procedure discussed in Section 3.3, x_{min} and x_{max} are set to be the smallest and largest data values in the set (Section 4.3), and entropy is estimated using Equation (9) for different selected values of N_Z . To account for sampling variability, the results are averaged over 500 different sample sets drawn randomly from the parent density.

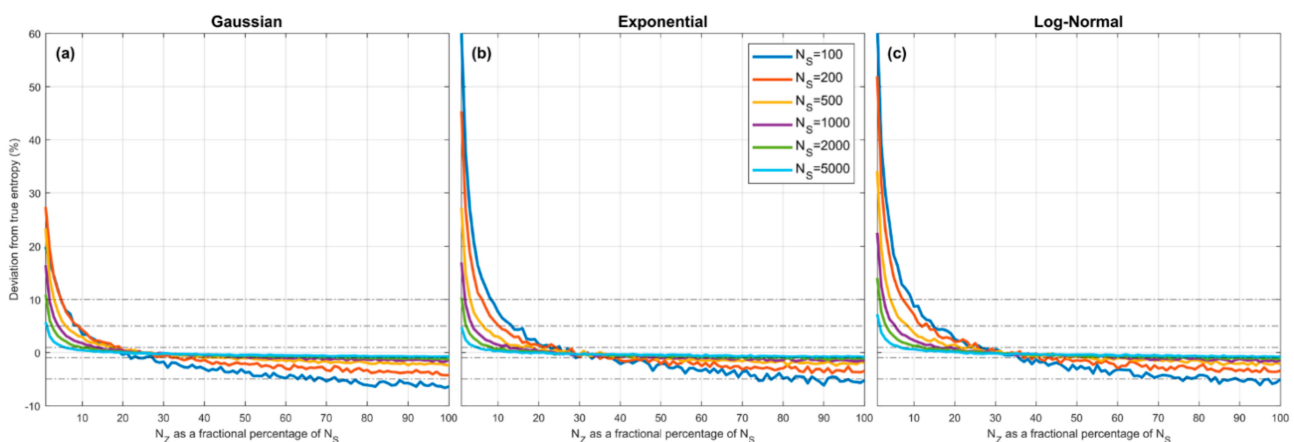


Figure 4. Plots showing expected percent error in the QS-based estimate of entropy derived from random samples, as a function of $\alpha = 100 * (N_Z/N_S)$, which expresses the number of quantiles N_Z as a fractional percentage of the sample size N_S . Results are averaged over 500 trials obtained by drawing sample sets of size N_S from the theoretical pdf, where x_{min} and x_{max} are set to be the smallest and largest data values in the particular sample. Results are shown for different sample sizes $N_S = [100, 200, 500, 1000, 2000, 5000]$, for the *Gaussian* (a), *Exponential* (b) and *Log-Normal* (c) densities. In each case, when α is small the estimation bias is positive (overestimation) and can be greater than 10% for $\alpha < 10\%$, and crosses zero to become negative (underestimation) when $\alpha > 25\text{--}35\%$. The marginal cost of setting α too large is low compared to setting α too small. As N_S increases, the bias diminishes. The optimal choice is $\alpha \approx 25\text{--}30\%$ and is relatively insensitive to pdf shape or sample size.

The plots show how percentage estimation error (bias) varies as α (and hence N_Z) changes as a fraction of sample size N_S , for different sample sizes from 100 to 5000. As might be expected, in each case when α is too small the estimation bias is positive (overestimation) and can be quite large due to the piecewise constant approximation. However, as α is increased the estimation bias decreases rapidly, crosses zero, and becomes negative (under-estimation) due to the combined effects of quantile position estimation bias and use of the smallest and largest sample values to approximate x_{min} and x_{max} . Most interesting is the fact that all of the curves cross zero at approximately $\alpha \approx 0.25\text{--}0.35$, and that this location does not seem to depend strongly on the sample size or shape of the pdf. Further, the marginal cost of setting α too large is low (less than -5% for $\alpha = 0.5$) compared to setting α too small. Overall, the expected bias error diminishes with increasing sample size N_S and the optimal choice for $\alpha \approx 0.25\text{--}0.35$.

Figure 5 illustrates both the bias and uncertainty in the estimate of entropy as a function of sample size N_S when we specified the number of quantiles N_Z to be 25% of the sample size (i.e., $\alpha = 0.25$). The uncertainty intervals are due to sampling variability, estimated by drawing 500 different sample sets from the parent population. The results

show that uncertainty due to sampling variability diminishes rapidly with sample size, becoming relatively small for large sample sizes.

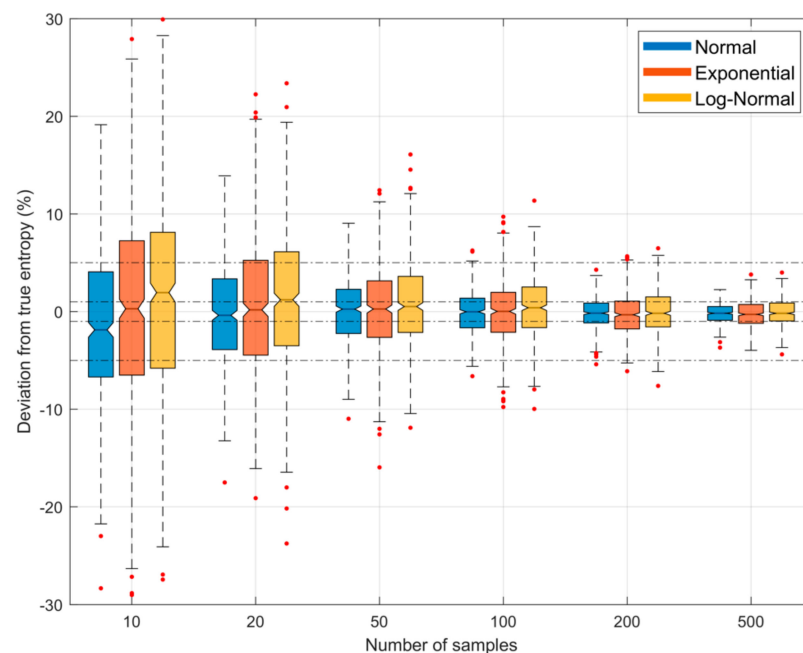


Figure 5. Plots showing bias and uncertainty in the QS-based estimate of entropy derived from random samples, as a function of sample size N_S , when the number of quantiles N_Z is set to 25% of the sample size ($\alpha = 0.25$), and x_{min} and x_{max} are respectively set to be the smallest and largest data values in the particular sample. The uncertainty shown is due to random sampling variability, estimated by drawing 500 different samples from the parent density. Results are shown for the *Gaussian* (blue), *Exponential* (red) and *Log-Normal* (orange) densities; box plots are shown side by side to improve legibility. As sample size N_S increases, the uncertainty diminishes.

