entropy

MDPI

Article

An Invariant Measure for Differential Entropy: From Kullback–Leibler Divergence to Scale-Invariant Information Theory

Félix Truong [1,*] **and Alexandre Giuliani** [1,2]

[1] Synchrotron SOLEIL, L'Orme des Merisiers, 91190 Saint-Aubin, France; alexandre.giuliani@synchrotron-soleil.fr
[2] UAR1008, Transform Department, INRAE, 44316 Nantes, France
* Correspondence: felix.servant7@gmail.com

Abstract

Shannon's differential entropy for continuous variables suffers from a fundamental limitation: it is not invariant under scale transformations. This makes entropy values dependent on the choice of measurement units rather than reflecting intrinsic properties of distributions. While Jaynes proposed the limiting density of discrete points (LDDP) as a theoretical solution, a concrete method for computing the required invariant measure has been lacking. This paper establishes a rigorous connection between Kullback–Leibler divergence and the invariant measure, providing theoretical proofs of invariance under affine transformations and a practical computational method. We prove that entropy normalized by the median of k-nearest neighbor distances is invariant under affine transformations (Theorems 1 and 2). The non-negativity of the resulting entropy has been validated empirically across all tested distribution families, though a complete theoretical proof remains an open question. This approach extends naturally to multivariate settings, enabling scale-invariant mutual information estimation. We provide open-source implementations in Julia (EntropyInvariant.jl) and Python (entropy_invariant) and demonstrate their advantages over traditional approaches, particularly for variables with disparate scales.

Keywords: differential entropy; invariant measure; Kullback–Leibler divergence; limiting density of discrete points; k-nearest neighbor; mutual information; scale invariance

1. Introduction

Information theory provides powerful tools for quantifying relationships between variables across scientific disciplines. Among these, mutual information (MI) stands out for its ability to capture both linear and nonlinear dependencies while remaining robust to small sample sizes [1–3]. Unlike correlation-based measures, MI is sensitive to the complete dependence structure between variables, making it particularly valuable for complex data analysis [4].

The foundation of MI lies in Shannon's entropy framework [5], originally defined for discrete variables and later extended to continuous variables as differential entropy. For a continuous random variable X with probability density function μ_X, the differential entropy is:

$$h(X) = -\int_X \mu_X(x) \cdot \log(\mu_X(x)) \cdot \mathrm{d}x \tag{1}$$

Academic Editor: Geert Verdoolaege

Received: 29 January 2026
Revised: 26 February 2026
Accepted: 28 February 2026
Published: 7 March 2026

This definition applies to both bounded and unbounded domains, provided the integral converges. Our invariant measure methodology requires that the median of k-nearest neighbor distances stabilizes with increasing sample size, a condition empirically satisfied even for heavy-tailed distributions such as Cauchy and Lévy (Table 2).

However, differential entropy suffers from a critical limitation: it is not invariant under scale transformations. Under a linear transformation $Y = aX$, the entropy transforms as:

$$h(aX) = h(X) + \log(|a|) \tag{2}$$

This scale-dependence means that entropy values depend on measurement units rather than reflecting intrinsic properties of the distribution. For instance, measuring temperature in Celsius versus Fahrenheit yields different entropy values, even though the underlying physical system is identical. This limitation extends to mutual information, potentially leading to misleading results when comparing variables with different scales.

MI is zero when X and Y are independent and increases with the dependence between variables. This measure is always non-negative and has no upper bound, expressed as $MI(X, Y) \geq 0$. MI can be expressed in terms of differential entropy as:

$$MI(X; Y) = h(X) + h(Y) - h(X, Y) \tag{3}$$

where $h(X, Y)$ denotes the joint differential entropy. Equivalently, $MI(X; Y) = D_{KL}(P_{X,Y} \| P_X \otimes P_Y)$, measuring the KL divergence from the joint distribution to the product of marginals. While MI shares some metric-like properties (non-negativity, symmetry), it does not satisfy the triangle inequality and is therefore not a true metric in the mathematical sense; we use the term "measure of dependence" throughout.

A significant challenge in many practical applications lies in estimating entropy from finite samples when the underlying probability density function (pdf) is unknown. Techniques for entropy estimation can be broadly classified into two main categories: parametric and nonparametric methods [6]. Parametric methods assume that the form of the pdf is known, reducing the problem to estimating the parameters from the data [1]. Nonparametric methods do not make assumptions about the form of the pdf, making them more versatile and widely applicable. Among these, popular approaches include histogram-based methods [7], kernel density estimators (KDEs) [8], and entropy-based statistical tests [9]. Another nonparametric approach is based on the k-nearest neighbors (kNN) method [3,6], which is computationally efficient and has been shown to be robust even with small sample sizes [10,11].

Despite their popularity, traditional nonparametric methods have significant limitations. Histogram-based methods are highly sensitive to the choice of bin width: too few bins lead to oversmoothing and information loss, while too many bins result in high variance and systematic bias that increases with the number of bins. The optimal bin width is data-dependent and difficult to determine a priori. KDE methods face similar bandwidth selection challenges. While the kNN method addresses some of these issues by using a single, interpretable parameter k and demonstrating good convergence properties, it remains sensitive to scale transformations. For comprehensive coverage of information-theoretic foundations, we refer readers to [12].

The root of the scale-dependence problem lies in the transition from discrete to continuous entropy. To understand this, consider the Kullback–Leibler (KL) divergence [13], which measures the dissimilarity between distributions P and Q:

$$D_{KL}(P||Q) = \int_X \mu_P(x) \cdot \log\left(\frac{\mu_P(x)}{\mu_Q(x)}\right) \cdot \mathrm{d}x \tag{4}$$

The KL divergence relates to differential entropy through:

$$D_{KL}(P||Q) = -h(P) - \mathbb{E}_P[\log(\mu_Q(X))] \tag{5}$$

When Q is a uniform distribution over an interval of length r, we obtain:

$$h(P) = \log(r) - D_{KL}(P||U_r) \tag{6}$$

This reveals that differential entropy implicitly compares the distribution to a uniform reference whose scale affects the entropy value. The scale-dependence arises because the reference scale r changes under transformations.

Jaynes [14] recognized this fundamental issue and proposed addressing it through an "invariant measure" $m(x)$ representing complete ignorance about the probability distribution. This leads to the limiting density of discrete points (LDDP):

$$H_c(X) = -\int_X \mu_X(x) \cdot \log\left(\frac{\mu_X(x)}{m(x)}\right) \cdot \mathrm{d}x \tag{7}$$

Comparing Equations (4) and (7), we observe that the LDDP can be interpreted as a KL divergence where the reference distribution is replaced by the invariant measure. However, Jaynes did not provide a concrete method for computing $m(x)$, leaving the central question unanswered: what properties must $m(x)$ satisfy, and how can it be estimated from data?

Recent work by Nagel and coworkers [15] proposed normalizing MI using an invariant measure, but their approach introduces inconsistencies: the marginal entropies vary depending on which variables are paired together, violating the fundamental principle that a variable's entropy should be intrinsic to that variable alone.

Our contributions: This paper establishes a rigorous connection between the KL divergence framework and Jaynes's invariant measure concept. We demonstrate that:

- The invariant measure $m(x)$ can be rigorously defined through specific mathematical properties and computed from data using k-nearest neighbor distances;
- This measure naturally emerges from requiring transformation invariance analogous to the KL divergence framework;
- The resulting invariant entropy corresponds to the LDDP and is truly scale-invariant;
- The approach generalizes naturally to multivariate settings, enabling consistent scale-invariant MI estimation;
- The median of nearest neighbor distances provides a robust estimator that avoids negative entropy values and identifies distribution families.