4.5. Implications of Informativeness of the Data Sample

For the results shown in Figures 4 and 5, we drew samples directly from $p(x)$. In practice, we must construct our entropy estimate by using a single data sample S of finite size N_S . Provided that S is a consistent and representative random sample from $p(x)$, with each element x_i being iid, then a sufficiently large sample size N_S should enable construction of an accurate approximation $\hat{p}(x|\hat{Z})$ of $p(x)$ via the QS method. However, if N_S is too small, it can (i) prevent setting a sufficiently large value for N_Z , and (ii) tend to make the sets Z_k sub-sampled from S to be insufficiently independent for accurate estimates of the quantile positions of $p(x)$ to be obtained. The overall effect will be to prevent $\hat{p}(x|\hat{Z})$ from approaching $p(x)$, leading to an unreliable estimate of its entropy.

Further, even if N_S is sufficiently large for $\hat{p}(x|\hat{Z}) \rightarrow p(x)$, sampling variability associated with randomly drawing S from $p(x)$ will result in the entropy estimate $H_{\hat{p}}(X|\hat{Z})$ being subject to statistical variability. Figure 6 shows how bootstrapped estimates of the uncertainty will differ from those shown in Figure 5 above, in which we drew different sample sets from the parent population. Here, each time a sample set is drawn from the parent density we draw $N_B = 500$ bootstrap samples of the same size N_S from that sample set, use these to obtain N_B different estimates of the associated entropy (using $\alpha = 0.25$), and compute the width of the resulting inter-quartile range (IQR). We then repeat this procedure for 500 different sample sets of the same size drawn from the parent population. Figure 6 shows the ratio of the IQR obtained using bootstrapping to that of the actual IQR for different sample sizes; the boxplots represent variability due to random sampling. Here, an expected (mean) ratio value of 1.0 and small width of the boxplot is ideal, indicating that bootstrapping provides a good estimate of the uncertainty to be associated with random sampling variability. The results show that for smaller sample sizes ($N_S < 500$) there is a

tendency to overestimate the width of the inter-quartile range, but that this slight positive bias disappears for larger sample sizes.

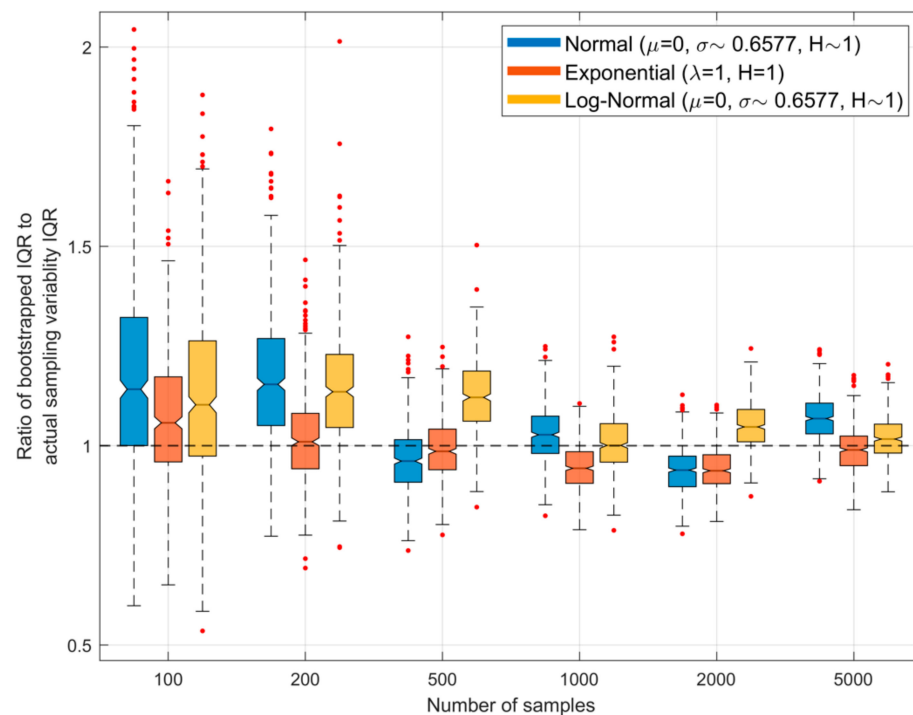


Figure 6. Plots showing, for different sample sizes and $\alpha = 25\%$, the ratio of the interquartile range (IQR) of the QS-based estimate of entropy obtained using bootstrapping to that of the actual IQR arising due to random sampling variability. Here, each sample set drawn from the parent density is bootstrapped to obtain $N_B = 500$ different estimates of the associated entropy, and the width of the resulting inter-quartile range is computed. The procedure is repeated for 500 different sample sets drawn from the parent population, and the graph shows the resulting variability as box-plots. The ideal result would be a ratio of 1.0.

4.6. Summary of Properties of the Proposed Quantile Spacing Approach

To summarize, bias in the estimate $\hat{H}_p(X|\hat{Z})$ can arise due to: (a) inadequacy of the piece-wise approximation of $p(x)$, (b) imperfect estimation of the quantile positions, (c) imperfect knowledge of the support interval, and (d) the sample S not being consistent, representative and sufficiently informative. Meanwhile, uncertainty in the estimate can arise due to: (a) random sampling variability associated with estimation of the quantiles, and (b) random sampling variability associated with drawing S from $p(x)$. For a given sample size N_S , and provided that the sample is consistent, representative and fully informative, the bias and uncertainty can be reduced by selecting sufficiently large values for N_Z and N_K (we recommend $N_K = 500$ and $N_Z = 0.25 \cdot N_S$), while the overall statistical variability associated with the estimate can be estimated by bootstrapping from S .

4.7. Algorithm for Estimating Entropy via the Quantile Spacing Approach

Given a sample set S of size N_S

- (1) Set $x_{min} = \min\{S\}$ and $x_{max} = \max\{S\}$
- (2) Select values for $\psi = \{N_Z, N_K, N_B\}$. Recommended default values are $N_Z = \alpha \cdot N_S$, $N_K = 500$ and $N_B = 500$, with $\alpha = 0.25$.
- (3) Bootstrap a sample set S_b of size N_S from S with replacement.
- (4) Compute the entropy estimate $\hat{H}_p(X|\hat{Z}_b)$ using Equation (9) and the procedure outlined in Section 3.

- (5) Repeat the above steps N_B times to generate the bootstrapped distribution of $\hat{H}_p(X|\hat{Z}_b)$ as an empirical probabilistic estimate $p(\hat{H}_p(X|S))$ of the Entropy $H_p(X)$ of $p(x)$ given S .

5. Relationship to the Bin Counting Approach

Because the proposed QS approach employs a piecewise constant approximation of $p(x)$, there are obvious similarities to BC. However, there are also clear differences. First, while BC typically employs equal-width binning along the support of X , with each bin having a different fraction of the total probability mass, QS uses variable width intervals (analogous to “bins”) each having an identical fraction of the total probability mass (so that the intervals are wider where $p(X)$ is small, and narrower where $p(X)$ is large). Both methods require specification of the support interval $[x_{min}, x_{max}]$.

Second, whereas BC requires counting samples falling within bins to estimate the probability masses associated with each bin, QS involves no “bin-counting”, and, instead, the data samples are used to estimate the positions of the quantiles. Since the probability mass estimates obtained by counting random samples falling within bins can be highly uncertain due to sampling variability, particularly for small sample sizes N_S , this translates into uncertainty regarding the shape of the pdf and thereby regarding its entropy. In QS, the effect of sampling variability is to consistently provide a pdf approximation that tends to be slightly more peaked than the true pdf, so that the bias in the entropy estimate tends to be slightly negative. This negative bias acts to counter the positive bias resulting from the piecewise constant approximation of the pdf.

Third, whereas BC requires selection of a bin-width hyper-parameter Δ that represents the appropriate bin-width required for “smoothing” to ensure an appropriate balance between bias and variance errors, QS requires selection of a hyper-parameter α that specifies the number of quantile positions N_Z as a fraction of the sample size N_S . As seen in Figure 4, an appropriate choice for α can effectively drive estimation bias to zero, while N_K controls the degree of uncertainty associated with the estimation of the positions of the quantiles. Further, if we desire estimates of the uncertainty in the computed value of entropy arising due to random sampling variability, we must specify the number of bootstraps N_B . In practice, the values selected for N_K and N_B can be made arbitrarily large, and our experiments suggest that setting $N_K = 500$ (or larger) and $N_B = 500$ (or larger) works well in practice. Accordingly, the QS hyper-parameter α takes the place of the BC hyper-parameter Δ in determining the accuracy of the Entropy estimate obtained from a given sample.

Our survey of the literature suggests that the problem of how to select the BC bin-width hyper-parameter Δ is not simple, and a number of different strategies have been proposed. [18] proposed to choose the number of bins based on sample size only. [8] estimated the optimal number of bins by minimizing the mean squared error between the sample histogram and the “true” form of the pdf (for which the shape must be assumed). [19] further developed this approach by estimating the shape of the true pdf from the interquartile range of the sample. More recently, [20] proposed a method that does not require choice of a hyper-parameter—using a Bayesian maximum likelihood approach, and assuming a piecewise-constant density model, the posterior probability for the number of bins is identified (this approach also provides uncertainty estimates for the related bin counts). Other BC approaches that provide uncertainty estimates based on the *Dirichlet*, *Multinomial*, and *Binomial* distributions are discussed by [17]. However, as shown in the next section, in practice the “optimal” fixed bin-width can vary significantly with shape of the pdf and with sample size.

6. Experimental Comparison with the Bin Counting Method

The right panel of Figure 1 shows the theoretical bias, due only to piecewise-constant approximation, associated with the estimate of $H_p(X)$ obtained using BC when the support interval $[x_{min}, x_{max}]$ is subdivided into equal-width intervals. We can compare the results

to the left panel of Figure 1 if we consider the number N_{Bin} of BC bins to be analogous to the number N_Z of spacings between quantiles for QS. Note that no random sampling variability or data informativeness issues are involved in the construction of these figures. For BC we use the theoretical fractions of probability mass associated with each of the equal-width bins, and for QS we use the theoretical quantile positions to compute the interval spacings. In both cases, to address the “infinite support” issue, we set $[x_{min}, x_{max}]$ to the locations where $P(z_0) = \epsilon$ and $P(z_{N_Z}) = 1 - \epsilon$ respectively, with $\epsilon = 1 \times 10^{-5}$. As with QS, the bias in entropy computed using the BC equal bin-width piecewise-continuous approximation declines to zero with increasing numbers of bins, and becomes less than 1% when $N_{Bin} \geq 100$; in fact, it can decline somewhat faster than for the QS approach. Clearly, for the *Gaussian* (blue) and *Exponential* (red) densities, the BC constant bin-width approximation can provide better entropy estimates with fewer bins than the QS variable bin-width approach. However, for the skewed *Log-Normal* density (orange) the behavior of the BC approximation is more complicated, whereas the QS approach shows an exponential rate of improvement with increasing number of bins for all three density types. This suggests that the variable bin-width QS approximation may provide a more consistent approach for more complex distributional forms (see Section 7).

Further, Figure 7 shows the results of a “naïve” implementation of BC where the value for N_{Bin} is varied as a fractional percentage of sample size N_S . As with QS, we specify the support interval by setting $x_{min} = \min\{S\}$ and $x_{max} = \max\{S\}$, but here the support interval is divided into equal-width bins so that $X_{BIN} = \{X_0, X_1, X_2, \dots, X_{N_{Bin}}\}$ represents the locations of the edges of the bins (where $X_0 = x_{min}$ and $X_{N_{Bin}} = x_{max}$), and therefore $\Delta = \frac{x_{max} - x_{min}}{N_{Bin}}$. We then assume that $p(x) \approx \hat{p}(x|X_{BIN}) = \frac{n_j}{N_S}$ for $X_{j-1} \leq X < X_j$ where n_j is the number of samples falling in the bin defined by $X_{j-1} \leq X < X_j$. Finally, we compute the BC estimate of entropy as $\hat{H}_{\hat{p}}(X|X_{BIN}) = - \sum_{j=1}^{N_{Bin}} \ln\left(\frac{n_j}{N_S}\right) \cdot \frac{n_j}{N_S} + \ln(\Delta)$, and follow the convention that $0 \cdot \ln(0) = 0$ to handle bins where the number of samples $n_j = 0$. Finally, we obtain estimates for 500 different sample sets drawn from the parent density and average the results. Results are shown for different sample sizes $N_S = \{100, 200, 500, 1000, 2000$ and $5000\}$. The yellow marker symbols indicate where each curve crosses the zero-bias line; clearly N_{Bin} is not a constant fraction of N_S , and for any given sample size the ratio of $\frac{N_{Bin}}{N_S}$ changes with form of the pdf.

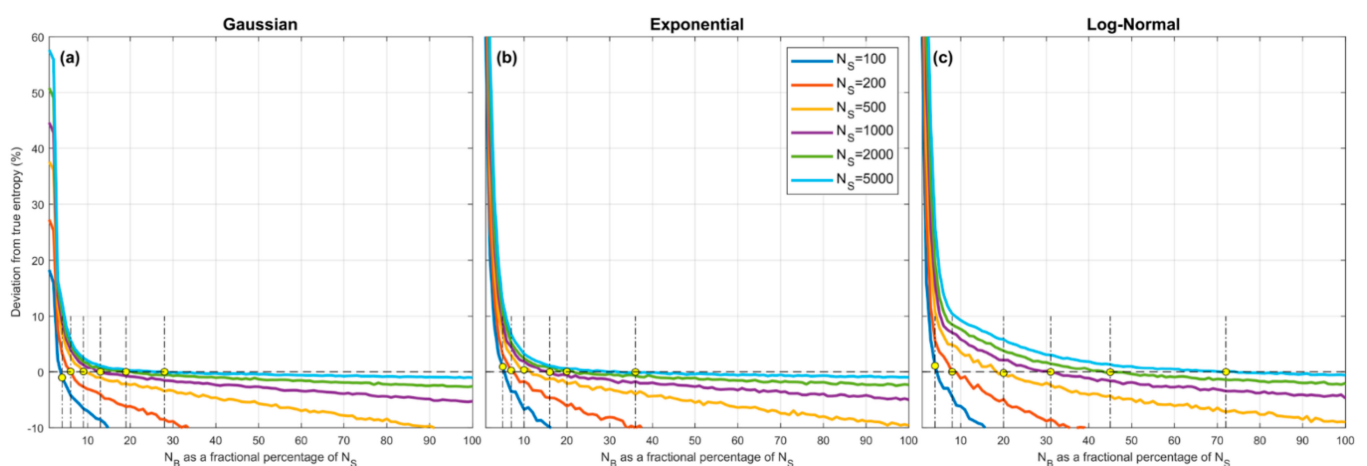


Figure 7. Plots showing how expected percentage error in the BC-based estimate of Entropy derived from random samples, varies as a function of the number of bins N_{Bin} for the (a) *Gaussian*, (b) *Exponential*, and (c) *Log-Normal* densities. Results are averaged over 500 trials obtained by drawing sample sets of size N_S from the theoretical pdf, where x_{min} and x_{max} are set to be the smallest and largest data values in the particular sample. Results are shown for different sample sizes $N_S = [100, 200, 500, 1000, 2000, 5000]$. When the number of bins is small the estimation bias is positive (overestimation) but rapidly declines to cross zero and become negative (underestimation) as the number of bins is increased. In general, the overall ranges of overestimation and underestimation bias are larger than for the QS method (see Figure 4).