The present manuscript is organized as follows. Section 2 develops the theoretical framework connecting KL divergence to the invariant measure, proves invariance properties, and presents the computational method. Section 3 validates the approach through simulations and demonstrates its advantages over traditional methods. Section 4 discusses connections to information geometry, maximum entropy principles, and broader implications. Section 5 concludes with future directions.

2. Materials and Methods

2.1. From Kullback–Leibler Divergence to Invariant Measure

To understand the connection between KL divergence and invariant entropy, we examine what happens when comparing a distribution to a uniform reference. We use a bounded interval here for pedagogical clarity; the resulting invariant measure framework applies broadly to distributions with bounded or unbounded support, as demonstrated

in Section 3. Consider a uniform distribution $U_{[a,b]}$ with density $\mu_U(x) = 1/(b-a)$ for $x \in [a,b]$.

For a distribution P with support contained in $[a,b]$, the KL divergence is:

$$D_{KL}(P||U_{[a,b]}) = -h(P) + \log(b-a) \tag{8}$$

Under an affine transformation $Y = aX + b$, the uniform reference transforms to $U_{[aa+b,ab+b]}$ with density $1/(|a|(b-a))$. The KL divergence becomes:

$$D_{KL}(P_Y||U_Y) = -h(X) + \log(b-a) = D_{KL}(P_X||U_X) \tag{9}$$

This demonstrates that the KL divergence to a uniform reference is invariant under affine transformations, but differential entropy is not because the log-volume term changes. The key insight is that we need a reference measure that adapts to the data scale in a way that removes this scale-dependence.

2.2. Definition of the Invariant Measure

We propose that the invariant measure $m(x)$ should satisfy properties that ensure scale invariance while remaining practically computable from data.

Proposition 1. *Let $m : \mathbb{R} \to \mathbb{R}^+$ be an invariant measure function. We require:*

(i) Positivity: $m(X) = r_X > 0$ for any random variable X;
(ii) Scale equivariance: $m(aX) = |a| \cdot m(X)$ for any $a \neq 0$;
(iii) Translation invariance: $m(X + b) = m(X)$ for any $b \in \mathbb{R}$;
(iv) Consistency with KL divergence: The measure should lead to an entropy-like quantity that behaves as a KL divergence from the data distribution to a reference distribution.

These properties ensure that $m(X)$ captures the intrinsic scale of the distribution independent of affine transformations. We now define the invariant differential entropy as:

$$h_c(X) := h\left(\frac{X}{m(X)}\right) \tag{10}$$

Theorem 1 (Invariance of h_c)**.** *Let $m(X)$ satisfy Properties (i)–(iii) of Proposition 1. Then, for any transformation $Y = aX + b$ with $a \neq 0$:*

$$h_c(Y) = h_c(X) \tag{11}$$

Proof. Let $Y = aX + b$. Based on Properties (iii) and (ii), we have:

$$m(aX + b) = m(aX) = |a| \cdot m(X) \tag{12}$$

Therefore:

$$h_c(Y) = h\left(\frac{Y}{m(Y)}\right) = h\left(\frac{aX + b}{|a| \cdot m(X)}\right) \tag{13}$$

Let $Z = X/m(X)$ be the normalized variable. Under the transformation $Y = aX + b$, we have:

$$\frac{Y}{m(Y)} = \frac{aX + b}{|a| \cdot m(X)} = \text{sgn}(a) \cdot \frac{X}{m(X)} + \frac{b}{|a| \cdot m(X)} = \text{sgn}(a) \cdot Z + c \tag{14}$$

where $c = b/(|a| \cdot m(X))$ is a constant and $\text{sgn}(a) = \pm 1$.

Now, we apply the change-of-variables formula for differential entropy. For a random variable Z with density $\mu_Z(z)$ and a transformation $W = g(Z)$ where g is differentiable and invertible, the density of W is:

$$\mu_W(w) = \mu_Z(g^{-1}(w)) \cdot \left| \frac{dg^{-1}}{dw} \right| \tag{15}$$

For a linear transformation $W = aZ + c$, we have $g^{-1}(w) = (w - c)/a$ and $|dg^{-1}/dw| = 1/|a|$, giving:

$$h(aZ + c) = -\int \mu_W(w) \log(\mu_W(w)) dw = -\int \frac{1}{|a|} \mu_Z\left(\frac{w - c}{a}\right) \log\left(\frac{1}{|a|} \mu_Z\left(\frac{w - c}{a}\right)\right) dw \tag{16}$$

Substituting $u = (w - c)/a$, $dw = |a|du$:

$$h(aZ + c) = -\int \mu_Z(u) \log\left(\frac{1}{|a|} \mu_Z(u)\right) |a| \frac{du}{|a|} = -\int \mu_Z(u) [\log(\mu_Z(u)) - \log(|a|)] du \tag{17}$$

$$= -\int \mu_Z(u) \log(\mu_Z(u)) du + \log(|a|) \int \mu_Z(u) du = h(Z) + \log(|a|) \tag{18}$$

However, for our normalized variable, $a = \text{sgn}(a)$ has $|a| = 1$, so:

$$h(\text{sgn}(a) \cdot Z + c) = h(Z) + \log(1) = h(Z) = h\left(\frac{X}{m(X)}\right) = h_c(X) \tag{19}$$

Therefore, $h_c(Y) = h_c(X)$, establishing invariance. □

2.3. Connection to Jaynes's LDDP

We now show that $h_c(X)$ is equivalent to Jaynes's LDDP. Using the change of variables $u = x/m(X)$, we have $x = m(X) \cdot u$ and $dx = m(X) \cdot du$. The density transforms as:

$$\mu_{X/m(X)}(u) = m(X) \cdot \mu_X(m(X) \cdot u) \tag{20}$$

Therefore:

$$\begin{aligned} h_c(X) = h\left(\frac{X}{m(X)}\right) &= -\int \mu_{X/m(X)}(u) \log(\mu_{X/m(X)}(u)) du \\ &= -\int m(X) \mu_X(m(X)u) \log(m(X)\mu_X(m(X)u)) du \\ &= -\int \mu_X(x) \log(m(X)\mu_X(x)) \frac{dx}{m(X)} \\ &= -\int \mu_X(x) [\log(\mu_X(x)) + \log(m(X))] dx \\ &= h(X) - \log(m(X)) \end{aligned} \tag{21}$$

This can be expressed as:

$$h_c(X) = -\int \mu_X(x) \log\left(\frac{\mu_X(x)}{1/m(X)}\right) dx \tag{22}$$

which is precisely Jaynes's expression (7) with constant invariant measure $m(x) = 1/m(X)$. Hence, as mentioned in the Introduction, the invariant entropy corresponds to comparing the distribution to a uniform reference over an interval of length $m(X)$, which adapts to the data scale.

2.4. Estimation of the Invariant Measure

To make this framework practical, we need a method to estimate $m(X)$ from data. We propose using the k-nearest neighbor (kNN) approach, which aligns naturally with kNN-based entropy estimation methods.

Given a sample $\{x_1, x_2, \ldots, x_n\}$ from X, we compute the distance from each point to its k-th nearest neighbor:

$$d_i^{(k)} = \min_{j \neq i}^{(k)} |x_i - x_j|, \quad i = 1, \ldots, n \tag{23}$$

where $\min^{(k)}$ denotes the k-th smallest value. For simplicity, we focus on $k = 1$ (nearest neighbor). The vector of nearest neighbor distances $\mathbf{d} = (d_1, \ldots, d_n)$ captures the local density structure. We propose:

$$m(X) = \text{median}(\mathbf{d}) \tag{24}$$

Justification for using the median:

(1) Robustness: The median is robust to outliers, reflecting "complete ignorance" about points far from the data bulk. Consider $X = \{2, 3, 5, 7, 11, 17, 19, 23, 29\}$ with nearest neighbor vector $\mathbf{d} = \{1, 1, 2, 2, 4, 2, 2, 4, 6\}$ (sorted: $\{1, 1, 2, 2, 2, 2, 4, 4, 6\}$), giving $m(X) = 2$ and $\text{mean}(\mathbf{d}) = 2.67$. Adding an outlier $x_{10} = 100$ yields $d_{10} = |100 - 29| = 71$ and $\mathbf{d}' = \{1, 1, 2, 2, 4, 2, 2, 4, 6, 71\}$, still giving $m(X') = 2$. The mean would change from 2.67 to 9.50, demonstrating inferior robustness.

(2) Scale equivariance: For scaled data aX, nearest neighbor distances scale as $|a|d_i$, so $\text{median}(|a|\mathbf{d}) = |a|\ \text{median}(\mathbf{d})$, satisfying Property (ii) of Proposition 1.

(3) Translation invariance: Translating data by b does not change distances, so $\text{median}(\mathbf{d})$ remains unchanged, satisfying Property (iii).

(4) Avoids negative entropy: Using the mean can lead to negative entropy. For example, the exponential and normal distributions yield negative values with the mean-based measure (Table 1):

Table 1. Comparison of entropy values using mean versus median invariant measure for exponential and normal distributions. We use standard parameterizations: $\mathcal{E}(\lambda)$ with density $\lambda e^{-\lambda x}$ for $x \geq 0$ (mean $= 1/\lambda$); $\mathcal{N}(\mu, \sigma)$ with density $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$. Numerical values use $\lambda = 1$ and $\mathcal{N}(0, 1)$.

Entropy (Mean)	Entropy (Median)
$h_{c_{\text{mean}}}(\mathcal{E}(\lambda)) = -0.06$	$h_{c_{\text{median}}}(\mathcal{E}(\lambda)) = 1.22$
$h_{c_{\text{mean}}}(\mathcal{N}(\mu, \sigma)) = -0.80$	$h_{c_{\text{median}}}(\mathcal{N}(\mu, \sigma)) = 1.14$

Empirically, the median avoids negative entropy values while preserving the desired invariance properties. This robustness likely stems from the median's optimality as an L^1 center, which is less sensitive to extreme values in the distance distribution than the mean (L^2 center). A rigorous proof establishing conditions under which the median measure guarantees non-negative entropy for all distribution classes remains an important open theoretical question.

Theoretical status: Our main results (Theorems 1 and 2) rigorously establish that the median-based invariant measure satisfies scale equivariance and translation invariance. The non-negativity of $h_c(X)$ has been verified empirically for all distribution families in Table 2, but a general proof guaranteeing $h_c(X) \geq 0$ for all distributions remains an important open question.

Table 2. Invariant entropy for common distributions. Each value represents the mean ± standard deviation over 100 simulations with 10,000 samples each. For bounded distributions: Arcsine$(0,1)$, Uniform$(0,1)$, Semicircle with $r=1$, Triangular$(0,1,0.5)$, Cosine$(\mu=0.5, s=1)$. For unbounded distributions: standard parameters were used (e.g., $\mathcal{N}(0,1)$, $\mathcal{E}(1)$). Due to the invariance property, these values remain constant for any choice of location and scale parameters within each distribution family.

Distribution	Invariant Entropy
Arcsine (a,b)	1.008 ± 0.006
Uniform (a,b)	1.060 ± 0.005
Semicircle (r)	1.073 ± 0.005
Triangular (μ,σ)	1.106 ± 0.005
Cosine (μ,σ)	1.114 ± 0.005
Normal (μ,σ)	1.150 ± 0.005
Rayleigh (σ)	1.135 ± 0.005
Chi (ν)	1.149 ± 0.005
Gumbel (μ,σ)	1.174 ± 0.005
Logistic (ν)	1.184 ± 0.005
Exponential (θ)	1.227 ± 0.005
Laplace (μ,σ)	1.227 ± 0.005
Cauchy (μ,σ)	1.517 ± 0.006
Levy (μ,σ)	1.973 ± 0.008

2.5. Multivariate and Multidimensional Generalization

We distinguish between the multivariate setting (multiple random variables $X_1,\ldots,X_n$, each potentially vector-valued) and the multidimensional setting (a single variable $\mathbf{X} \in \mathbb{R}^d$). For a multidimensional variable, $m(\mathbf{X})$ is computed from kNN distances in $\mathbb{R}^d$. For multiple variables, each $m(X_i)$ is computed from the marginal distribution independently, and the joint measure uses the product form $m(X_1,\ldots,X_n) = \prod_i m(X_i)$.

The extension to multiple variables follows naturally from the KL divergence perspective. For two random variables (X,Y), the joint LDDP is:

$$H_c(X,Y) = -\int\int \mu_{X,Y}(x,y)\log\left(\frac{\mu_{X,Y}(x,y)}{m(x,y)}\right)dxdy \tag{25}$$

Proposition 2 (Separable invariant measure)**.** *For independent scale transformations, the invariant measure for the joint distribution should satisfy:*

$$m(x,y) = m_X(x)\cdot m_Y(y) \tag{26}$$

where m_X and m_Y are the marginal invariant measures.

Theorem 2 (Joint invariant entropy)**.** *Let (X,Y) be jointly distributed random variables with joint density $\mu_{X,Y}$, marginal densities μ_X and μ_Y, and marginal invariant measures $m(X)$ and $m(Y)$ computed from the respective marginal samples. Define the normalized variables $X' = X/m(X)$ and $Y' = Y/m(Y)$. The invariant joint entropy is:*

$$h_c(X,Y) = h\left(\frac{X}{m(X)}, \frac{Y}{m(Y)}\right) \tag{27}$$

This is invariant under independent affine transformations $(X,Y) \to (a_1X+b_1, a_2Y+b_2)$, for any $a_1,a_2 \neq 0$ and $b_1,b_2 \in \mathbb{R}$.

Geometric interpretation via kNN distances:

The kNN entropy estimator in 2D uses Euclidean distances. For points $A = (x_i, y_i)$ and $B = (x_j, y_j)$:

$$d(A, B) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{28}$$

Under uniform transformation $(x, y) \to (kx, ky)$, distances scale as $d \to kd$. However, for independent transformations $(x, y) \to (k_1 x, k_2 y)$:

$$d(k_1 A, k_2 B) = \sqrt{k_1^2 (x_i - x_j)^2 + k_2^2 (y_i - y_j)^2} \tag{29}$$

which is not simply proportional to $d(A, B)$.