To more clearly compare these BC results with the results shown in Figure 4 for the QS approach, Figure 8 shows a plot indicating the sampling variability distribution of the optimal number of bins (i.e., the value of N_{Bin} for which the expected entropy estimation error is zero) as a function of sample size for the *Gaussian*, *Exponential* and *Log-Normal* densities. We see clearly that, in contrast to QS, the “expected optimal” number of bins to achieve zero bias is neither a constant fraction of the sample size or independent of the pdf shape, but instead declines as the sample size increases, and is different for different pdf shapes. Further, the sampling variability associated with the optimal fractional number of bins can be quite large, and is highly skewed at smaller sample sizes. This is in contrast with QS where the optimal fractional number of bins is approximately constant at $\alpha \approx 25\text{--}35\%$ for different sample sizes and pdf shapes.

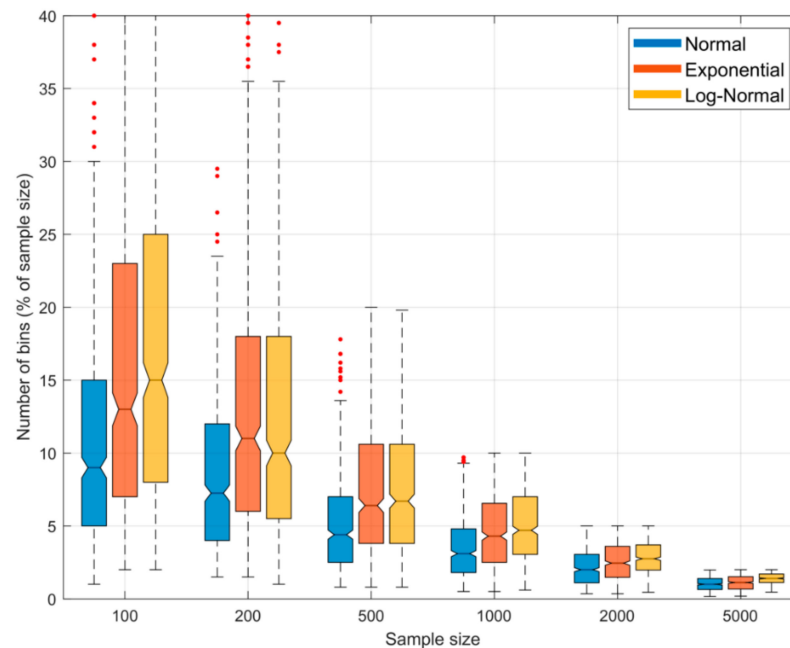


Figure 8. Boxplots showing the sampling variability distribution of optimal fractional number of bins (as a percentage of sample size) to achieve zero bias, when using the BC method for estimating entropy from random samples. Results are shown for the *Gaussian* (blue), *Exponential* (red) and *Log-Normal* (orange) densities. The uncertainty estimates are computed by drawing 500 different sample data sets of a given size from the parent distribution. Note that the expected optimal fractional number of bins varies with shape of the pdf, and is not constant but declines as the sample size increases. This is in contrast with the QS method where the optimal fractional number of bins is constant at $\sim 25\%$ for different sample sizes and pdf shapes. Further, the variability in optimal fractional number of bins can be large and highly skewed at smaller sample sizes.

7. Testing on Multi-Modal PDF Forms

While the types of pdf forms tested in this paper are far from exhaustive, they represent differently shapes and degrees of skewness, including infinite support on both sides (*Gaussian*), and infinite support on only one side (*Exponential* and *Log-Normal*). However, all three forms are “unimodal”, and so we conducted an additional test for a multimodal distributional form.

Figure 9 shows results for a *Bimodal* pdf (Figure 9a) constructed using a mixture of two *Gaussians* $N(1,5)$ and $N(5,1)$. Since its theoretical entropy value is unknown, we used the piecewise constant approximation method with true (known) quantile positions to compute its entropy by progressively increasing the number of quantiles N_Z until the estimate converged to within three decimal places (Figure 9b) to the value $H_p(X) \approx 2.265$ for $N_Z > 2000$. Figure 9c,d show that QS estimation bias declines exponentially with fractional number of quantiles and crosses zero at $\alpha \approx 25\%$ to 35% , in a manner similar to

the *Unimodal* pdfs tested previously (Figure 4). Figure 9c shows the results for $N_S = 5000$ samples, along with the distribution due to sampling variability (500 repetitions), showing that the IQR falls within $\pm 1\%$ of the correct value and the whiskers (± 2 sigma) fall within $\pm 3\%$. Figure 9d shows the expected bias (estimated by averaging over 500 repetitions) for varying sample size N_S ; for smaller sample sizes, the optimal value for α is closer to 20%, while for $N_S \geq 200$ the value of $\alpha \approx 25\%$ to 35% seems to work quite well, while being relatively insensitive to the choice of value within this range.