By normalizing each coordinate by its invariant measure:

$$A' = \left(\frac{x_i}{m(X)}, \frac{y_i}{m(Y)} \right), \quad B' = \left(\frac{x_j}{m(X)}, \frac{y_j}{m(Y)} \right) \tag{30}$$

The distance becomes:

$$\begin{aligned} d(k_1 A', k_2 B') &= \sqrt{\left(k_1 \frac{x_i}{m(k_1 X)} - k_1 \frac{x_j}{m(k_1 X)} \right)^2 + \left(k_2 \frac{y_i}{m(k_2 Y)} - k_2 \frac{y_j}{m(k_2 Y)} \right)^2} \\ &= \sqrt{\left(\frac{x_i - x_j}{m(X)} \right)^2 + \left(\frac{y_i - y_j}{m(Y)} \right)^2} \\ &= d(A', B') \end{aligned} \tag{31}$$

This shows that the normalized coordinates create a natural reference frame where distances are invariant under independent scale transformations—precisely what is needed for invariant entropy estimation.

By normalizing each coordinate by its invariant measure, we create a natural reference frame where distances are invariant under independent scale transformations. The invariant mutual information is then:

$$MI_c(X, Y) = h_c(X) + h_c(Y) - h_c(X, Y) \tag{32}$$

3. Results

3.1. Validation with Standard Distributions

To validate our invariant measure approach, we performed extensive simulations comparing it with traditional kNN and histogram methods. The purpose of Figure 1 is to test the core invariance property: distributions differing only in scale parameters should yield identical invariant entropy. Each column corresponds to a different distribution family (uniform, normal, exponential), with three scale parameter values overlaid. The top panels show convergence with sample size, while the bottom panels show stability across the number of neighbors k.

The key observation is that distributions with the same shape but different scale parameters yield identical invariant entropy values (overlapping curves in Figure 1). This confirms that the invariant measure successfully removes scale dependence. The convergence is rapid, with the estimation stabilizing at approximately 1500 samples. Moreover, the standard deviation is small and remains consistent across different parameter values, demonstrating the robustness of the approach.

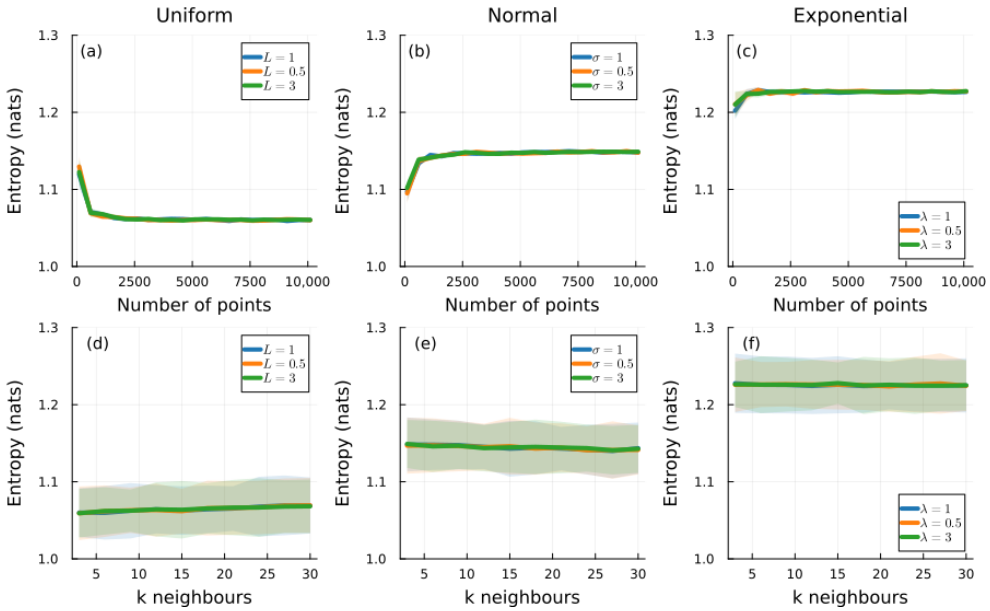

Figure 1. Invariant entropy estimation for: (**a**,**d**) uniform distribution $\mathcal{U}(0, b)$ with $b = [1, 0.5, 3]$; (**b**,**e**) normal distribution $\mathcal{N}(0, \sigma)$ with $\sigma = [1, 0.5, 3]$; (**c**,**f**) exponential distribution $\mathcal{E}(\lambda)$ with $\lambda = [1, 0.5, 3]$. Top panels show entropy versus sample size; bottom panels show entropy versus number of neighbors k. The green shaded area denotes the standard deviation from 100 simulations. The uniform distribution has bounded support $[0, b]$; the normal and exponential distributions have unbounded support but rapidly decaying tails, ensuring stable estimation of $m(X)$ from finite samples.

3.2. Distribution Identification via Invariant Entropy

The median-based invariant entropy identifies distribution families independent of their location and scale parameters. Table 2 presents invariant entropy values for common continuous distributions.

Table 2 presents invariant entropy values for 14 common distribution families, ordered from lowest to highest. The ordering reflects a spectrum from highly predictable local structure (arcsine, 1.008) to highly unpredictable local structure (Lévy, 1.973). Several observations emerge. First, the invariant entropy uniquely characterizes each distribution family: all normal distributions share the same value (1.150), all exponential distributions share the same value (1.227), etc. The parameter-free nature of the invariant entropy is demonstrated in Figure 1, where curves for different scale parameters ($\sigma = [0.5, 1, 3]$) overlap completely. Second, the arcsine distribution has the lowest invariant entropy (1.008), even lower than the uniform distribution (1.060). This reflects the arcsine distribution's distinctive property: its probability density concentrates at the boundaries $x = a$ and $x = b$, where $\mu_{\text{Arcsine}}(x) = 1/(\pi\sqrt{(x-a)(b-x)})$. Given the typical nearest-neighbor spacing captured by $m(X)$, points near the boundaries are highly predictable relative to the invariant measure scale, resulting in lower entropy. Third, heavy-tailed distributions like Cauchy and Levy have higher invariant entropy, reflecting their greater unpredictability at the scale of typical nearest-neighbor distances.

3.3. Scale-Invariant Mutual Information

To demonstrate the practical advantage of invariant MI, we simulated three independent variables with vastly different scales: $X \sim \mathcal{N}(0, 0.1)$, $Y \sim \mathcal{N}(0, 1)$, and $Z \sim \mathcal{N}(0, 10)$. These differing standard deviations highlight the advantages of the invariant entropy estimation. Since the variables are independent, the theoretical MI between any pair should be zero.

Figure 2 demonstrates striking differences between methods. The histogram method (panels a,d) achieves scale invariance, with all three MI estimates superposed in panel

(a). However, convergence toward the theoretical value of zero is extremely slow as sample size increases (panel a), requiring thousands of points to approach the correct value. Furthermore, panel (d) reveals significant systematic bias that varies with the number of bins, making parameter selection critical and problematic. The bias increases substantially with larger bin counts, demonstrating a fundamental limitation of the binning approach.

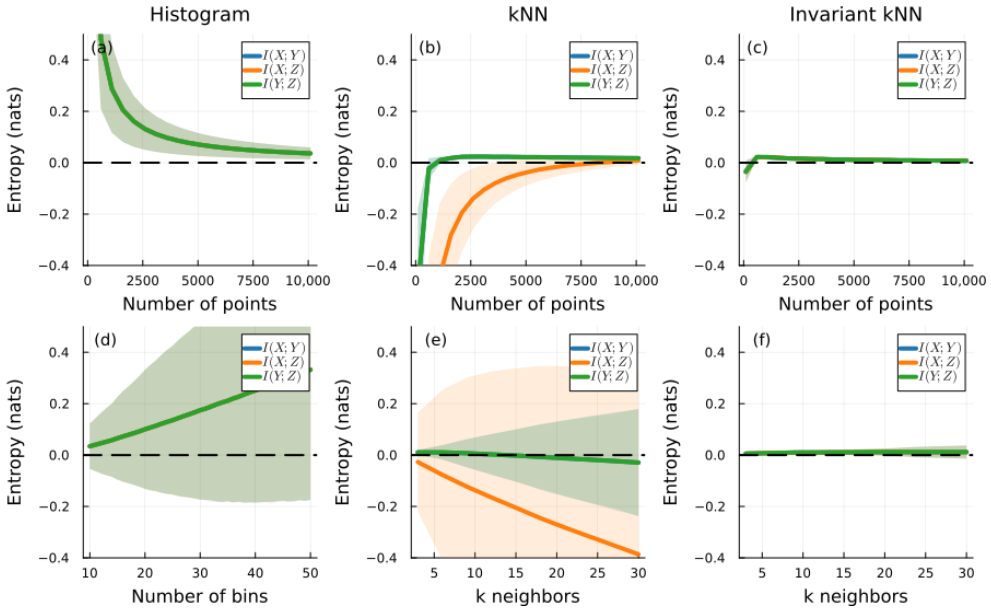

Figure 2. MI estimation between three independent random variables as a function of the number of points (**a**–**c**) and the number of bins/neighbors (**d**–**f**). The histogram method results are shown in (**a**,**d**), the kNN method in (**b**,**e**), and the invariant estimation in (**c**,**f**). The variables follow $X \sim \mathcal{N}(0, 0.1)$, $Y \sim \mathcal{N}(0, 1)$, and $Z \sim \mathcal{N}(0, 10)$. The green shaded area denotes the standard deviation from 100 simulations. All variables are normally distributed with unbounded support; the invariant measure stabilizes due to rapid tail decay despite the 100-fold scale differences. Since the variables are independent, the true MI is zero for all pairs; a successful estimator should yield $MI \approx 0$ regardless of the scale difference.

The kNN method (panels b,e) shows faster convergence than histograms. In panel (b), $I(X;Y)$ and $I(Y;Z)$ are superposed and converge relatively quickly to zero. Panel (e) reveals no bias for $k < 20$—the typical range preferred in the literature where smaller k values are standard. However, a severe breakdown emerges for $I(X;Z)$: in panel (e), this estimate diverges toward increasingly negative values, a physical impossibility since mutual information is non-negative by definition. This failure is not a minor numerical artifact but a fundamental problem: the divergence persists across all values of k (panel e). The breakdown occurs because the kNN estimator implicitly assumes comparable scales when computing joint nearest-neighbor distances. The 100-fold magnitude difference between X (scale ~ 0.1) and Z (scale ~ 10) causes the Z coordinate to completely dominate the Euclidean distance metric, corrupting the joint entropy estimation.

In contrast, the invariant method (panels c,f) demonstrates superior performance across all metrics. In panel (c), all three MI estimates ($I(X;Y)$, $I(Y;Z)$, and $I(X;Z)$) are perfectly superposed, confirming complete scale invariance regardless of the magnitude differences between variables. The convergence toward zero is faster than both the histogram and kNN methods, achieving accurate estimates with fewer than 2500 samples. Crucially, panel (f) shows no bias regardless of the number of neighbors k, eliminating the need for careful parameter tuning. The estimates remain stable and centered near zero across the entire range $k \in [3, 30]$. While finite-sample estimation errors can occasionally produce

small negative values near zero due to statistical fluctuations, the invariant method avoids the divergence that afflicts standard kNN estimation.

These simulations establish the superiority of the invariant approach across multiple dimensions. First, it achieves faster convergence than competing methods, requiring fewer samples to reach accurate estimates. Second, it demonstrates complete scale invariance with all MI pairs collapsing to a single curve, independent of scale differences spanning four orders of magnitude (0.1 to 10). Third, it exhibits no parameter-dependent bias, maintaining stability across a wide range of k values without requiring careful tuning. Fourth, it avoids the catastrophic divergence to negative values that plague traditional kNN estimation when variables have disparate scales.

3.4. Computational Efficiency

The computational complexity of our method is identical to standard kNN entropy estimation: $O(N \log N)$ for sorting and $O(N)$ for nearest neighbor search using KD-trees or ball trees. The additional computation of $m(X)$ via median is $O(N)$, making it negligible compared to the nearest neighbor search. The Julia package EntropyInvariant.jl provides an efficient implementation with performance comparable to standard kNN methods (see the Appendix A for detailed usage examples).

4. Discussion

4.1. Theoretical Contributions

This work establishes a rigorous connection between Kullback–Leibler divergence and Jaynes's limiting density of discrete points. Our main theoretical contributions are:

1. Formalization of the invariant measure: We have shown that the invariant measure $m(X)$ can be understood through the lens of KL divergence as a data-adaptive reference scale. When differential entropy is expressed relative to a uniform distribution, the implicit scale factor $\log(r)$ introduces scale-dependence. By comparing to a reference distribution with characteristic scale $m(X)$ estimated from the data itself, we remove this implicit scale dependence. This is analogous to using an empirical prior in Bayesian statistics: the reference incorporates information about the typical scale of the phenomenon.
2. Bridge between discrete and continuous entropy: The KL divergence framework naturally connects Shannon's discrete entropy to differential entropy. Our invariant measure provides the missing piece for making the continuous case behave like the discrete case with respect to invariance properties. Just as Shannon entropy for discrete variables is invariant to relabeling of outcomes, our invariant entropy for continuous variables is invariant to rescaling of measurement units.
3. Rigorous change-of-variables proof: Unlike previous informal arguments, our proof of invariance (Theorem 1) explicitly uses the change-of-variables formula for probability densities. This establishes the result on solid mathematical foundations and clarifies the role of the Jacobian in transformation properties.
4. Multivariate generalization with geometric interpretation: The extension to joint distributions follows naturally from the separability principle in KL divergence (Proposition 2). The geometric interpretation shows that normalizing by marginal invariant measures creates a coordinate system where Euclidean distances are invariant under independent scale transformations, a property essential for multivariate kNN entropy estimation.

https://doi.org/10.3390/e28030301

4.2. Comparison with Existing Approaches

Traditional histogram methods: As demonstrated in the original work and Figure 2, histogram methods suffer from systematic bias that increases with the number of bins. There is no principled way to choose the optimal bin width, and the method shows severe scale sensitivity. Our approach eliminates these issues by using a data-adaptive scale.

Standard kNN estimator: The standard kNN estimator [3] provides consistent entropy estimates and is computationally efficient. However, it is not scale-invariant, as evidenced by the negative MI values in Figure 2d. Our approach modifies this by normalizing data by $m(X)$ before applying the kNN estimator, yielding $\hat{h}_c(X) = \hat{h}(X/m(X))$. This simple modification preserves all the advantages of kNN methods while adding scale invariance.

Kernel density estimators: KDE methods [8] face similar bandwidth selection challenges as histogram methods. While sophisticated adaptive bandwidth selection procedures exist, they add computational complexity. Our median-based invariant measure provides a simple, robust alternative that requires no tuning beyond the standard k parameter.