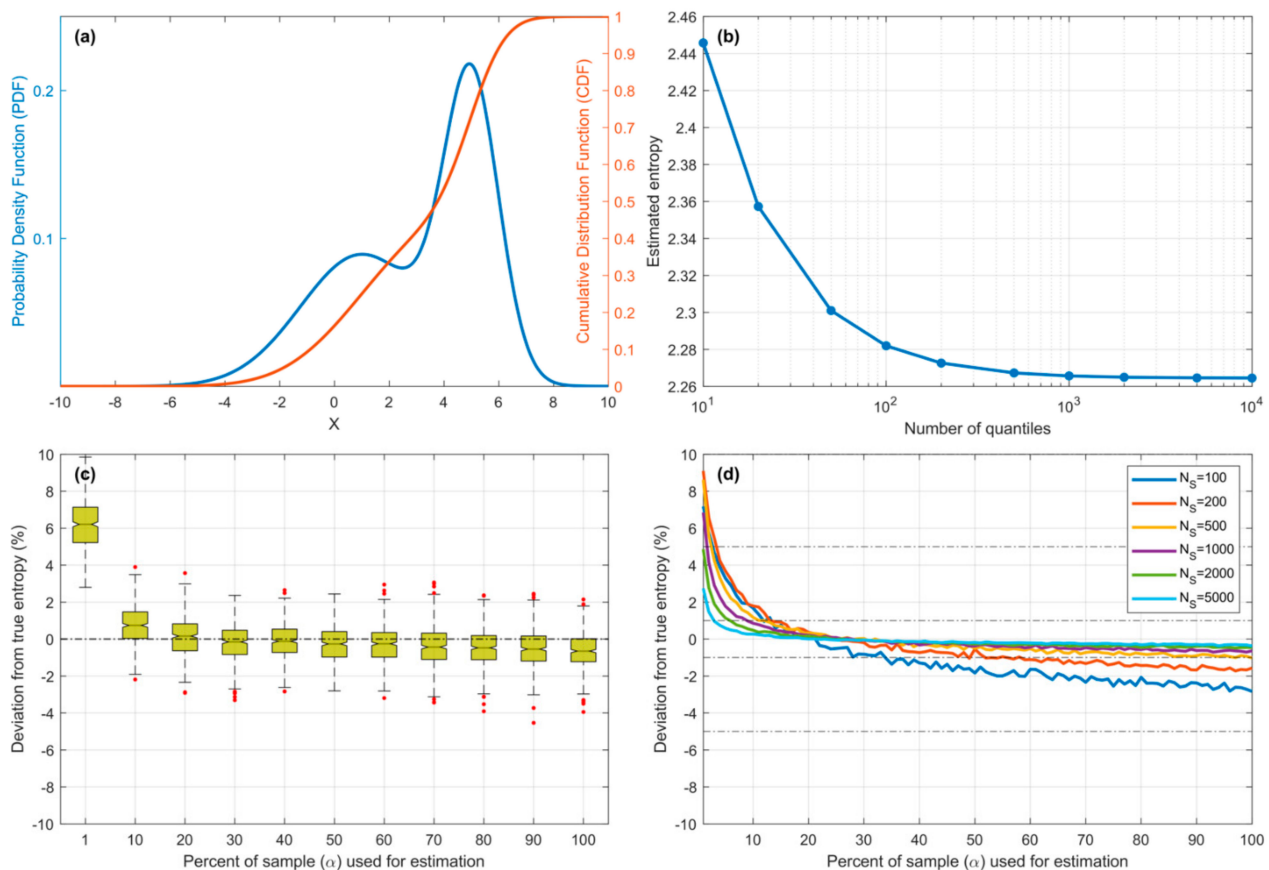


Figure 9. Plots showing results for the *Bimodal* pdf. (a) Pdf and Cdf for the *Gaussian Mixture* model. (b) Showing convergence of entropy computed using piecewise constant approximation as the number of quantiles N_Z is increased. (c) Bias and sampling variability of the QS-based estimate of entropy plotted against N_Z as a percentage of sample size. (d) Expected bias of QS-based estimate of entropy plotted against N_Z as a percentage of sample size, for different sample sizes $N_S = [100, 200, 500, 1000, 2000, 5000]$.

Interestingly, comparing Figure 9d (*Bimodal Gaussian Mixture*) with Figure 4a (*Unimodal Gaussian*), we see that QS actually converges more rapidly for the *Bimodal* density. One possible explanation is that the *Bimodal* density is in some sense “closer” in shape to a *Uniform* density, for which the piecewise constant representation is a better approximation.

8. Experimental Comparison with the Kernel Density Method

Figures 10 and 11 show performance of the KD method on the same four distributions (*Gaussian*, *Exponential*, *Log-Normal* and *Bimodal*). We used the KD method developed by [21] and the code provided at webee.technion.ac.il/~yoav/research/blind-separation.html (accessed on 13 May 2021), which provides an efficient implementation of the non-parametric Parzen-window density estimator [4,22]. To align well with the known (exponential-type) shapes of the four example distributions, we used a *Gaussian* kernel, so that the corresponding hyper-parameter to be tuned/selected is the standard deviation σ_K of each kernel. Consistent with our implementation of QS and BC, the code specifies the support interval

by setting $x_{min} = \min\{S\}$ and $x_{max} = \max\{S\}$. As discussed below, the results clearly reinforce the findings of [17] that KD performance can depend strongly on distribution shape and sample size.

According to Parzen-window theory, the appropriate choice for the kernel standard deviation σ_K is a function of the sample size N_S such that $\sigma_K = K/\sqrt{N_S}$, where K is some unknown constant; this makes sense because for smaller sample sizes the data points are spaced further apart and we require more smoothing (larger σ_K), while for larger sample sizes the data points are more closely spaced and we require less smoothing (smaller σ_K). Accordingly, Figure 10 shows the expected percentage estimation error (bias) on the y -axis plotted as a function of $K = \sigma_K \cdot \sqrt{N_S}$ for different sample sizes (averaged over 500 repetitions) for each of the four distributions. The yellow marker symbols indicate where each curve crosses the zero-bias line. Clearly, in practice, the optimal value for the KD hyper-parameter K is not constant and varies as a function of both sample size and distributional form.