Nagel et al.'s approach: Nagel and coworkers [15] proposed normalizing MI by subtracting a scaling factor computed from marginal entropies. However, their normalization has a fundamental flaw: it affects marginal entropies differently depending on which variables are paired together. For example, the normalized entropy of X when computed with Y differs from its normalized entropy when computed with Z. This violates the principle that a variable's entropy should be an intrinsic property. Our approach provides consistent entropy for each variable regardless of which other variables it is paired with, because each variable has its own invariant measure $m(X)$ computed from its marginal distribution.

4.3. Interpretation of Results

Our simulation results demonstrate several key properties:

1. Fast convergence with small samples: The invariant estimator converges faster than traditional methods, particularly when variables have different scales (Figure 2, panels e,f versus c,d). This occurs because normalization by $m(X)$ and $m(Y)$ brings variables to comparable scales before computing joint entropy. The kNN distance calculations then operate in a balanced space where all dimensions contribute equally.
2. Distribution identification: Table 2 shows that distributions maintain consistent invariant entropy values regardless of location and scale parameters. From the KL divergence perspective, this makes sense: distributions in the same family (e.g., all normal distributions) differ only in location parameter μ and scale parameter σ. The invariant measure removes precisely these parameters, leaving only the intrinsic "shape" of the distribution. This property enables distribution classification based solely on shape characteristics.
3. Boundary concentration in arcsine distribution: The arcsine distribution has the lowest invariant entropy (1.008), even lower than uniform (1.060). This initially surprising result reflects a deep property of the arcsine distribution. Its density $\mu(x) = 1/(\pi\sqrt{(x-a)(b-x)})$ diverges at the boundaries, meaning probability mass concentrates there. When we measure entropy relative to the typical nearest-neighbor spacing $m(X)$, points near boundaries are highly predictable, and their neighbors must also be near the boundary. This local predictability, captured by the invariant measure, results in lower entropy despite the distribution appearing "spread out" over $[a, b]$.

4.4. Relationship to Maximum Entropy Principle

The Maximum Entropy (MaxEnt) principle and our invariant entropy framework address different questions. MaxEnt is a distribution selection principle: given constraints

(e.g., fixed mean, fixed variance), it selects the distribution that maximizes entropy, yielding the "least biased" distribution compatible with the constraints. Invariant entropy is an entropy measurement principle: given a distribution (known or empirically observed), it computes an entropy value that is invariant to measurement units. These frameworks are complementary rather than competing.

The MaxEnt principle yields different distributions depending on the constraint structure:

Bounded support constraint: Among all distributions with support contained in a fixed interval $[a, b]$, the uniform distribution $\mathcal{U}(a, b)$ maximizes differential entropy:

$$h_{\text{uniform}} = \log(b - a)$$

For the uniform distribution, the variance $\sigma^2 = (b - a)^2/12$ is determined by the interval bounds, not independently specified.

Fixed variance constraint on $\mathbb{R}$: Among all distributions on $\mathbb{R}$ with fixed variance σ^2, the normal distribution $\mathcal{N}(\mu, \sigma)$ maximizes differential entropy:

$$h_{\text{normal}} = \frac{1}{2}\log(2\pi e\sigma^2)$$

This is fundamentally different: the domain is unbounded, and variance is an independent constraint.

From the invariant entropy perspective, all uniform distributions yield $h_c \approx 1.060$ and all normal distributions yield $h_c \approx 1.150$, regardless of their parameters (Table 2). The uniform has relatively low invariant entropy (1.060). This reflects what the invariant measure captures: the median nearest-neighbor distance reflects the typical local spacing of points. For a uniform distribution, this spacing is highly regular—when we normalize by $m(X)$, all points lie within a predictable range relative to their typical spacing. In contrast, heavy-tailed distributions like Cauchy and Levy exhibit extreme variability in local density: some regions have tightly clustered points while others are sparse. This variability persists after normalization, yielding higher invariant entropy. Distributions with high standard differential entropy (given appropriate constraints) also tend to have high invariant entropy, suggesting consistency between the two frameworks rather than competition.

4.5. Connections to Information Geometry

From the perspective of information geometry [16], our approach can be understood as choosing a coordinate system on the manifold of probability distributions. The Fisher information metric provides a natural Riemannian structure on this manifold, and geodesics correspond to exponential families.

The invariant measure $m(X)$ defines a natural coordinate chart that makes entropy calculations coordinate-independent, analogous to working in canonical coordinates in differential geometry. For a family of distributions $\{p_\theta : \theta \in \Theta\}$, the Fisher metric is:

$$g_{ij}(\theta) = \mathbb{E}\left[\frac{\partial \log p_\theta}{\partial \theta_i}\frac{\partial \log p_\theta}{\partial \theta_j}\right] \tag{33}$$

For location-scale families $p_{\mu,\sigma}(x) = \frac{1}{\sigma}p_0\left(\frac{x-\mu}{\sigma}\right)$, the invariant measure removes the (μ, σ) dependence, effectively projecting onto the "shape manifold" of distributions. Future work could explore whether there is a canonical connection between our invariant measure and the Fisher–Rao metric, potentially leading to a geometric interpretation of the LDDP as arc length on the shape manifold.

https://doi.org/10.3390/e28030301

4.6. Limitations and Future Directions

1. Beyond affine transformations: Our current framework handles affine transformations $(x, y) \to (ax + b, cy + d)$. Extending to general diffeomorphisms $\phi : \mathbb{R}^d \to \mathbb{R}^d$ would require incorporating the Jacobian determinant $|\det D\phi|$. The invariant measure would need to transform as $m(\phi(X)) = m(X) \cdot |\det D\phi|^{1/d}$ to preserve invariance. This generalization could connect to the theory of differential forms and volume elements in differential geometry.

2. Theoretical optimality of the median: While our empirical results demonstrate that the median of nearest-neighbor distances avoids negative entropy and satisfies all required invariance properties, a complete mathematical characterization remains to be established. The median is optimal as an L^1 center (minimizing $\sum |d_i - c|$), while the mean is optimal as an L^2 center (minimizing $\sum (d_i - c)^2$). Different L^p centers induce different geometries on the distance distribution, each preserving the geometric invariance properties but potentially yielding different entropy values. Establishing the minimax optimality of the median among all L^p centers for minimizing the probability of negative entropy across relevant distribution classes, and characterizing the tail conditions under which $P(h_c < 0) \to 0$ as $N \to \infty$, would provide a rigorous theoretical foundation. Such an analysis could parallel the development of M-estimators in robust statistics, potentially establishing conditions under which the L^1 geometry provides the natural reference scale for entropy measurement.

3. Connection to rate-distortion theory: Rate-distortion theory [12] involves minimizing $I(X; Y)$ subject to distortion constraints $\mathbb{E}[d(X, Y)] \leq D$. The invariant MI might provide new insights into scale-invariant coding schemes where the distortion metric itself adapts to the data scale. This could have applications in lossy compression where preservation of "shape" is more important than absolute accuracy.

4. Applications in causality: Invariance under interventions is central to causal inference. Pearl's do-calculus and the invariance principle of Peters et al. both rely on identifying relationships that remain stable across environments. Our scale-invariant MI might help identify causal relationships that persist across different measurement scales or units, potentially improving causal discovery algorithms when variables are measured inconsistently across datasets.

5. Extensions to discrete–continuous mixtures: Many real-world datasets contain both discrete and continuous variables. The MI between discrete and continuous variables is well-defined, but estimating it is challenging [2]. The invariant measure framework could potentially be extended to mixed data types by defining appropriate reference measures for discrete components.