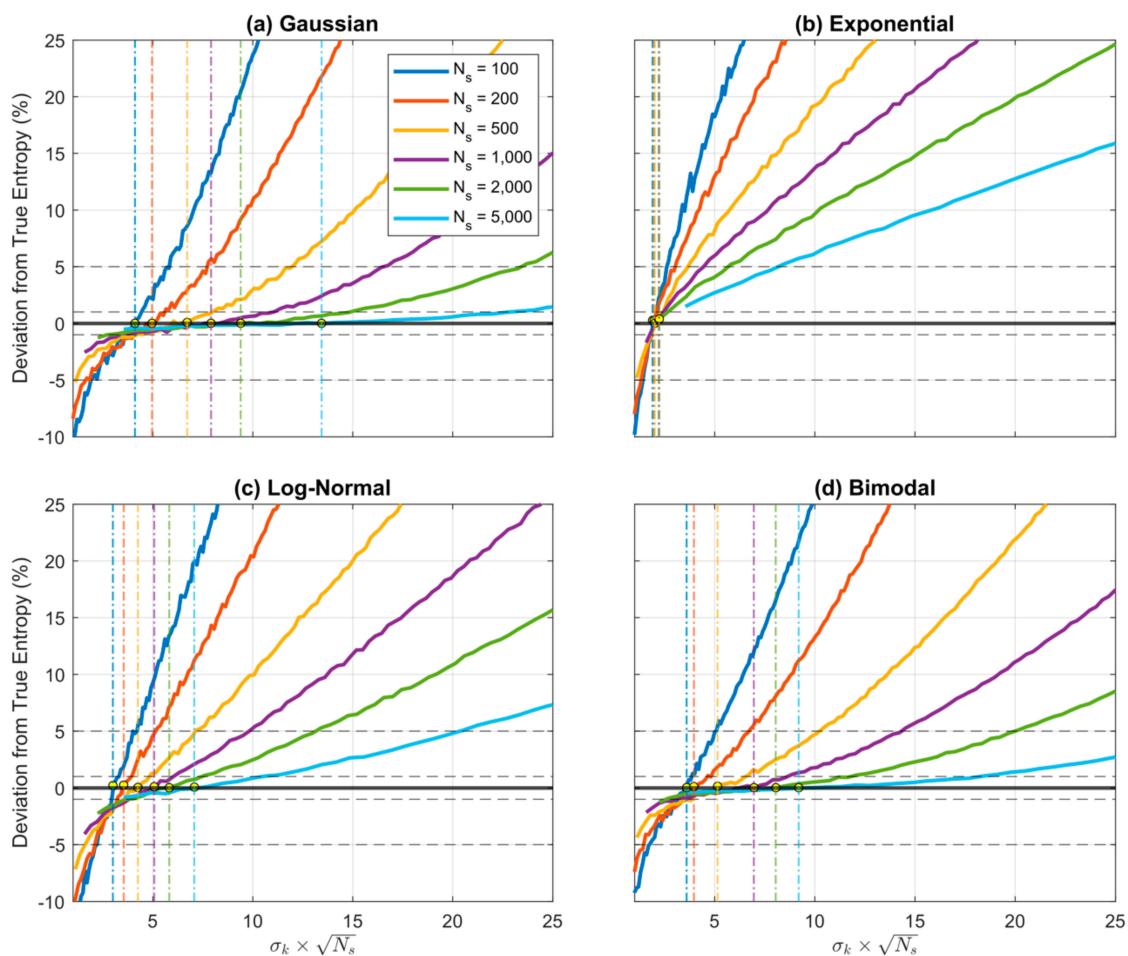


Figure 10. Plot showing how expected percentage error in the KD-based estimate of Entropy derived from random samples, varies as a function of $K = \sigma_k \cdot \sqrt{N_S}$ when using a Gaussian kernel. Results are averaged over 500 trials obtained by drawing sample sets of size N_S from the theoretical pdf, where x_{min} and x_{max} are set to be the smallest and largest data values in the particular sample. Results are shown for different sample sizes $N_S = [100, 200, 500, 1000, 2000, 5000]$, for the (a) Gaussian, (b), Exponential, (c) Log-Normal, and (d) Bimodal densities. When the kernel standard deviation σ_k (and hence K) is small the estimation bias is negative (underestimation) but rapidly increases to cross zero and become positive (overestimation) as the kernel standard deviation is increased. The location of the crossing point (corresponding to optimal value for K (and hence σ_k)) varies with sample size and shape of the pdf.

This variation is illustrated clearly by Figure 11 where the optimal K is plotted as a function of sample size N_S for each of the four distributions. So, as with the BC (but in contrast with the QS), practical implementation of KD requires both selection of an appropriate kernel type *and* tuning to determine the optimal value of the kernel hyper-parameter (either K or σ_K); this can be done by optimizing σ_K to maximize the *Likelihood* of the data. Note that for smaller sample sizes the entropy estimate can be very sensitive to the choice of this hyper-parameter, as evidenced by the steeper slopes of the curves in Figure 10. In contrast, with QS no kernel-type needs to be selected and no hyper-parameter tuning seems to be necessary, regardless of distribution shape and sample size, and sensitivity of the entropy estimate to precise choice of the number of quantiles N_Z is relatively small even for smaller sample sizes (see Figure 4).

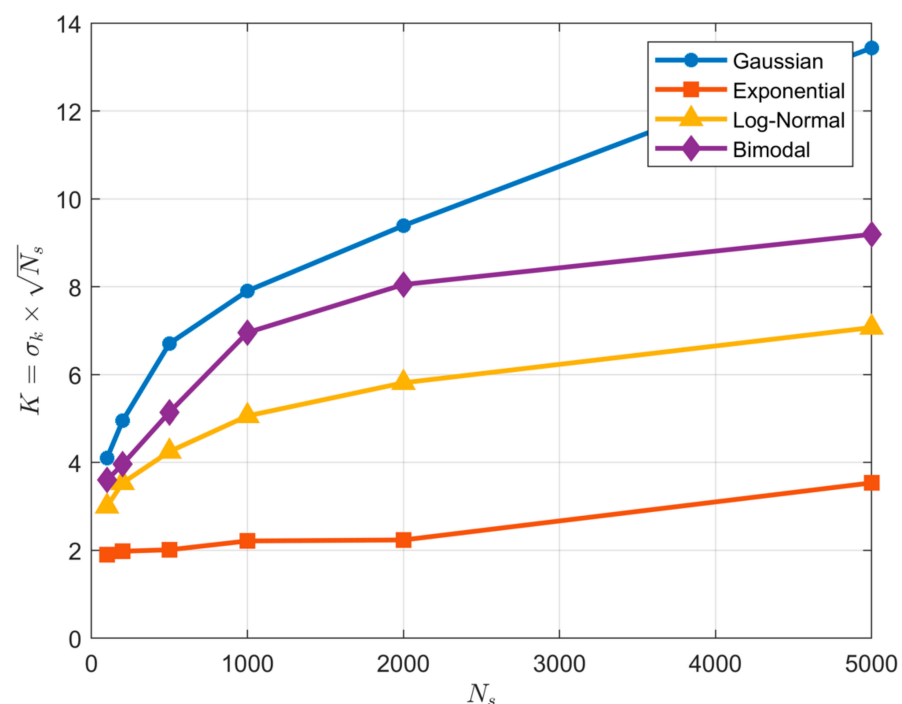


Figure 11. Plot showing how the optimal value of the KD hyper-parameter $K = \sigma_k \cdot \sqrt{N_S}$ varies as a function of sample size N_S and pdf type when using a *Gaussian* kernel. In disagreement with Parzen-window theory, the optimal value for K does not remain approximately constant as the sample size N_S is varied. Further, the value of K varies significantly with shape of the underlying pdf.

9. Discussion and Conclusions

The QS approach provides a relatively simple method for obtaining accurate estimates of entropy from data samples, along with an idea of the estimation-uncertainty associated with sampling variability. It appears to have an advantage over BC and KD since the most important hyper-parameter to be specified, the number of quantiles N_Z , does not need to be tuned and can apparently be set to a fixed fraction (~ 25 – 35%) of the sample size, regardless of pdf shape or sample size. In contrast, for BC the optimal number of bins N_{Bin} varies with pdf shape and sample size and, since the underlying pdf shape is usually not known beforehand, it can be difficult to come up with a general rule for how to accurately specify this value. Similarly, for KD, without prior knowledge of the underlying pdf shape (and especially when the pdf may be non-smooth) it can be difficult to know what kernel-type and hyper-parameter settings to use.

Besides being simpler to apply, the QS approach appears to provide a more accurate estimate of the underlying data generating pdf than either BC or KD, particularly for smaller sample sizes. This is illustrated clearly by Figure 12 where the expected percent entropy estimation error is plotted as a function of sample size N_S for each of the three

methods. For QS, the fractional number of bins was fixed at $\alpha = 25\%$ regardless of pdf form or sample size; in other words, no hyperparameter tuning was performed. For each of the other methods, the corresponding hyperparameter (kernel standard deviation σ_K for KD, and bin width Δ for BC) was optimized for each random sample, by finding the value that maximizes the *Likelihood* of the sample. As can clearly be seen, the QS-based estimates remain relatively unbiased even for samples as small as 100 data points, whereas the KD- and BC-based estimates tend to get progressively worse (negatively biased) as sample sizes are decreased. Overall, QS is both easier to apply (no hyper-parameter tuning required) and likely to be more accurate than BC or KD when applied to data from an unknown distributional form, particularly since the the piecewise linear interpolation between CDF points makes it applicable to pdfs of any arbitrary shape, including those with sharp discontinuities in slope and/or magnitude. A follow-up study investigating the accuracy of these methods when faced with data drawn from complex, arbitrarily shaped, pdfs is currently in progress and will be reported in due course.