5. Conclusions

This work establishes a rigorous theoretical foundation for invariant entropy estimation by connecting Jaynes's limiting density of discrete points to the Kullback–Leibler divergence framework. By defining the invariant measure $m(X)$ as the median of nearest-neighbor distances and proving that $h_c(X) = h(X/m(X))$ is truly scale-invariant, we provide the first practical method for computing Jaynes's LDDP. The approach extends naturally to multivariate settings and demonstrates superior performance compared to standard methods, particularly avoiding catastrophic failures when variables have disparate scales.

The invariant entropy represents a logical evolution of Shannon's information theory for continuous variables. Just as Shannon entropy for discrete variables is invariant to relabeling of outcomes, our invariant entropy for continuous variables is invariant to rescaling of measurement units. This makes it a more natural measure of uncertainty for physical quantities that can be measured in different units—temperature in Celsius

versus Fahrenheit, distance in meters versus feet, concentration in molarity versus parts-per-million.

Beyond practical utility, the invariant entropy offers theoretical insights. Table 2 reveals that the arcsine distribution, despite appearing spread over an interval, has the lowest invariant entropy due to boundary concentration effects visible at the nearest-neighbor scale. Conversely, heavy-tailed distributions like Cauchy and Levy have high invariant entropy, reflecting fundamental unpredictability even at local scales. These observations suggest that invariant entropy captures the intrinsic properties of distribution families independent of parametrization.

By grounding this concept in the well-established KL divergence framework, we provide both theoretical justification and practical tools for scale-invariant information-theoretic analysis. The connection to information geometry suggests deeper links between invariant measures and the Fisher–Rao metric on distribution manifolds, opening avenues for future research.

We hope this work will enable new applications across diverse scientific fields where scale invariance is essential: feature selection in machine learning with mixed-unit data, network inference in systems biology where genes have vastly different expression scales, time series analysis comparing signals with different amplitudes, and causal discovery across heterogeneous datasets. Open-source implementations in Julia (EntropyInvariant.jl) and Python (entropy_invariant) make these methods readily accessible to the research community.

Author Contributions: Conceptualization, F.T. and A.G.; methodology, F.T.; software, F.T.; validation, F.T. and A.G.; formal analysis, F.T.; investigation, F.T.; writing—original draft preparation, F.T.; writing—review and editing, F.T. and A.G.; supervision, A.G.; project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: F.T. acknowledges a PhD grant from Synchrotron SOLEIL and INRAE.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Data Availability Statement: The methods described in this paper are implemented in two open-source packages: the EntropyInvariant.jl Julia package (https://github.com/Entropy-Invariant/EntropyInvariant.jl (accessed on 28 January 2026)) and the entropy_invariant Python package (available via PyPI: `pip install entropy_invariant`). Both implementations provide identical functionality and produce numerically equivalent results. All simulation code for reproducing the figures is provided in the Appendix A.

Acknowledgments: F.T. acknowledges a PhD grant from Synchrotron SOLEIL and INRAE. Special thanks go to Laurent Nahon for his unwavering support throughout the project.

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

Abbreviations

The following abbreviations are used in this manuscript:

LDDP	Limiting Density of Discrete Points
kNN	k-Nearest Neighbor
MI	Mutual Information
KL	Kullback–Leibler
KDE	Kernel Density Estimators
IQR	Inter-Quartile Range

Appendix A. Computational Details and Package Usage

The calculations and examples in this manuscript have been implemented using the Julia programming language and the EntropyInvariant.jl package (https://github.com/Entropy-Invariant/EntropyInvariant.jl (accessed on 28 January 2026)). The package is registered through the Julia package manager and can be installed as follows:

```
julia> ] add EntropyInvariant
```

Appendix A.1. Main Entropy Function

The EntropyInvariant package provides a unified interface for computing entropy using histogram, kNN, and invariant methods:

```
entropy(mat::Matrix{<:Real};
        method::String = "inv",
        nbins::Int = 10,
        k::Int = 3,
        base::Real = e,
        verbose::Bool = false,
        degenerate::Bool = false,
        dim::Int = 1
        )::Real
```

The `method` parameter selects the estimation method: ‘‘`inv`’’ for invariant estimation (default), ‘‘`knn`’’ for standard kNN method, or ‘‘`histogram`’’ for histogram-based estimation.

Appendix A.2. Algorithmic Details of Invariant Measure Estimation

The computation of the invariant entropy $h_c(X) = h(X/m(X))$ requires estimating the invariant measure $m(X)$ from data. For a sample $\{x_1, x_2, \ldots, x_n\}$, the algorithm proceeds as follows:

Step 1: Compute nearest-neighbor distances. For each dimension d, sort the data and compute the distance from each point to its nearest neighbor:

$$d_i = \min_{j \neq i} |x_i^{(d)} - x_j^{(d)}|, \quad i = 1, \ldots, n \tag{A1}$$

Step 2: Extract the invariant measure. The median of the distance distribution provides a robust estimate of the characteristic scale:

$$m(X^{(d)}) = \text{median}(\{d_1, d_2, \ldots, d_n\}) \tag{A2}$$

The median is preferred over the mean because it satisfies the required invariance properties and avoids negative entropy values. For the exponential distribution, the mean-based measure yields $h_c = -0.06$ (negative), whereas the median-based measure gives $h_c = 1.22$ (positive).

Step 3: Normalize the data. Each dimension is divided by its invariant measure:

$$X'^{(d)} = \frac{X^{(d)}}{m(X^{(d)})} \tag{A3}$$

Step 4: Apply kNN entropy estimator. The normalized data X' is used in the standard Kraskov estimator, yielding $h_c(X) = h_{\text{kNN}}(X')$.

This procedure ensures that $h_c(aX + b) = h_c(X)$ for any affine transformation, as proven in Theorem 1.

Appendix A.3. Python Implementation

A Python port providing identical functionality is available via PyPI:

```
pip install entropy_invariant
```

The API mirrors the Julia interface and produces numerically identical results. Example usage:

```
from entropy_invariant import entropy, mutual_information
h = entropy(data, method="inv", k=3)
```

Appendix A.4. Additional Functions

The package provides additional information-theoretic quantities beyond those demonstrated in this appendix: `conditional_entropy`, `conditional_mutual_information`, `normalized_mutual_information`, `interaction_information`, and Partial Information Decomposition functions (`redundancy`, `unique`, `synergy`). Complete documentation is available in the package repositories.

Basic usage examples:

```
julia> using EntropyInvariant
julia> data = rand(100, 2)  # 100 points in 2D

# Using the invariant method (default)
julia> h_inv = entropy(data, method="inv", k=3)

# Using the kNN method
julia> h_knn = entropy(data, method="knn", k=5)

# Using the histogram method
julia> h_hist = entropy(data, method="histogram", nbins=10)
```

Appendix A.5. Scale Invariance Demonstration

The following example demonstrates the scale invariance property of the invariant entropy:

```
julia> using EntropyInvariant, Random
julia> Random.seed!(42)
julia> n = 1000
julia> p1 = rand(n)

julia> println("Entropy invariant:")
julia> println(entropy(p1))
julia> println(entropy(1e5 * p1 .- 123.465))
julia> println(entropy(1e-5 * p1 .+ 654.321))

Entropy invariant:
1.087225165954103
1.087225165954023
1.087224434765357
```

All three entropy values are essentially identical despite dramatic scale transformations ($\times 10^5$ and $\times 10^{-5}$) and translations, confirming the invariance property stated in Theorem 1.