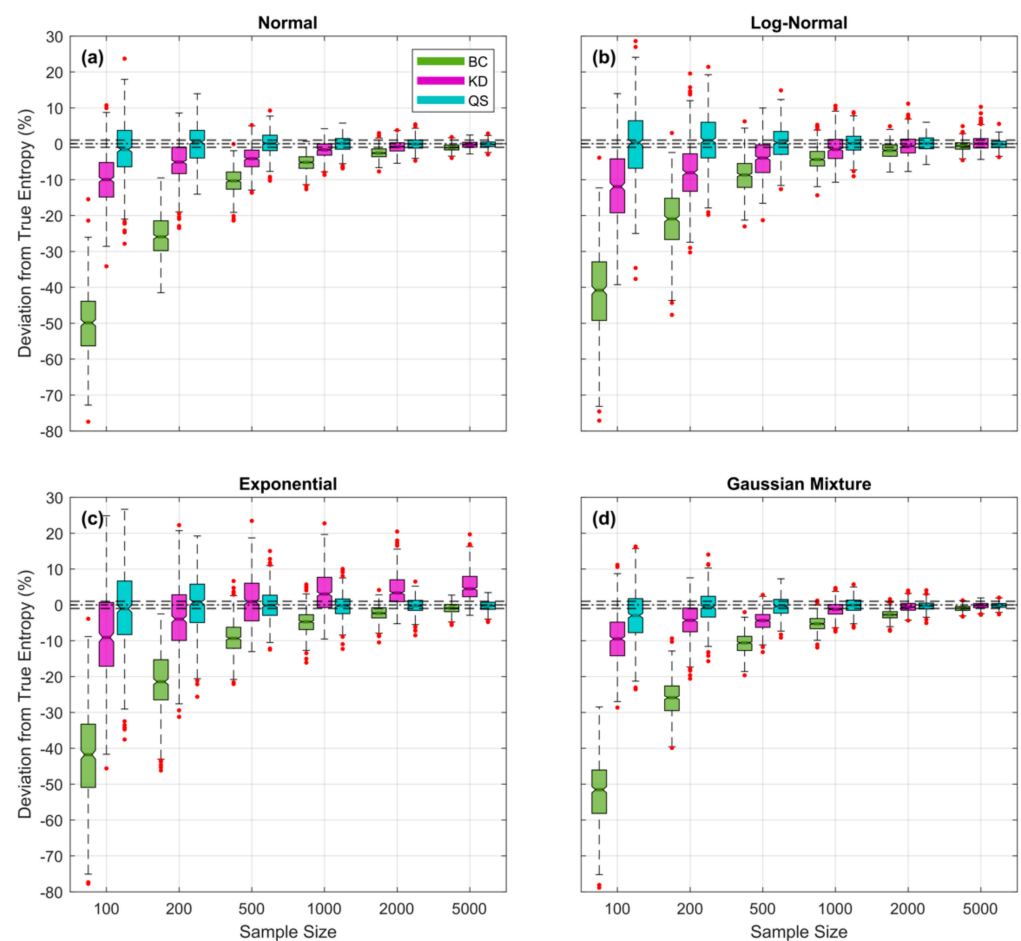


Figure 12. Plots showing expected percent error in the QS- (blue), KD- (purple) and BC-based (green) estimates of entropy derived from random samples, as a function of sample size N_S for the (a) *Gaussian*, (b) *Log-Normal*, (c) *Exponential*, and (d) *Bimodal* densities; box plots are shown side by side to improve legibility. Results are averaged over 500 trials obtained by drawing sample sets of size N_S from the theoretical pdf, where x_{min} and x_{max} are set to be the smallest and largest data values in the particular sample. For QS, the fractional number of bins was fixed at $\alpha = 25\%$ regardless of pdf form or sample size. For KD and BC, the corresponding hyperparameter (kernel standard deviation σ_K and bin width Δ respectively) was optimized for each random sample by finding the value that maximizes the *Likelihood* of the sample. Results show clearly that QS-based estimates are relatively unbiased, even for small sample sizes, whereas KD- and BC-based estimates can have significant negative bias when sample sizes are small.

The fact that QS differs from BC in one very important way may help to explain the properties noted above. Whereas in BC we choose the “bin” size and locations, and then compute the “probability mass” estimates for each bin from the data, in QS we instead choose the “probability mass” size (by specifying the number of quantiles) and then compute the “bin” sizes and locations (to conform to the spacings between quantiles) from the data. In doing so, BC uses only the samples falling within a particular bin to compute each probability mass estimate, which value can (in principle) be highly sensitive to sampling variability unless the number of samples (in each bin) is sufficiently large. In contrast, QS uses a potentially large number of samples from the data to generate a smoothed (via subsampling and averaging) estimate of the position of each quantile. As shown in Figure 2, the estimation bias and uncertainty are small for most of the quantiles and may only be significant near the extreme tails of the density, and for smaller sample sizes. In principle, therefore, with its focus on estimating quantiles rather than probability masses, the QS method seems to provide a more efficient use of the information in the data, and thereby a more robust approximation of the shape of the pdf.

In this paper, we have used a simple, perhaps naïve, way of estimating the locations of the quantiles. Future work could investigate more sophisticated methods where the bias associated with extreme quantiles is accounted for and corrected. These include both Kernel and non-parametric methods. The simplest non-parametric methods are the empirical quantile estimator based on a single order statistic, or the extension based on two consecutive order statistics [23], for which the variance can be large. Quantile estimators based on L-statistics have been explored as a way to reduce estimation variance [24–26], and include Kernel quantile estimators [27–31]. However, performance of the latter can be very sensitive to the choice of bandwidth. More recently, quantile L-estimators intended to be efficient at small sample sizes for estimating quantiles in the tails of a distribution have also been proposed [32,33]. Finally, quantile estimators based on Bernstein polynomials [34,35] and importance sampling (see [36] and references therein) have been also investigated.