Appendix A.6. Reproducing Figure 1: Invariant Entropy Estimation

The following code reproduces the results shown in Figure 1:

```
julia> using EntropyInvariant, Random, Distributions

julia> Random.seed!(42)
julia> N = 100:500:10,100  # sample sizes
julia> KNN = 3:3:30  # k values
julia> NB = 100  # number of simulations

# Initialize storage arrays
julia> inv_knn_nor = zeros(NB, length(KNN), length(N))
julia> inv_knn_uni = zeros(NB, length(KNN), length(N))
julia> inv_knn_exp = zeros(NB, length(KNN), length(N))

# Run simulations
julia> for nb in 1:NB
           for (idx_knn, k) in enumerate(KNN)
               for (idx_n, n) in enumerate(N)
                   # Normal distribution
                   nor_ = rand(Normal(0, 1), n)
                   inv_knn_nor[nb, idx_knn, idx_n] =
                       entropy(nor_, k=k, method="inv")

                   # Uniform distribution
                   uni_ = rand(Uniform(0, 1), n)
                   inv_knn_uni[nb, idx_knn, idx_n] =
                       entropy(uni_, k=k, method="inv")

                   # Exponential distribution
                   exp_ = rand(Exponential(1), n)
                   inv_knn_exp[nb, idx_knn, idx_n] =
                       entropy(exp_, k=k, method="inv")
               end
           end
       end

# Compute means and standard deviations
julia> mean_nor = mean(inv_knn_nor, dims=1)
julia> std_nor = std(inv_knn_nor, dims=1)
# Similar for uniform and exponential...
```

Appendix A.7. Mutual Information Estimation

The package provides a mutual information function that uses the invariant entropy by default:

```
mutual_information(x::Vector{<:Real},
                   y::Vector{<:Real};
                   method::String = "inv",
                   k::Int = 3,
                   nbins::Int = 10)::Real
```

Example: Computing MI between independent variables with extreme scale differences

```
julia> using EntropyInvariant, Random, Distributions

julia> Random.seed!(42)
julia> n = 10,000

# Three independent variables with different scales
julia> X = rand(Normal(0, 0.01), n)
julia> Y = rand(Normal(0, 1), n)
julia> Z = rand(Normal(0, 100), n)

# Compute mutual information (should be \approx 0)
julia> MI_XY_inv = mutual_information(X, Y, method="inv", k=5)
julia> MI_XZ_inv = mutual_information(X, Z, method="inv", k=5)
julia> MI_YZ_inv = mutual_information(Y, Z, method="inv", k=5)

julia> println("MI(X;Y) = ", MI_XY_inv)
julia> println("MI(X;Z) = ", MI_XZ_inv)
julia> println("MI(Y;Z) = ", MI_YZ_inv)

MI(X;Y) = 0.0132
MI(X;Z) = 0.0116
MI(Y;Z) = 0.0115
```

Compare with standard kNN method showing catastrophic failure:

```
julia> MI_XY_knn = mutual_information(X, Y, method="knn", k=5)
julia> MI_XZ_knn = mutual_information(X, Z, method="knn", k=5)
julia> MI_YZ_knn = mutual_information(Y, Z, method="knn", k=5)

julia> println("MI(X;Y) = ", MI_XY_knn)
julia> println("MI(X;Z) = ", MI_XZ_knn)
julia> println("MI(Y;Z) = ", MI_YZ_knn)

MI(X;Y) = 0.0277
MI(X;Z) = -1.3216  # Catastrophic failure: negative MI!
MI(Y;Z) = 0.0226
```

The negative MI value for $I(X;Z)$ in the standard kNN method demonstrates how severely scale differences can corrupt estimation. This physically impossible result (MI must be non-negative) occurs because the kNN estimator computes joint nearest-neighbor distances in the (X, Z) space, where the 100-fold scale difference causes Z to dominate the distance metric. The invariant method avoids this by normalizing each variable to its intrinsic scale before computing joint entropy.

Appendix A.8. Reproducing Figure 2: MI Comparison

The following code generates the data for Figure 2, comparing histogram, kNN, and invariant methods:

```
julia> using EntropyInvariant, Random, Distributions
```

```
julia> Random.seed!(42)
julia> N = 100:500:10,100
julia> BINS = 5:5:50
julia> KNN = 3:3:30
julia> NB = 100

# Initialize storage for histogram method
julia> mi_hist_XY = zeros(NB, length(BINS), length(N))
julia> mi_hist_XZ = zeros(NB, length(BINS), length(N))
julia> mi_hist_YZ = zeros(NB, length(BINS), length(N))

julia> for nb in 1:NB
           for (idx_bins, bins) in enumerate(BINS)
               for (idx_n, n) in enumerate(N)
                   X = rand(Normal(0, 0.1), n)
                   Y = rand(Normal(0, 1), n)
                   Z = rand(Normal(0, 10), n)

                   mi_hist_XY[nb, idx_bins, idx_n] =
                       mutual_information(X, Y,
                           method="histogram", nbins=bins)
                   mi_hist_XZ[nb, idx_bins, idx_n] =
                       mutual_information(X, Z,
                           method="histogram", nbins=bins)
                   mi_hist_YZ[nb, idx_bins, idx_n] =
                       mutual_information(Y, Z,
                           method="histogram", nbins=bins)
               end
           end
       end

# Similar loops for kNN and invariant methods...
```

References

1. Brillinger, D.R. Some data analyses using mutual information. *Braz. J. Probab. Stat.* **2004**, *18*, 163–182. Available online: http://www.jstor.org/stable/43601047 (accessed on 28 January 2026).
2. Ross, B.C. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef] [PubMed]
3. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
4. Steuer, R.; Kurths, J.; Daub, C.O.; Weise, J.; Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **2002**, *18*, S231–S240. [CrossRef] [PubMed]
5. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
6. Lombardi, D.; Pant, S. Nonparametric *k*-nearest-neighbor entropy estimator. *Phys. Rev. E* **2016**, *93*, 013310. [CrossRef] [PubMed]
7. Moddemeijer, R. On estimation of entropy and mutual information of continuous distributions. *Signal Process.* **1989**, *16*, 233–248. [CrossRef]
8. Moon, Y.-I.; Rajagopalan, B.; Lall, U. Estimation of mutual information using kernel density estimators. *Phys. Rev. E* **1995**, *52*, 2318–2321. [CrossRef] [PubMed]
9. Vasicek, O. A Test for Normality Based on Sample Entropy. *J. R. Stat. Soc. B* **1976**, *38*, 54–59. [CrossRef]
10. Khan, S.; Bandyopadhyay, S.; Ganguly, A.R.; Saigal, S.; Erickson, D.J.; Protopopescu, V.; Ostrouchov, G. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E* **2007**, *76*, 026209. [CrossRef] [PubMed]
11. Alizadeh Noughabi, H. Entropy Estimation Using Numerical Methods. *Ann. Data Sci.* **2015**, *2*, 231–241. [CrossRef]
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012; pp. 224–238.

13. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86. [CrossRef]
14. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241. [CrossRef]
15. Nagel, D.; Diez, G.; Stock, G. Accurate estimation of the normalized mutual information of multidimensional data. *J. Chem. Phys.* **2024**, *161*, 054108. [CrossRef]
16. Amari, S. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016. [CrossRef]