Note that the small-sample efficiency of the QS method may be affected by the fact that the entropy fractions associated with the extreme upper and lower end “bins” (quantile spacings) can be quite large when a small number of quantiles is used (see Figure 3). Our implementation of QS seems to successfully compensate for lack of exact knowledge of z_0 and z_{N_z} by using as empirical estimates the values x_{min} and x_{max} in the data sample. Intuitively, one would expect that wider “bins” could be used in regions where the slopes of the entropy fraction curves are flatter (e.g., the center of the *Gaussian* density), and narrower “bins” in the tails (more like the BC method). Taken to its logical conclusion, an ideal approach might be to use “bin” locations and widths such that the cumulative value of $-\ln(p) \cdot p$ for any given pdf is subdivided into equal intervals (i.e., equal fractional entropy spacings) such that each bin then contributes approximately the same amount to the summation in Equation (9). To achieve this, the challenge is to estimate the edge locations of these “bins” (analogous to locations of the quantiles) from the sample data; we leave the possibility of such an approach for future investigation.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We are grateful to the two anonymous reviewers who provided comments and suggestions that helped to improve this paper, and especially to reviewer #2 who insisted that we provide a comparison with the Parzen-window kernel density method. We also acknowledge the many members of the *GeoInfoTheory* community (<https://geoinftheory.org> accessed on 1 March 2021) who have provided both moral support and extensive discussion of matters related to Information Theory and its applications to the Geosciences; without their existence and enthusiastic engagement it is unlikely that the ideas leading to this manuscript would have occurred to us. The first author acknowledges partial support by the *Australian Research Council Centre of Excellence for Climate Extremes* (CE170100023). The second and sixth authors acknowledge partial support by the University of Arizona Earth Dynamics Observatory, funded by the Office of Research, Innovation and Impact. The QS and BC algorithms used in this work are freely accessible for non-commercial use at <https://github.com/rehsani/Entropy> accessed on 20 February 2021. The KD algorithm used in this work is downloadable from <https://webee.technion.ac.il/~yoav/research/blind-separation.html> accessed on 13 May 2021.

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

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2006.
3. Gong, W.; Yang, D.; Gupta, H.V.; Nearing, G. Estimating information entropy for hydrological data: One dimensional case. *Tech. Note Water Resour. Res.* **2014**, *50*, 5003–5018. [CrossRef]
4. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: New York, NY, USA, 1986.
5. Beirlant, J.; Dudewicz, E.J.; Gyorfi, L.; van der Meulen, E.C. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17–39.
6. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley: New York, NY, USA, 2008.
7. Loftsgarden, D.O.; Quesenberry, C.D. A Nonparametric Estimate of a Multivariate Density Function. *Ann. Math. Stat.* **1965**, *36*, 1049–1051. [CrossRef]
8. Scott, D.W. Optimal and data-based histograms. *Biometrika* **1979**, *66*, 605–610. [CrossRef]
9. Scott, D.W.; Thompson, J.R. Probability Density Estimation in Higher Dimensions. In *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*; Gentle, J.E., Ed.; North-Holland: Amsterdam, The Netherlands, 1983; pp. 173–179.
10. Scott, D.W. *Handbook of Computational Statistics—Concepts and Methods*; Springer: New York, NY, USA, 2004.
11. Scott, D.W. Averaged shifted histograms—Effective nonparametric density estimators in several dimensions. *Ann. Stat.* **1985**, *13*, 1024–1040. [CrossRef]
12. Wegman, E.J. Maximum likelihood estimation of a probability density function. *Sankhyā Indian J. Stat.* **1975**, *3*, 211–224.
13. Denison, D.G.T.; Adams, N.M.; Holmes, C.C.; Hand, D.J. Bayesian partition modelling. *Comput. Stat. Data Anal.* **2002**, *38*, 475–485. [CrossRef]
14. Jackson, B.; Scargle, J.; Barnes, D.; Arabhi, S.; Alt, A.; Gioumousis, P.; Gwin, E.; Sangtrakulcharoen, P.; Tan, L.; Tsai, T.T. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **2005**, *12*, 105–108. [CrossRef]
15. Endres, A.; Foldiak, P. Bayesian bin distribution inference and mutual information. *IEEE Trans. Inf. Theory* **2005**, *51*, 3766–3779. [CrossRef]
16. Hutter, M. Exact Bayesian regression of piecewise constant functions. *Bayesian An.* **2007**, *2*, 635–664. [CrossRef]
17. Darscheid, P.; Guthke, A.; Ehret, U. A maximum-entropy method to estimate discrete distributions from samples ensuring nonzero probabilities. *Entropy* **2018**, *20*, 601. [CrossRef]
18. Sturges, H.A. The choice of a class interval. *J. Am. Stat. Assoc.* **1926**, *21*, 65–66. [CrossRef]
19. Freedman, D.; Diaconis, P. On the histogram as a density estimator: L2 theory. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* **1981**, *57*, 453–476. [CrossRef]
20. Knuth, K.H. Optimal data-based binning for histograms and histogram-based probability density models. *Digit. Signal Process.* **2019**, *95*, 102581. [CrossRef]
21. Schwartz, S.; Zibulevsky, M.; Schechner, Y.Y. Fast kernel entropy estimation and optimization. *Signal Process.* **2005**, *85*, 1045–1058. [CrossRef]
22. Viola, P.A. Alignment by Maximization of Mutual Information. Ph.D. Thesis, Massachusetts Institute of Technology—Artificial Intelligence Laboratory, Cambridge, MA, USA, 1995.
23. Hyndman, R.J.; Fan, Y. Sample quantiles in statistical packages. *Am. Stat.* **1996**, *50*, 361–365.
24. Harrell, F.E.; Davis, C.E. A new distribution-free quantile estimator. *Biometrika* **1982**, *69*, 635–640. [CrossRef]
25. Kaigh, W.D.; Lachenbruch, P.A. A Generalized Quantile Estimator. *Commun. Stat. Part A Theory Methods* **1982**, *11*, 2217–2238.
26. Brodin, E. On quantile estimation by bootstrap. *Comput. Stat. Data Anal.* **2006**, *50*, 1398–1406. [CrossRef]

27. Parzen, E. Nonparametric Statistical Data Modeling. *J. Am. Stat. Assoc.* **1979**, *74*, 105–131. [[CrossRef](#)]
28. Sheather, J.S.; Marron, J.S. Kernel quantile estimators. *J. Am. Stat. Assoc.* **1990**, *85*, 410–416. [[CrossRef](#)]
29. Cheng, C.; Parzen, E. Unified estimators of smooth quantile and quantile density functions. *J. Stat. Plan. Inference* **1997**, *59*, 291–307. [[CrossRef](#)]
30. Park, C. *Smooth Nonparametric Estimation of a Quantile Function under Right Censoring Using Beta Kernels*; Technical Report (TR 2006-01-CP); Department of Mathematical Sciences, Clemson University: Clemson, SC, USA, 2006.
31. Yang, S.S. A smooth nonparametric estimator of a quantile function. *J. Am. Stat. Assoc.* **1985**, *80*, 1004–1011. [[CrossRef](#)]
32. Sfakianakis, M.E.; Verginis, D.G. A new family of nonparametric quantile estimators. *Commun. Stat. Simul. Comput.* **2008**, *37*, 337–345. [[CrossRef](#)]
33. Navruz, G.; Özdemir, A.F. A new quantile estimator with weights based on a subsampling approach. *Br. J. Math. Stat. Psychol.* **2020**, *73*, 506–521. [[CrossRef](#)] [[PubMed](#)]
34. Cheng, C. The Bernstein polynomial estimator of a smooth quantile function. *Stat. Probab. Lett.* **1995**, *24*, 321–330. [[CrossRef](#)]
35. Pepelyshev, A.; Rafajłowicz, E.; Steland, A. Estimation of the quantile function using Bernstein–Durrmeyer polynomials. *J. Nonparametric Stat.* **2014**, *26*, 1–20. [[CrossRef](#)]
36. Kohler, M.; Tent, R. Nonparametric quantile estimation using surrogate models and importance sampling. *Metrika* **2020**, *83*, 141–169. [[CrossRef](#)]